RESEARCH CENTER
Rennes - Bretagne-Atlantique

FIELD

Activity Report 2018

Section New Results

Edition: 2019-03-07

CAIRN Project-Team

7. New Results

7.1. Reconfigurable Architecture Design

7.1.1. Algorithmic Fault Tolerance for Timing Speculative Hardware

Participants: Thibaut Marty, Tomofumi Yuki, Steven Derrien.

Timing speculation, also known as overclocking, is a well known approach to increase the computational throughput of processors and hardware accelerators. When used aggressively, timing speculation can lead to incorrect/corrupted results. As reported in the literature, timing errors can cause large numerical errors in the computation, and such occasional large errors can have devastating effect on the final output. The frequency of such errors depends on a number of factors, including the intensity of overclocking, operating temperature, voltage drops, variability within and across boards, input data, and so on. This makes it extremely difficult to determine a "safe" overclocking speed analytically or empirically. Several circuit-level error mitigating techniques have been proposed, but they are difficult to implement in modern FPGAs, and often involve significant area overhead. Instead of resorting to circuit level technique, we propose to rely on light-weight algorithm-level error detections techniques. This allows us to augment accelerators with low overhead mechanism to protect against timing errors, enabling aggressive timing speculation. We have demonstrated in [36] the validity of our approach for convolutional neural networks, where we use overclocking for the convolution stages. Our prototype on ZC706 demonstrated 68-77% computational throughput with negligible (<1%) area overhead.

7.1.2. 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 with 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 [43], [27] show that, compared to a software implementation, our approach offers speed-up by 8$\times$ while consuming 18$\times$ less energy. We have also shown how our approach 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. Our first experimental results have shown that this approach leads to best-case performance and energy efficiency when compared against static VLIW configurations [42].

7.1.3. 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 of heterogeneous multicore architectures using the HAsim/LEAP infrastructure. Two aspects are currently under development. The first one concerns the deployment of simulator models on the hybrid Xeon CPU-Arria 10 FPGA Intel platforms. The second one concerns the definition of simulation models of hardware accelerators. The core processor brick is a RISCV core.

7.1.4. Dynamic Fault-Tolerant Scheduling onto 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. Temporal and/or spatial redundancy is currently used to tackle this issue. 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 several run-time algorithms based on the primary/backup approach which is commonly used for its minimal resources utilization and high reliability [31], [30]. The aim of our work was to reduce the complexity of the algorithm in order to target real-time embedded systems without sacrificing reliability. This work is done in collaboration with Oliver Sinnen, PARC Lab., the University of Auckland.

7.1.5. 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 [21] 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.

7.1.6. Energy Constrained and Real-Time Scheduling and Assignment on Multicores

Participants: Olivier Sentieys, Angeliki Kritikakou, Lei Mo.

Multicore architectures have been used to enhance computing capabilities, but the energy consumption is still an important concern. Embedded application domains usually require less accurate, but always in-time, results. Imprecise Computation (IC) can be used to divide a task into a mandatory subtask providing a baseline Quality-of-Service (QoS) and an optional subtask that further increases the baseline QoS. Combining dynamic voltage and frequency scaling, task allocation and task adjustment, we can maximize the system QoS under real-time and energy supply constraints. However, the nonlinear and combinatorial nature of this problem makes it difficult to solve. In [25], we formulate a Mixed-Integer Non-Linear Programming (MINLP) problem to concurrently carry out task-to-processor allocation, frequency-to-task assignment and optional task adjustment. We provide a Mixed-Integer Linear Programming (MILP) form of this formulation without performance degradation and we propose a novel decomposition algorithm to provide an optimal solution with reduced computation time compared to state-of-the-art optimal approaches (22.6% in average). We also propose a heuristic version that has negligible computation time. In [24], we focus on QoS maximizing for dependent IC-tasks under real-time and energy constraints. Compared with existing approaches, we consider the joint-design problem, where task-to-processor allocation, frequency-to-task assignment, task scheduling and task adjustment are optimized simultaneously. The joint-design problem is formulated as an NP-hard Mixed-Integer Non-Linear Programming and it is safely transformed to a Mixed-Integer Linear Programming (MILP) without performance degradation. Two methods (basic and accelerated version) are proposed to find the optimal solution to MILP problem. They are based on problem decomposition and provide a controllable way to trade-off the quality of the solution and the computational complexity. The optimality of the proposed methods is proved rigorously, and the experimental results show reduced computation time (23.7% in average) compared with existing optimal methods. Finally, in [50] we summarize the problem and the methods for imprecise computation task mapping on multicore Wireless Sensor Networks.

7.1.7. Real-Time Energy-Constrained Scheduling in Wireless Sensor and Actuator Networks

Participants: Angeliki Kritikakou, Lei Mo.

Wireless Sensor and Actuator Networks (WSANs) are emerging as a new generation of Wireless Sensor Networks (WSNs). Due to the coupling between the sensing areas of the sensors and the action areas of the actuators, the efficient coordination among the nodes is a great challenge. In our work in [23] we address the problem of distributed node coordination in WSANs aiming at meeting the user's requirements on the states of the Points of Interest (POIs) in a real-time and energy-efficient manner. The node coordination problem is formulated as a non-linear program. To solve it efficiently, the problem is divided into two correlated subproblems: the Sensor-Actuator (S-A) coordination and the Actuator-Actuator (A-A) coordination. In the S-A coordination, a distributed federated Kalman filter-based estimation approach is applied for the actuators to collaborate with their ambient sensors to estimate the states of the POIs. In the A-A coordination, a distributed Lagrange-based control method is designed for the actuators to optimally adjust their outputs, based on the estimated results from the S-A coordination. The convergence of the proposed method is proved rigorously. As the proposed node coordination scheme is distributed, we find the optimal solution while avoiding high computational complexity. The simulation results also show that the proposed distributed approach is an efficient and practically applicable method with reasonable complexity. In addition, the design of fast and effective coordination among sensors and actuators in Cyber-Physical Systems (CPS) is a fundamental, but challenging issue, especially when the system model is a priori unknown and multiple random events can simultaneously occur. In [37], we propose a novel collaborative state estimation and actuator scheduling algorithm with two phases. In the first phase, we propose a Gaussian Mixture Model (GMM)-based method using the random event physical field distribution to estimate the locations and the states of events. In the second phase, based on the number of identified events and the number of available actuators, we study two actuator scheduling scenarios and formulate them as Integer Linear Programming (ILP) problems with the objective to minimize the actuation delay. We validate and demonstrate the performance of the proposed scheme through both simulations and physical experiments for a home temperature control application.

7.1.8. 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. By including this parameter in the operating system, it is then possible to develop some 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 including P identical cores. We assumed that they can be reconfigured dynamically by authorizing the addition and/or removal of periodic tasks and each core schedules its local tasks by using the EDF algorithm [20]. This work is extended to address heterogeneous systems for which each task has different execution parameters. The objective of this extension work is to develop a new strategy for mapping N tasks to P heterogeneous cores of a given distributed system [32]. For these two architectures models, we formulated the problem as an Integer Linear Program (ILP) optimization problem. Assuming that the energy consumed by the communication is dependent on the distance between cores, we proposed a mapping strategy to minimize the total cost of communication between cores 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-core architecture case. Experimental results reveal the effectiveness of the proposed strategy by comparing the derived heuristics with the optimal ones, obtained by solving an ILP problem

7.1.9. Improving the Reliability of Wireless NoC

Participants: Olivier Sentieys, Joel Ortiz Sosa.

Wireless Network-on-Chip (WiNoC) is one of the most promising solutions to overcome multi-hop latency and high power consumption of modern many/multi core System-on- Chip (SoC). However, the design of efficient wireless links faces challenges to overcome multi-path propagation present in realistic WiNoC channels. In order to alleviate such channel effect, we propose a Time-Diversity Scheme (TDS) to enhance the reliability of on-chip wireless links using a semi-realistic channel model in [45]. First, we study the significant performance degradation of state-of-the-art wireless transceivers subject to different levels of multi-path propagation. Then we investigate the impact of using some channel correction techniques adopting standard performance metrics. Experimental results show that the proposed Time-Diversity Scheme significantly improves Bit Error Rate (BER) compared to other techniques. Moreover, our TDS allows for wireless communication links to be established in conditions where this would be impossible for standard transceiver architectures. Results on the proposed complete transceiver, designed using a 28-nm FDSOI technology, show a power consumption of 0.63mW at 1.0V and an area of 317 μm^2. Full channel correction is performed in one single clock cycle.

7.1.10. Optical Network-on-Chip (ONoC) for 3D Multiprocessor Architectures

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

Photonics on silicon is now a technology that offers real opportunities in the context of multiprocessor interconnect. 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 latency. 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. We formulated the crosstalk noise and latency models and then proposed a Wavelength Allocation (WA) method in a ring-based WDM ONoC to reach performance and energy trade-offs based on application constraints. We show that for a 16-cluster ONoC architecture using 12 wavelengths, more than 10^5 allocation solutions exist and only 51 are on a Pareto front giving a tradeoff between latency and energy per bit derived from the BER. These optimized solutions reduce the execution time of the application by 37% and the energy from 7.6fJ/bit to 4.4fJ/bit. In

[22], we define high-level mechanisms which can handle wavelength allocation protocol of the communication medium for each data transfer between tasks. 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. We produce the communication configurations which are defined by the number of wavelengths for each communication, the level of quality for the communications, and the laser power levels. In [35], we define an Optical-Network-Interface (ONI) to connect a cluster of processors to the optical communication medium. This interface, constrained by the 10 Gb/s data-rate of the lasers, integrates Error Correcting Codes (ECC), laser drivers, and a communication manager. The ONI can select, at run-time, the communication mode to use depending on performance, latency or power constraints. The use of ECC is based on redundant bits which increases the transmission time, but saves power for a given Bit Error Rate (BER). Furthermore, the use of several wavelengths in parallel reduces latency and increases bandwidth, but also increases communication loss.

7.2. Compilation and Synthesis for Reconfigurable Platform

7.2.1. Compile Time Simplification of Sparse Matrix Code Dependences

Participant: Tomofumi Yuki.

Analyzing array-based computations to determine data dependences is useful for many applications including automatic parallelization, race detection, computation and communication overlap, verification, and shape analysis. For sparse matrix codes, array data dependence analysis is made more difficult by the use of index arrays that make it possible to store only the nonzero entries of the matrix (e.g., in $A[B[i]]$, B is an index array). Here, dependence analysis is often stymied by such indirect array accesses due to the values of the index array not being available at compile time. Consequently, many dependences cannot be proven unsatisfiable or determined until runtime. Nonetheless, index arrays in sparse matrix codes often have properties such as monotonicity of index array elements that can be exploited to reduce the amount of runtime analysis needed. In this work, we contribute a formulation of array data dependence analysis that includes encoding index array properties as universally quantified constraints. This makes it possible to leverage existing SMT solvers to determine whether such dependences are unsatisfiable and significantly reduces the number of dependences that require runtime analysis in a set of eight sparse matrix kernels. Another contribution is an algorithm for simplifying the remaining satisfiable data dependences by discovering equalities and/or subset relationships. These simplifications are essential to make a runtime-inspection-based approach feasible.

7.2.2. Automatic Parallelization Techniques for Time-Critical Systems

Participants: Steven Derrien, Mickael Dardaillon.

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 lend 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 [34]. Our current research direction aims [41] at studying the impact of compiler optimization on WCET estimates, and develop specific WCET aware compiler optimization flows. More specifically, we explore the use of iterative compilation (WCET-directed program optimization to explore the optimization space), with the objective to (i) allow flow facts to be automatically found and (ii) select optimizations that result in the lowest WCET estimates. We also explore to which extent code outlining helps, by allowing the selection of different optimization options for different code snippets of the application.

7.2.3. Design of High Throughput Mathematical Function Evaluators

Participant: Silviu Ioan Filip.

The evaluation of mathematical functions is a core component in many computing applications and has been a core topic in computer arithmetic since the inception of the field. In [28], we proposed an automatic method for the evaluation of functions via polynomial or rational approximations and its hardware implementation, on FPGAs. These approximations are evaluated using Ercegovac's iterative E-method adapted for FPGA implementation. The polynomial and rational function coefficients are optimized such that they satisfy the constraints of the E-method. It allows for an effective way to perform design space exploration when targeting high throughput.

7.2.4. Robust Tools for Computing Rational Chebyshev Approximations

Participant: Silviu Ioan Filip.

Rational functions are useful in a plethora of applications, including digital signal processing and model order reduction. They are nevertheless known to be much harder to work with in a numerical context than other, potentially less expressive families of approximating functions, like polynomials. In [19] we have proposed the use of a numerically robust way of representing rational functions, the barycentric form (*i.e.,* a ratio of partial fractions sharing the same poles). We use this form to develop scalable iterative algorithms for computing rational approximations to functions which minimize the uniform norm error. Our results are shown to significantly outperform previous state of the art approaches.

CELTIQUE Project-Team

5. New Results

5.1. Software Fault Isolation

Participants: Frédéric Besson, Thomas Jensen, Julien Lepiller.

Software Fault Isolation (SFI) consists in transforming untrusted code so that it runs within a specific address space, (called the sandbox) and verifying at load-time that the binary code does indeed stay inside the sandbox. Security is guaranteed solely by the SFI verifier whose correctness therefore becomes crucial. Existing verifiers enforce a very rigid, almost syntactic policy where every memory access and every control-flow transfer must be preceded by a sandboxing instruction sequence, and where calls outside the sandbox must implement a sophisticated protocol based on a shadow stack. We have defined SFI as a defensive semantics, with the purpose of deriving semantically sound verifiers that admit flexible and efficient implementations of SFI. We derive an executable analyser, that works on a per-function basis, which ensures that the defensive semantics does not go wrong, and hence that the code is well isolated. Experiments show that our analyser exhibits the desired flexibility: it validates correctly sandboxed code, it catches code breaking the SFI policy, and it can validate programs where redundant instrumentations are optimised away [8].

5.2. Compilation and Side-Channels

Participants: Frédéric Besson, Alexandre Dang, Thomas Jensen.

The usual guarantee provided by compilers is that the input/output observable behaviour of the target program is one of the possible behaviours of the source program. In the context of security, the notion of observable behaviour needs to be revisited in order to take into account side-channels *i.e.* observations beyond input/output that are leaked by the program execution to a potential attacker.

For instance, a common security recommendation is to reduce the in-memory lifetime of secret values, in order to reduce the risk that an attacker can obtain secret data by probing memory. To mitigate this risk, secret values can be overwritten, at source level, after their last use. However, as secret values are never used afterwards, a compiler may remove this mitigation during a standard Dead Store Elimination pass. We propose a formal definition of Information Flow Preserving transformation [7] which ensures that secret values are not easier to obtain at assembly level than at source level. Using the notion of Attacker Knowledge, we relate the information leak of a program before and after the transformation. We consider two classic compiler passes (Dead Store Elimination and Register Allocation) and show how to validate and, if needed, modify these transformations in order to be information flow preserving.

5.3. Semantic study of the Sea-of-Nodes form

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

As part of the PhD of Yon Fernandez de Retana [2], we started to study the the Sea-of-Nodes form. This intermediate representation was introduced by Cliff Click in the mid 90s [21] as an enhanced SSA form. It improves on the initial SSA form by relaxing the total order on instructions in basic blocks into explicit data and control dependencies. This makes programs more flexible to optimize. While Sea-of-node is popular in many production-size compiler (Sun's HotSpot, Graal...), it is still not very well understood, from a semantic, foundational point of view. We have defined a simple but rigorous formal semantics for a Sea-of-Nodes form. It comprises a denotational component to express data computation, and an operational component to express control flow. We prove a fundamental, dominance-based semantic property on Sea-of-Nodes programs which determines the regions of the graph where the values of nodes are preserved. Finally, we apply our results to prove the semantic correctness of a redundant zero-check elimination optimization. All the necessary semantic properties have been mechanically verified in the Coq proof assistant. These results have been published in [11]. A more detailed account can be found in Yon Fernandez de Retana's PhD manuscript [2].

5.4. Certified Concurrent Garbage Collector

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

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 [6] as an extended version of [22].

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 [20].

5.5. Formalization of Higher-Order Process Calculi

Participants: Guillaume Ambal, Sergueï Lenglet, Alan Schmitt.

Guillaume Amabal, Sergueï Lenglet, Alan Schmitt have continued exploring how to formalize HOπ in Coq, in particular how to deal with the different kinds of binders used in the calculus. We have studied and compared several approaches, such as locally nameless, De Bruijn indices, and nominal binders. We have discovered that the locally nameless approach introduces a lot of complexity, as name restriction allows reduction under binders, introducing the need for numerous renaming lemmas. The nominal approach is quite elegant and very close to the pen and paper definitions, but it still requires many technical lemmas to be proven. The de Bruijn approach is the most concise. A first version of this work has been published at CPP 2018 [18] and a journal version is submitted for publication. The Coq scripts can be found at http://passivation.gforge.inria.fr/hopi/.

5.6. Certified Semantics and Analyses for JavaScript

Participants: Samuel Risbourg, 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. Since September 2018, Samuel Risbourg has been hired to continue developing the tool. It is publicly available at https://github.com/jscert/jsexplain and is described in a publication at The Web Conference 2018 [10]. The tool is regularly presented at the TC39 committee standardizing JavaScript to solicit feedback.

5.7. Skeletal Semantics

Participants: Nathanael Courant, Thomas Jensen, Alan Schmitt.

Alan Schmitt and Thomas Jensen, in collaboration with Martin Bodin and Philippa Gardner at Imperial College London, have designed a new meta-language to formally describe semantics. A fundamental idea behind this approach to semantics is that this description can be used to derive several *interpretations* corresponding to different kinds of semantics, such as a big-step semantics, an abstract interpretation, or a control-flow analysis. The correctness of these semantics is proven independently of the language considered. This work has been accepted at POPL 2019 [4] and is formalized in Coq (see the skeletal semantics web site for more detail). Nathanael Courant is currently extending this work to generate analyses automatically and to facilitate the way in which to adjust their precision.

5.8. Verification of High-Level Transformations with Inductive Refinement Types

Participant: Thomas Jensen.

High-level transformation languages like Rascal or Stratego include expressive features for manipulating large abstract syntax trees: first-class traversals, expressive pattern matching, backtracking and generalized iterators. We have designed and implemented an abstract interpretation tool, Rabit, for verifying inductive type and shape properties for transformations written in such languages. We describe how to perform abstract interpretation based on operational semantics, specifically focusing on the challenges arising when analyzing the expressive traversals and pattern matching. We have evaluated Rabit on a series of transformations (normalization, desugaring, refactoring, code generators, type inference, etc.) showing that we can effectively verify stated properties.

This work was done in collaboration with researchers at the IT University of Copenhagen. The paper [19] presenting these results won the Best Paper award at GPCE 2018.

5.9. 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. This year we successfully extended this result to the case of higher-order functional programs [15], [16]. Moreover, the approximation automaton can be certified using an efficient Coq-extracted checker that we developped in 2008. Thus, we have an automatic static analysis procedure for higher-order functional programs whose results are certified by the Coq proof assistant. The algorithm presented in [15] has been implemented in Timbuk [14] and gives very encouraging experimental results http://people.irisa.fr/Thomas.Genet/timbuk/funExperiments/. Besides, we have shown the completeness of this approach, i.e., that any regular approximation of the image of a function can be found using completion [13].

CIDRE Project-Team

6. New Results

6.1. Axis 1 : Attack comprehension

6.1.1. Attacks stay possible even when programs seem not vulnerable

The protection of any software starts at the hardware level. In ,K. Bukasa, L. Claudepierre, J.-L. Lanet, in collaboration with R. Lashermes from SED Inria Rennes – Bretagne Atlantique, explore how Electromagnetic Fault Injection (EMFI) can disturb the behavior of a chip and undermine the security of the information handled by the target. They demonstrate the possibilities to create software vulnerabilities with hardware fault injection (with EM pulses), not against crypto-systems but targeting regular software running on IoT devices. Experimentations are conducted on an ARMv7-M (Cortex-M3) microcontroller, present at the heart of a wide-range of embedded systems, to prove that a fault attack is able to create a vulnerability in a code where there is none in the usual software security meaning. Protecting against vulnerabilities must thus encompass protecting against both software and hardware attacks.

6.2. Axis 2 : Attack detection

6.2.1. Intrusion detection in sequential control systems.

Sophisticated process-aware attacks targeting industrial control systems require adequate detection measures taking into account the physical process. In [20], 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 the mining efficiency and attack detection capabilities of our approach.

6.2.2. Hardware-based Information Flow Tracking

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.

Existing hardware DIFT approaches have not been widely used neither by research community nor by hardware vendors. It is due to two major reasons: current hardware DIFT solutions lack support for multi-threaded applications and implementations for hardcore processors. In [10] we addresse both issues by introducing an approach with some unique features: DIFT for multi-threaded software, virtual memory protection (rather than physical memory as in related works) and Linux kernel support using an information flow monitor called RFBlare. These goals are accomplished by taking advantage of a notable feature of ARM CoreSight components (context ID) combined with a custom DIFT coprocessor and RFBlare. The communication time overhead, major source of slowdown in total DIFT time overhead, is divided by a factor 3.8 compared to existing solutions with similar software constraints as in this work. The area overhead of this work is lower than 1% and power overhead is 16.2% on a middle-class Xilinx Zynq SoC.

Most of hardware-assisted solutions for software security, program monitoring, and event-checking approaches require instrumentation of the target software, an operation which can be performed using an SBI (Static Binary Instrumentation) or a DBI (Dynamic Binary Instrumentation) framework. Hardware-assisted instrumentation can use one of these two solutions to instrument data to a memory-mapped register. Both these approaches require an in-depth knowledge of frameworks and an important amount of software modifications in order to instrument a whole application. In [11] we propose a novel way to instrument an application, at the source code level, taking advantage of underlying hardware debug components such as CS (CoreSight) components available on Xilinx Zynq SoCs. As an example, the instrumentation approach proposed in this work is used to detect a double free security attack. Furthermore, it is evaluated in terms of runtime and area overhead.

6.2.3. Alert correlation in intrusion detection.

In distributed systems and in particular in industrial SCADA environments, alert correlation systems are necessary to identify complex multi-step attacks within the huge amount of alerts and events. In [22] we describe an automata-based correlation engine developed in the context of a European project where the main stakeholder was an energy distribution company. The behavior of the engine is extended to fit new requirements. In the proposed solution, a fully automated process generates thousands of correlation rules. Despite this major scalability challenge, the designed correlation engine exhibits good performance. Expected rates of incoming low level alerts approaching several hundreds of elements per second are tolerated. Moreover, the data structures chosen allow to quickly handle dynamic changes of the set of correlation rules. As some attack steps are not observed, the correlation engine can be tuned to raise an alert when all the attack steps except k of them have been detected. To be able to react to an ongoing attack by taking countermeasures, alerts must also be raised as soon as a significant prefix of an attack scenario is recognized. Fulfilling these additional requirements leads to an increase in the memory consumption. Therefore purge mechanisms are also proposed and analyzed. An evaluation of the tool is conducted in the context of a SCADA environment.

6.2.4. Most recent and frequent items in distributed streams for DDoS detection.

The need to analyze in real time large-scale and distributed data streams has recently became tremendously important to detect attacks (DDoS), anomalies or performance issues. In particular the identification of recent heavy-hitters (or hot items) is essential but highly challenging. Actually, this problem has been heavily studied during the last decades with both exact and probabilistic solutions. While simple to state and fundamental for advanced analysis, answering this issue over a sliding time window and among distributed nodes is still an active research field. The distributed detection of frequent items over a sliding time window presents two extra challenging aspects with respect to the centralized detection of frequent items since the inception of the stream: (i) Treat time decaying items as they enter and exit the sliding window; (ii) Produce mergeable local stream summaries in order to obtain a system-wide summary. In [12], we propose a sliding window-based solution of the top k most frequent items based on a deterministic counting of the most over-represented items in the data streams, which are themselves probabilistically identified using a dynamically defined threshold. Performance of our new algorithm are astonishingly good, despite any items order manipulation or distributed execution.

6.2.5. Propagation of information.

Together with Yves Mocquard and Bruno Sericola, we have worked on the well studied dissemination of information in large scale distributed networks through pairwise interactions. The information to be propagated can simply be a bit of information to any code, including viruses. This problem, originally called rumor mongering, and then rumor spreading has mainly been investigated in the synchronous model. This model 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 [4]. In addition to this study, we have proposed an algorithm that allows, through simple pairwise interactions, each node of the large scale and dynamic system to build a global clock which allows any node to maintain with high probability a common temporal referential [25]. By combining this global clock together with the rumor spreading algorithm, we have proposed a mechanism that allows each node to locally detect that the system has converged to a sought configuration with high probability. We have also shown the applicability of our convergence detection mechanism to many other pairwise interaction-based protocols. For instance, our construction can be applied to a leader election protocol provided that its convergence time is known with high probability [26].

6.3. Axis 3 : Attack resistance

6.3.1. Connectivity in an inter-MANET network.

New generation radio equipment, used by soldiers and vehicles on the battlefield, form ad hoc networks and specifically, Mobile Ad hoc NETworks (MANET). The battlefields where these equipments are deployed include a majority of coalition communication. Each group on the battleground may communicate with other members of the coalition and establish inter-MANET links. These inter-MANET links are governed by routing policies that can be summarized as Allowed or Denied link. However, if more than two groups form a coalition, blocked multi-hop communications and non-desired transmissions due to these restrictive policies would appear. In [19], we present these blocking cases and theoretically evaluate their apparition frequency. Then, we present two alternatives to extend the binary policies and decrease the number of blocking cases. Finally, we describe an experimental scenario containing a blocking case and evaluate our propositions and their performance.

6.3.2. Permissionless ledgers for decentralized cryptocurrency systems (blockchain).

The goal of decentralized cryptocurrency systems is to offer a medium of exchange secured by cryptography, without the need of a centralized banking authority. An increasing number of distributed cryptocurrency systems are emerging, and among them Bitcoin, which is often designated as the pioneer of this kind of systems. Bitcoin circumvents the absence of a global trusted third-party by relying on a blockchain, an append-only data-structure, publicly readable and writable, in which all the valid transactions ever issued in the system are progressively appended through the creation of cryptographically linked blocks. In [15], we propose a new way to organise both transactions and blocks in a distributed ledger to address the performance issues of permissionless ledgers. In contrast to most of the existing solutions in which the ledger is a chain of blocks extracted from a tree or a graph of chains, we present a distributed ledger whose structure is a balanced directed acyclic graph of blocks. We call this specific graph a SYC-DAG. We show that a SYC-DAG allows us to keep all the remarkable properties of the Bitcoin blockchain in terms of security, immutability, and transparency, while enjoying higher throughput and self-adaptivity to transactions demand.

6.3.3. Modular verification of Programs with Effects and Effect Handlers in Coq

Modern computing systems have grown in complexity, and the attack surface has increased accordingly. Even though system components are generally carefully designed and even verified by different groups of people, the composition of these components is often regarded with less attention. This paves the way for architectural attacks, a class of security vulnerabilities where the attacker is able to threaten the security of the system even if each of its components continues to act as expected. In [24], we introduce FreeSpec, a formalism built upon the key idea that components can be modelled as programs with algebraic effects to be realized by other components. FreeSpec allows for the modular modelling of a complex system, by defining idealized components connected together, and the modular verification of the properties of their composition. In addition, we have implemented a framework for the Coq proof assistant based on FreeSpec.

GALLINETTE Project-Team

6. New Results

6.1. Logical Foundations of Programming Languages

Participants: Rémi Douence, Ambroise Lafont, Étienne Miquey, Xavier Montillet, Guillaume Munch-Maccagnoni, Nicolas Tabareau, Pierre Vial.

6.1.1. Classical Logic

6.1.1.1. A sequent calculus with dependent types for classical arithmetic.

In a recent paper, Herbelin developed a calculus $dPA\omega$ in which constructive proofs for the axioms of countable and dependent choices could be derived via the encoding of a proof of countable universal quantification as a stream of it components. However, the property of normalisation (and therefore the one of soundness) was only conjectured. The difficulty for the proof of normalisation is due to the simultaneous presence of dependent dependent types (for the constructive part of the choice), of control operators (for classical logic), of coinductive objects (to encode functions of type $N \rightarrow A$ into streams $(a_0, a_1, \cdots)$) and of lazy evaluation with sharing (for these coinductive objects).Building on previous works, we introduce in [14], [26] a variant of $dPA\omega$ presented as a sequent calculus. On the one hand, we take advantage of a variant of Krivine classical realisability we developed to prove the normalisation of classical call-by-need. On the other hand, we benefit of dL, a classical sequent calculus with dependent types in which type safety is ensured using delimited continuations together with a syntactic restriction. By combining the techniques developed in these papers, we manage to define a realisability interpretation à la Krivine of our calculus that allows us to prove normalisation and soundness.

6.1.1.2. Realisability Interpretation and Normalisation of Typed Call-by-Need λ-calculus With Control.

In [13], we define a variant of realisability where realisers are pairs of a term and a substitution. This variant allows us to prove the normalisation of a simply-typed call-by-need λ-calculus with control due to Ariola et al. Indeed, in such call-by-need calculus, substitutions have to be delayed until knowing if an argument is really needed. In a second step, we extend the proof to a call-by-need λ-calculus equipped with a type system equivalent to classical second-order predicate logic, representing one step towards proving the normalisation of the call-by-need classical second-order arithmetic introduced by the second author to provide a proof-as-program interpretation of the axiom of dependent choice.

6.1.2. Lambda Calculus

6.1.2.1. Every λ-Term is Meaningful for the Infinitary Relational Model.

Infinite types and formulas are known to have really curious and unsound behaviours. For instance, they allow to type Ω, the auto-autoapplication and they thus do not ensure any form of normalisation/productivity. Moreover, in most infinitary frameworks, it is not difficult to define a type R that can be assigned to every λ-term. However, these observations do not say much about what coinductive (i.e. infinitary) type grammars are able to provide: it is for instance very difficult to know what types (besides R) can be assigned to a given term in this setting. In [17], we begin with a discussion on the expressivity of different forms of infinite types. Then, using the resource-awareness of sequential intersection types (system S) and tracking, we prove that infinite types are able to characterise the arity of every λ-terms and that, in the infinitary extension of the relational model, every term has a " meaning " i.e. a non-empty denotation. From the technical point of view, we must deal with the total lack of guarantee of productivity for typable terms: we do so by importing methods inspired by first order model theory.

6.1.2.2. High-level signatures and initial semantics.

In [9], we present a device for specifying and reasoning about syntax for datatypes, programming languages, and logic calculi. More precisely, we consider a general notion of signature for specifying syntactic constructions. Our signatures subsume classical algebraic signatures (i.e., signatures for languages with variable binding, such as the pure lambda calculus) and extend to much more general examples. In the spirit of Initial Semantics, we define the syntax generated by a signature to be the initial object—if it exists—in a suitable category of models. Our notions of signature and syntax are suited for compositionality and provide, beyond the desired algebra of terms, a well-behaved substitution and the associated inductive/recursive principles. Our signatures are general in the sense that the existence of syntax is not automatically guaranteed. In this work, we identify a large class of signatures which do generate a syntax. This paper builds upon ideas from a previous attempt by Hirschowitz-Maggesi (FICS 2012), which, in turn, was directly inspired by some earlier work of Ghani-Uustalu and Matthes-Uustalu. The main results presented in the paper are computer-checked within the UniMath system.

6.1.3. Models of programming languages mixing effects and resources

6.1.3.1. A resource modality for RAII

Systems programming languages C++11 and Rust have developed techniques and idioms for the safe management of resources called *"Resource acquisition is initialisation"* (RAII) and *move semantics*. We have related resources from systems programming to the notion of resource put forward by linear logic, by giving a construction in terms of categorical semantics for a resource modality that model RAII and move semantics. This work was presented by at the workshop LOLA 2018 in Oxford [20].

6.1.3.2. Resource polymorphism

Thanks to a new logical and semantic understanding of resource-management techniques in systems programming languages, we have proposed [27] a design for an extension of functional programming language towards systems programming, centred on the OCaml language, and based on a notion of resource polymorphism inspired by the C++11 language and by the works on polarisation in proof theory.

6.1.4. Distributed Programming

6.1.4.1. Chemical foundations of distributed aspects.

Distributed applications are challenging to program because they have to deal with a plethora of concerns, including synchronisation, locality, replication, security and fault tolerance. Aspect-oriented programming (AOP) is a paradigm that promotes better modularity by providing means to encapsulate cross-cutting concerns in entities called aspects. Over the last years, a number of distributed aspect-oriented programming languages and systems have been proposed, illustrating the benefits of AOP in a distributed setting. Chemical calculi are particularly well-suited to formally specify the behaviour of concurrent and distributed systems. The join calculus is a functional name-passing calculus, with both distributed and object-oriented extensions. It is used as the basis of concurrency and distribution features in several mainstream languages like C# (Polyphonic C#, now Cω), OCaml (JoCaml), and Scala Joins. Unsurprisingly, practical programming in the join calculus also suffers from modularity issues when dealing with crosscutting concerns. We propose the Aspect Join Calculus [8], an aspect-oriented and distributed variant of the join calculus that addresses crosscutting and provides a formal foundation for distributed AOP. We develop a minimal aspect join calculus that allows aspects to advise chemical reactions. We show how to deal with causal relations in pointcuts and how to support advanced customisable aspect weaving semantics.

6.2. Type Theory and Proof Assistants

Participants: Simon Boulier, Eric Finster, Gaëtan Gilbert, Pierre-Marie Pédrot, Nicolas Tabareau, Théo Winterhalter.

6.2.1. Type Theory

6.2.1.1. Effects in Type Theory.

In [16] , we define the exceptional translation, a syntactic translation of the Calculus of Inductive Constructions (CIC) into itself, that covers full dependent elimination. The new resulting type theory features call-by-name exceptions with decidable type-checking and canonicity, but at the price of inconsistency. Then, noticing parametricity amounts to Kreisel's realisability in this setting, we provide an additional layer on top of the exceptional translation in order to tame exceptions and ensure that all exceptions used locally are caught, leading to the parametric exceptional translation which fully preserves consistency. This way, we can consistently extend the logical expressivity of CIC with independence of premises, Markov's rule, and the negation of function extensionality while retaining η-expansion. As a byproduct, we also show that Markov's principle is not provable in CIC. Both translations have been implemented in a Coq plugin, which we use to formalise the examples.

6.2.1.2. Eliminating Reflection from Type Theory.

Type theories with equality reflection, such as extensional type theory (ETT), are convenient theories in which to formalise mathematics, as they make it possible to consider provably equal terms as convertible. Although type-checking is undecidable in this context, variants of ETT have been implemented, for example in NuPRL and more recently in Andromeda. The actual objects that can be checked are not proof-terms, but derivations of proof-terms. This suggests that any derivation of ETT can be translated into a typecheckable proof term of intensional type theory (ITT). However, this result, investigated categorically by Hofmann in 1995, and 10 years later more syntactically by Oury, has never given rise to an effective translation. In [18], we provide the first syntactical translation from ETT to ITT with uniqueness of identity proofs and functional extensionality. This translation has been defined and proven correct in Coq and yields an executable plugin that translates a derivation in ETT into an actual Coq typing judgment. Additionally, we show how this result is extended in the context of homotopy to a two-level type theory.

6.2.1.3. Foundations of Dependent Interoperability.

Full-spectrum dependent types promise to enable the development of correct-by-construction software. However, even certified software needs to interact with simply-typed or untyped programs, be it to perform system calls, or to use legacy libraries. Trading static guarantees for runtime checks, the dependent interoperability framework provides a mechanism by which simply-typed values can safely be coerced to dependent types and, conversely, dependently-typed programs can defensively be exported to a simply-typed application. In [2], we give a semantic account of dependent interoperability. Our presentation relies on and is guided by a pervading notion of type equivalence, whose importance has been emphasised in recent work on homotopy type theory. Specifically, we develop the notion of type-theoretic partial Galois connections as a key foundation for dependent interoperability, which accounts for the partiality of the coercions between types. We explore the applicability of both monotone and antitone type-theoretic Galois connections in the setting of dependent interoperability. A monotone partial Galois connection enforces a translation of dependent types to runtime checks that are both sound and complete with respect to the invariants encoded by dependent types. Conversely, picking an antitone partial Galois connection instead lets us induce weaker, sound conditions that can amount to more efficient runtime checks. Our framework is developed in Coq; it is thus constructive and verified in the strictest sense of the terms. Using our library, users can specify domain-specific partial connections between data structures. Our library then takes care of the (sometimes, heavy) lifting that leads to interoperable programs. It thus becomes possible, as we shall illustrate, to internalise and hand-tune the extraction of dependently-typed programs to interoperable OCaml programs within Coq itself.

6.2.1.4. Equivalences for Free: Univalent Parametricity for Effective Transport.

Homotopy Type Theory promises a unification of the concepts of equality and equivalence in Type Theory, through the introduction of the univalence principle. However, existing proof assistants based on type theory treat this principle as an axiom, and it is not yet clear how to extend them to handle univalence internally. In [7], we propose a construction grounded on a univalent version of parametricity to bring the benefits of univalence to the programmer and prover, that can be used on top of existing type theories. In particular, univalent parametricity strengthens parametricity to ensure preservation of type equivalences. We present a lightweight framework implemented in the Coq proof assistant that allows the user to transparently transfer definitions and theorems for a type to an equivalent one, as if they were equal. Our approach handles both type and term dependency. We study how to maximise the effectiveness of these transports in terms of computational behaviour, and identify a fragment useful for certified programming on which univalent transport is guaranteed to be effective. This work paves the way to easier-to-use environments for certified programming by supporting seamless programming and proving modulo equivalences.

6.2.1.5. Special Issue on Homotopy Type Theory and Univalent Foundations.

The preface [4] introduces the first special issue out of a series of workshops on Homotopy Type Theory and Univalent Foundations. This recent area of research finds its roots in the seminal work of Martin Hofmann and Thomas Streicher on the structure of Martin-Löf identity types. But the main research program has been foreseen by Vladimir Voevodsky, who, from its initial motivation of formalising his results in homotopy theory, has initiated what is now called the univalent foundations program. Borrowing ideas from homotopy theory, the goal of the univalent foundations program is to leverage dependent Type Theory to a formal framework that could replace Set Theory for the foundations of mathematics. This special issue gathers research contributions of some of the most prominent researchers of the field.

6.2.1.6. Goodwillie's Calculus of Functors and Higher Topos Theory

In [1], we develop an approach to Goodwillie's calculus of functors using the techniques of higher topos theory. Central to our method is the introduction of the notion of fiberwise orthogonality, a strengthening of ordinary orthogonality which allows us to give a number of useful characterisations of the class of n-excisive maps. We use these results to show that the pushout product of a P_n-equivalence with a P_m-equivalence is a P_{m+n+1}-equivalence. Then, building on our previous work, we prove a Blakers-Massey type theorem for the Goodwillie tower. We show how to use the resulting techniques to rederive some foundational theorems in the subject, such as delooping of homogeneous functors.

6.2.2. Proof Assistants

6.2.2.1. Typed Template Coq – Certified Meta-Programming in Coq.

Template-Coq [19], [10] is a plugin for Coq, originally implemented by Malecha, which provides a reifier for Coq terms and global declarations , as represented in the Coq kernel, as well as a denotation command. Initially, it was developed for the purpose of writing functions on Coq's AST in Gallina. Recently, it was used in the CertiCoq certified compiler project, as its front-end language, to derive parametricity properties, and to extract Coq terms to a CBV λ-calculus. However, the syntax lacked semantics, be it typing semantics or operational semantics, which should reflect, as formal specifications in Coq, the semantics of Coq's type theory itself. The tool was also rather bare bones, providing only rudimentary quoting and unquoting commands. We generalise it to handle the entire Calculus of Inductive Constructions (CIC), as implemented by Coq, including the kernel's declaration structures for definitions and inductives, and implement a monad for general manipulation of Coq's logical environment. We demonstrate how this setup allows Coq users to define many kinds of general purpose plugins, whose correctness can be readily proved in the system itself, and that can be run efficiently after extraction. We give a few examples of implemented plugins, including a parametricity translation. We also advocate the use of Template-Coq as a foundation for higher-level tools.

6.2.2.2. Definitional Proof-Irrelevance without K.

Definitional equality—or conversion—for a type theory with a decidable type checking is the simplest tool to prove that two objects are the same, letting the system decide just using computation. Therefore, the more things are equal by conversion, the simpler it is to use a language based on type theory. Proof-irrelevance, stating that any two proofs of the same proposition are equal, is a possible way to extend conversion to make a type theory more powerful. However, this new power comes at a price if we integrate it naively, either by making type checking undecidable or by realising new axioms—such as uniqueness of identity proofs (UIP)—that are incompatible with other extensions, such as univalence. In [3], taking inspiration from homotopy type theory, we propose a general way to extend a type theory with definitional proof irrelevance, in a way that keeps type checking decidable and is compatible with univalence. We provide a new criterion to decide whether a proposition can be eliminated over a type (correcting and improving the so-called singleton elimination of Coq) by using techniques coming from recent development on dependent pattern matching without UIP. We show the generality of our approach by providing implementations for both Coq and Agda, both of which are planned to be integrated in future versions of those proof assistants.

6.3. Program Certifications and Formalisation of Mathematics

Participants: Danil Annenkov, Assia Mahboubi, Étienne Miquey.

6.3.1. Certified Compilation of Financial Contracts.

In [11], we present an extension to a certified financial contract management system that allows for templated declarative financial contracts and for integration with financial stochastic models through verified compilation into so-called payoff-expressions. Such expressions readily allow for determining the value of a contract in a given evaluation context, such as contexts created for stochastic simulations. The templating mechanism is useful both at the contract specification level, for writing generic reusable contracts, and for reuse of code that, without the templating mechanism, needs to be recompiled for different evaluation contexts. We report on the effect of using the certified system in the context of a GPGPU-based Monte Carlo simulation engine for pricing various over-the-counter (OTC) financial contracts. The full contract-management system, including the payoff-language compilation, is verified in the Coq proof assistant and certified Haskell code is extracted from our Coq development along with Futhark code for use in a data-parallel pricing engine.

6.3.2. Static interpretation of higher-order modules in Futhark: functional GPU programming in the large.

In [12], we present a higher-order module system for the purely functional data-parallel array language Futhark. The module language has the property that it is completely eliminated at compile time, yet it serves as a powerful tool for organising libraries and complete programs. The presentation includes a static and a dynamic semantics for the language in terms of, respectively, a static type system and a provably terminating elaboration of terms into terms of an underlying target language. The development is formalised in Coq using a novel encoding of semantic objects based on products, sets, and finite maps. The module language features a unified treatment of module type abstraction and core language polymorphism and is rich enough for expressing practical forms of module composition.

6.3.3. Formalising Implicative Algebras in Coq.

In [15], we present a Coq formalisation of Alexandre Miquel's implicative algebras, which aim at providing a general algebraic framework for the study of classical realisability models. We first give a self-contained presentation of the underlying implicative structures, which roughly consists of a complete lattice equipped with a binary law representing the implication. We then explain how these structures can beturned into models by adding separators, giving rise to the so-called implicative algebras. Additionally, we show how they generalise Boolean and Heyting algebras as well as the usual algebraic structures used in the analysis of classical realisability.

6.3.4. Formally Verified Approximations of Definite Integrals.

Finding an elementary form for an antiderivative is often a difficult task, so numerical integration has become a common tool when it comes to making sense of a definite integral. Some of the numerical integration methods can even be made rigorous: not only do they compute an approximation of the integral value but they also bound its inaccuracy. Yet numerical integration is still missing from the toolbox when performing formal proofs in analysis. In [5], we present an efficient method for automatically computing and proving bounds on some definite integrals inside the Coq formal system. Our approach is not based on traditional quadrature methods such as Newton-Cotes formulas. Instead, it relies on computing and evaluating antiderivatives of rigorous polynomial approximations, combined with an adaptive domain splitting. Our approach also handles improper integrals, provided that a factor of the integrand belongs to a catalog of identified integrable functions. This work has been integrated to the CoqInterval library.

HYCOMES Project-Team

6. New Results

6.1. Hybrid Systems Modeling and Verification

6.1.1. Building a Hybrid Systems Modeler on Synchronous Languages Principles

Participants: Albert Benveniste, Benoît Caillaud.

Hybrid systems modeling languages that mix discrete and continuous time signals and systems are widely used to develop Cyber-Physical systems where control software interacts with physical devices. Compilers play a central role, statically checking source models, generating intermediate representations for testing and verification, and producing sequential code for simulation and execution on target platforms. In [5], Albert Benveniste, Timothy Bourke (PARKAS team Inria/ENS Paris), Benoît Caillaud, Jean-Louis Colaço, Cédric Pasteur (ANSYS/Esterel Technologies, Toulouse) and Marc Pouzet (PARKAS team Inria/ENS Paris) propose a comprehensive study of hybrid systems modeling languages (formal semantics, causality analysis, compiler design, ...). This paper advocates a novel approach to the design and implementation of these languages, built on synchronous language principles and their proven compilation techniques. The result is a hybrid systems modeling language in which synchronous programming constructs can be mixed with Ordinary Differential Equations (ODEs) and zero-crossing events, and a runtime that delegates their approximation to an off-the-shelf numerical solver. We propose an ideal semantics based on non standard analysis, which defines the execution of a hybrid model as an infinite sequence of infinitesimally small time steps. It is used to specify and prove correct three essential compilation steps: (1) a type system that guarantees that a continuous-time signal is never used where a discrete-time one is expected and conversely; (2) a type system that ensures the absence of combinatorial loops; (3) the generation of statically scheduled code for efficient execution. Our approach has been evaluated in two implementations: the academic language Zélus, which extends a language reminiscent of Lustre with ODEs and zero-crossing events, and the industrial prototype Scade Hybrid, a conservative extension of Scade 6.

6.1.2. Structural Analysis of Differential-Algebraic Equations (DAE), State-of-the-Art

Participants: Khalil Ghorbal, Mathias Malandain.

In a deliverable [0] for the FUI ModeliScale collaborative project, Mathias Malandain and Khalil Ghorbal discuss the state-of-the-art methods for performing what is called structural index reduction for differential-algebraic equations, that is equations involving both differential and algebraic equality constraints. Index reduction is one of the basic required methods implemented in any DAE-based modelling language (like Modelica). It is a mandatory step to perform prior to calling a numerical solver to effectively advance time by integrating the set of equations. We cover in particular a recent work that tackles extended models involving several modes, each of which is encoded as a standard DAE.

6.1.3. Multi-Mode DAE Models: Challenges, Theory and Implementation

Participants: Albert Benveniste, Benoît Caillaud, Khalil Ghorbal.

[0] Modeliscale project, deliverable M2.1.1 1, Structural Analysis of Differential-Algebraic Equations (DAE), State-of-the-Art.

The modeling and simulation of Cyber-Physical Systems (CPS) such as robots, vehicles, and power plants often require models with a time varying structure, due to failure situations or due to changes in physical conditions. These are called multi-mode models. In [17], Albert Benveniste, Benoît Caillaud, Hilding Elmqvist (Mogram AB, Lund, Sweden), Khalil Ghorbal, Martin Otter (DLR-SR, Oberpfaffenhofen, Germany) and Marc Pouzet (PARKAS team, Inria/ENS Paris) are interested in multi-domain, component-oriented modeling as performed, for example, with the modeling language Modelica that leads naturally to Differential Algebraic Equations (DAEs). This paper is thus about multi-mode DAE systems. In particular, new methods are introduced to overcome one key problem that was only solved for specific subclasses of systems before: How to switch from one mode to another one when the number of equations may change and variables may exhibit impulsive behavior? An evaluation is performed both with the experimental modeling and simulation system Modia, a domain specific language extension of the programming language Julia, and with SunDAE, a novel structural analysis library for multi-mode DAE systems.

6.1.4. Vector Barrier Certificates and Comparison Systems

Participant: Khalil Ghorbal.

Vector Lyapunov functions are a multi-dimensional extension of the more familiar (scalar) Lyapunov functions, commonly used to prove stability properties in systems of non-linear ordinary differential equations (ODEs). In [7], Kahlil Ghorbal and Andrew Sogokon (CMU, Pittsburgh, USA) explore an analogous vector extension for so-called barrier certificates used in safety verification. As with vector Lyapunov functions, the approach hinges on constructing appropriate comparison systems, i.e., related differential equation systems from which properties of the original system may be inferred. The paper presents an accessible development of the approach, demonstrates that most previous notions of barrier certificate are special cases of comparison systems, and discusses the potential applications of vector barrier certificates in safety verification and invariant synthesis.

6.2. Contract-based Reasoning for Cyper-Physical Systems Design

6.2.1. Contracts for Cyper-Physical Systems Design

Participants: Albert Benveniste, Benoît Caillaud.

Contract-based reasoningn has been proposed as an “orthogonal” approach that complements methodologies proposed so far to cope with the complexity of cyber-physical systems design. Contract-based reasoning provides a rigorous framework for the verification, analysis, abstraction/refinement, and even synthesis of cyber-physical systems. A number of results have been obtained in this domain but a unified treatment of the topic that can help put contract-based design in perspective was missing. In [6], Albert Benveniste, Benoît Caillaud and co-authors provide a unified theory where contracts are precisely defined and characterized so that they can be used in design methodologies with no ambiguity. This monograph gathers research results of the former S4 inria team. It identifies the essence of complex system design using contracts through a *mathematical meta-theory*, where all the properties of the methodology are derived from an abstract and generic notion of contract. We show that the meta-theory provides deep and enlightening links with existing contract and interface theories, as well as guidelines for designing new theories. Our study encompasses contracts for both software and systems, with emphasis on the latter. We illustrate the use of contracts with two examples: requirement engineering for a parking garage management, and the development of contracts for timing and scheduling in the context of the Autosar methodology in use in the automotive sector.

6.2.2. Cyber-Physical Systems Design: from Natural Language Requirements

In his current PhD work, co-supervised by Benoît Caillaud and Annie Forêt (SemLIS, IRISA, Rennes, France), Aurélien Lamercerie explores the construction of formal representations of natural language texts. The mapping from a natural language to a logical representation is realized with a grammatical formalism, linking the syntactic analysis of the text to a semantic representation. In [44], Aurélien Lamercerie targets behavioral specifications of cyber-physical systems, ie any type of system in which software components interact closely with a physical environment. The objective is the simulation and formal verification, by automatic or assisted methods, of system level requirements expressed in a controled fragment of a natural language.

PACAP Project-Team

7. New Results

7.1. Compilation and Optimization

Participants: Arif Ali Ana-Pparakkal, Loïc Besnard, Rabab Bouziane, Sylvain Collange, Byron Hawkins, Imane Lasri, Kévin Le Bon, Erven Rohou.

7.1.1. Optimization in the Presence of NVRAM

Participants: Rabab Bouziane, Erven Rohou, Bahram Yarahmadi.

Energy-efficiency has become one major challenge in both embedded and high-performance computing. Different approaches have been investigated to solve the challenge, e.g., heterogeneous multicore, system runtime and device-level power management, or deployment of emerging non-volatile memories (NVMs), such as Spin-Transfer Torque RAM (STT-RAM), which inherently have quasi-null leakage. This enables to reduce the static power consumption, which tends to become dominant in modern systems. The usage of NVM in memory hierarchy comes however at the cost of expensive write operations in terms of latency and energy.

7.1.1.1. Silent-Stores

We propose [31] 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.

7.1.1.2. Variable Retention Time

In order to mitigate expensive writes, we leverage the notion of δ-*worst-case execution time* (δ-WCET), which consists of partial WCET estimates [32]. From program analysis, δ-WCETs are determined and used to safely allocate data to NVM memory banks with variable data retention times. The δ-WCET analysis computes the WCET between any two locations in a function code, i.e., between basic blocks or instructions. Our approach is validated on the Mälardalen benchmark suite and significant memory dynamic energy reductions (up to 80 %, and 66 % on average) are reported.

This research is done in collaboration with Abdoulaye Gamatié at LIRMM (Montpellier) within the context the ANR project CONTINUUM. Results are detailed in the PhD thesis document of Rabab Bouziane, defended in December 2018 [20].

7.1.1.3. Efficient checkpointing for intermittently-powered systems

Future internet of things (IoT) ultra low-power micro controllers do not have any battery. Instead they harvest energy from the environment such as solar or radio and store it to a capacitor. However, one of the unique problems with these energy harvesting devices is the unstable energy supply which causes frequent power failures during the execution of the program. As a result, a program may not be able to terminate with one power cycle. A solution to this problem consists in using non-volatile memory (NVM) such as FLASH or Ferroelectric RAM (FRAM) and checkpointing the volatile state of the program to the non-volatile memory regularly. The program can resume its execution when the power is back. However, checkpointing regularly at runtime has overhead, and poses some memory inconsistency problems [47]. By statically analyzing the program binary code, we propose to insert checkpoints in the proper places in the program to decrease the amount of checkpointing overhead at runtime. Also, a compiler can give hints to runtime system about whether to do the checkpoint or not. Concerning the reduction of checkpointing overhead, we are analyzing the binary code of the program for estimating the energy of each section of the program. We rely on Heptane, which is originally designed for estimating worst-case execution time. However, by giving energy cost to each ISA instruction, we can estimate the energy consumption of the sections of the program for processors like MSP430 or Arm-cortex m0+ which are typical low-power embedded processors. With this information, we insert checkpoints into the LLVM IR and also apply optimizations in order to have better performance and energy efficiency at runtime.

This research is done within the context of the project IPL ZEP.

7.1.2. Dynamic Binary Optimization

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

Modern hardware features can boost the performance of an application, but software vendors are often limited to the lowest common denominator to maintain compatibility with the spectrum of processors used by their clients. Given more detailed information about the hardware features, a compiler can generate more efficient code, but even if the exact CPU model is known, manufacturer confidentiality policies leave substantial uncertainty about precise performance characteristics. In addition, the activity of other programs colocated in the same runtime environment can have a dramatic effect on application performance. For example, if a shared CPU cache is being heavily used by other programs, memory access latencies may be orders of magnitude longer than those recorded during an isolated profiling session, and instruction scheduling based on such profiles may lose its anticipated advantages. Program input can also drastically change the efficiency of statically compiled code, yet in many cases is subject to total uncertainty until the moment the input arrives during program execution. We have developed FITTCHOOSER [30] to defer optimization of a program's most processor-intensive functions until execution time. FITTCHOOSER begins by profiling the application to determine the performance characteristics that are in effect for the present execution, then generates a set of candidate variations and dynamically links them in succession to empirically measure which of them performs best. The underlying binary instrumentation framework Padrone allows for selective transformation of the program without otherwise modifying its structure or interfering with the flow of execution, making it possible for FITTCHOOSER to minimize the overhead of its dynamic optimization process. Our experimental evaluation demonstrates up to 19 % speedup on a selection of programs from the SPEC CPU 2006 and PolyBench suites while introducing less than 1 % overhead. The FITTCHOOSER prototype achieves these gains with a minimal repertoire of optimization techniques taken from the static compiler itself, which not only testifies to the effectiveness of dynamic optimization, but also suggests that further gains can be achieved by expanding FITTCHOOSER's repertoire of program transformations to include more diverse and more advanced techniques.

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

Nowadays almost every device has parallel architecture, hence parallelization is almost always desirable. However, parallelizing legacy running programs is very challenging. That is due to the fact that usually source code is not available, and runtime parallelization is challenging. Also, detecting parallelizable code is difficult, due to possible dependencies and different execution paths that are undecidable statically. Therefore, speculation is a typical approach whereby wrongly parallelized code is detected and rolled back at runtime. We proposed [27] utilizing processes to implement speculative parallelization using on-stack replacement, allowing for generally simple and portable design where forking a new process enters the speculative state, and killing a faulty process simply performs the roll back operation. While the cost of such operations are high, the approach is promising for cases where the parallel section is long and dependency issues are rare. Also, our proposed system performs speculative parallelization on binary code at runtime, without the need for source code, restarting the program or special hardware support. Initial experiments show about 2× to 3× speedup for speculative execution over serial, when three fourth of loop iterations are parallelizable. Maximum speculation overhead over pure parallel execution is measured at 5.8 %.

This research was partially done within the context of the project PHC IMHOTEP.

7.1.3. Autotuning

Participants: Loïc Besnard, Imane Lasri, Pierre Le Meur, Erven Rohou.

The ANTAREX project relies on a Domain Specific Language LARA [0] of the Clava environment [0]. This DSL is based on Aspect Oriented Programming concepts to allow applications to enforce extra functional properties such as energy-efficiency and performance and to optimize Quality of Service in an adaptive way. The DSL approach allows the definition of energy-efficiency, performance, and adaptivity strategies as well as their enforcement at runtime through application autotuning and resource and power management [28], [29].

In this context, this year we have integrated in Clava some technologies: the memoization, the precision tuning and the loop splitting compilation.

7.1.3.1. Memoization

The concept of memoization essentially involves saving the results of functions together with their inputs so that when the input repeats, the result is taken from a look-up table. This technique, whose objective is to improve sequential performance, has been implemented for C and C++ languages. The support library of this technology allows in particular flexibility for the table management. This work has been submitted for publication in the Elsevier journal SoftwareX. The support library is available at https://gforge.inria.fr/projects/memoization (registered with APP under number IDDN.FR.001.250029.000.S.P.2018.000.10800)

7.1.3.2. Precision tuning

The developed aspects on the type precision consist in the parametrization of the applications in terms of types. Indeed, error-tolerating applications are increasingly common in the emerging field of real-time HPC. Thus, recent works investigated the use of customized precision in HPC as a way to provide a breakthrough in power and performance. This parametrization allows to test easily and quickly different type representations (such as `double`, `float`, `fixed-point`).

7.1.3.3. Loop splitting

The loop splitting technique takes advantage of long running loops to explore the impact of several optimization sequences at once, thus reducing the number of necessary runs. We rely on a variant of loop peeling which splits a loop into into several loops, with the same body, but a subset of the iteration space. New loops execute consecutive chunks of the original loop. We then apply different optimization sequences on each loop independently. Timers around each chunk observe the performance of each fragment. This technique may be generalized to combine compiler options and different implementations of a function called in a loop. It is useful when, for example, the profiling of the application shows that a function is critical in term of time of execution. In this case, the user must try to find the best implementation of their algorithm.

This research is done within the context of the ANTAREX FET HPC collaborative project. The software is being registered with APP.

7.1.4. Hardware/Software JIT Compiler

Participant: Erven Rohou.

In order to provide dynamic adaptation of the performance/energy trade-off, systems today rely on heterogeneous multi-core architectures (different micro-architectures on a chip). These systems are limited to single-ISA approaches to enable transparent migration between the different cores. To offer more trade-offs, we can integrate statically scheduled micro-architecture and use Dynamic Binary Translation (DBT) for task migration. However, in a system where performance and energy consumption are a prime concern, the translation overhead has to be kept as low as possible. We propose Hybrid-DBT [26], an open-source, hardware accelerated DBT system targeting VLIW cores. Three different hardware accelerators have been designed to speed-up critical steps of the translation process. Experimental study shows that the accelerated steps are two orders of magnitude faster than their software equivalent. The impact on the total execution time of applications and the quality of generated binaries are also measured.

Our proposed DBT framework targets the RISC-V ISA, for which both OoO and in-order implementations exist. Our experimental results [37] show that our approach can lead to best-case performance and energy efficiency when compared against static VLIW configurations.

[0] https://web.fe.up.pt/~specs/projects/lara/doku.php
[0] http://specs.fe.up.pt/tools/clava

This work is part of the PhD of Simon Rokicki [22], co-advised by Erven Rohou.

7.1.5. Qubit allocation for quantum circuit compilers

Participants: Sylvain Collange, Marcos Siraichi, Victor Careil.

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. In collaboration with Vinícius Fernandes dos Santos, Fernando Pereira and Marcos Yukio Siraichi at UFMG, 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 was presented at the Code Generation and Optimization (CGO) conference [40].

7.2. Processor Architecture

Participants: Sylvain Collange, Niloofar Charmchi, Kleovoulos Kalaitzidis, Pierre Michaud, Daniel Rodrigues Carvalho, André Seznec, Anita Tino.

7.2.1. Value prediction

Participants: Kleovoulos Kalaitzidis, André Seznec.

For the 1st Championship on Value Prediction (CVP1), we have explored the performance limits of value prediction for small value predictors (8KB and 32KB) in the context of a processor assuming a large instruction window (256-entry ROB), a perfect branch predictor, fetching 16 instructions per cycle, an unlimited number of functional units, but a large value misprediction penalty with a complete pipeline flush at commit on a value misprediction

Our proposition EVES, for Enhanced VTAGE Enhanced Stride, combines two predictor components which do not use on the result of the last occurrence of the instruction to compute the prediction. We use an enhanced version of the VTAGE predictor [11], E-VTAGE. Second, we propose an enhanced version of the stride predictor, E-Stride. E-Stride computes the prediction from the last committed occurrence of the instruction and the number of speculative inflight occurrences of the instruction in the pipeline. The prediction flowing out from E-Stride or E-VTAGE is used only when its confidence is high. A major contribution of this study is the algorithm to assign confidence to predictions depending on the expected benefit/loss from the prediction.

The EVES predictor won the three tracks of CVP1 [39].

7.2.2. Compressed caches

Participants: Daniel Rodrigues Carvalho, Niloofar Charmchi, André Seznec.

Recent advances in research on compressed caches make them an attractive design point for effective hardware implementation for last-level caches. For instance, the yet another compressed cache (YACC) layout [14] leverages both spatial and compression factor localities to pack compressed contiguous memory blocks from a 4-block super-block in a single cache block location. YACC requires less than 2 % extra storage over a conventional uncompressed cache. Performance of LLC is also highly dependent on its cache block replacement management. This includes allocation and bypass decision on a miss as well as replacement target selection which is guided by priority insertion policy on allocation and priority promotion policy on a hit. YACC uses the same cache layout as a conventional set-associative uncompressed cache Therefore the LLC cache management policies that were introduced during the past decade can be transposed to YACC. However, YACC features super-block tags instead of block tags. For uncompressed block, these super-block tags can be used to monitor the reuse behavior of blocks from the same super-block. We introduce the First

In Then First Use Bypass (FITFUB) allocation policy for YACC. With FITFUB, a missing uncompressed block that belongs to a super-block that is already partially valid in the cache is not stored in the cache on its first use, but only on its first reuse if any. FITFUB can be associated with any priority insertion/promotion policy. YACC+FITFUB with compression turned off, achieves an average 6.5%/8% additional performance over a conventional LLC, for single-core/multi-core workloads, respectively. When compression is enabled, the performance benefits associated with compression and FITFUB are almost additive reaching 12.7%/17%. This leads us to call this design the Synergistic cache layout for Reuse and Compression (SRC). SRC reaches the performance benefit that would be obtained with a 4× larger cache, but with less than 2 % extra storage [34].

7.2.3. The Omnipredictor

Participant: André Seznec.

Modern superscalar processors heavily rely on out-of-order and speculative execution to achieve high performance. The conditional branch predictor, the indirect branch predictor and the memory dependency predictor are among the key structures that enable efficient speculative out-of-order execution. Therefore, processors implement these three predictors as distinct hardware components. In [35] we propose the omnipredictor that predicts conditional branches, memory dependencies and indirect branches at state-of-the-art accuracies without paying the hardware cost of the memory dependency predictor and the indirect jump predictor. We first show that the TAGE prediction scheme based on global branch history can be used to concurrently predict both branch directions and memory dependencies. Thus, we unify these two predictors within a regular TAGE conditional branch predictor whose prediction is interpreted according to the type of the instruction accessing the predictor. Memory dependency prediction is provided at almost no hardware overhead. We further show that the TAGE conditional predictor can be used to accurately predict indirect branches through using TAGE entries as pointers to Branch Target Buffer entries. Indirect target prediction can be blended into the conditional predictor along with memory dependency prediction, forming the omnipredictor.

7.2.4. Branch prediction

Participant: Pierre Michaud.

The branch predictor is the keystone of modern superscalar micro-architectures. The TAGE predictor, introduced by André Seznec and Pierre Michaud in 2006, is the most storage-efficient conditional branch predictor known today [16]. Although TAGE is very accurate, it does not exploit its input information perfectly, as significant prediction accuracy improvements are obtained by complementing TAGE with a perceptron-based *statistical corrector* using the same input information [18]. The statistical corrector, even small, makes the whole predictor more complex. We proposed an alternative TAGE-like predictor, called BATAGE, making statistical correction superfluous. BATAGE has the same global structure as TAGE but uses a different tagged-entry format and different prediction and update algorithms. The main reason for TAGE needing statistical correction is the *cold-counter* problem, that is, the fact that recently created tagged entries contain little branch history. To solve the cold-counter problem, we replaced the up-down counter in the tagged entry with two counters counting the *taken* and *not-taken* occurrences separately, and we introduced Bayesian confidence estimation based on Laplace's rule of succession. We also introduced a method called *Controlled Allocation Throttling* for adjusting the rate of creation of tagged entries dynamically. The resulting predictor, BATAGE, obviates the need for external statistical correction [25].

7.2.5. 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 $\times$ 4-lane 4-issue DITVA architecture with a realistic bank-interleaved cache achieves 1.55$\times$ 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 published in the Journal of Parallel and Distributed Computing [6].

7.2.6. *Toward out-of-order SIMT micro-architecture*

Participants: Sylvain Collange, Anita Tino.

Prior work highlights the continued importance of maintaining adequate sequential performance within throughput-oriented cores [49]. 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.

7.3. WCET estimation and optimization

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

7.3.1. *WCET estimation for many core processors*

Participants: Imen Fassi, Damien Hardy, Viet Anh Nguyen, Isabelle Puaut, Benjamin Rouxel, Stefanos Skalistis.

7.3.1.1. 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 is 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 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. An implementation on the Kalray MPPA machine allowed to identify implementation-related overheads. All results are described in the PhD thesis document of Viet Anh Nguyen [21], defended in February 2018.

This research is part of the PIA Capacités project.

7.3.1.2. Shared resource contentions and WCET estimation

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.

The work along this direction is part of the PhD thesis of Benjamin Rouxel, defended in December 2018. The new results in 2018 concern scheduling/mapping of parallel applications on multi-core systems using ScratchPad Memories (SPMs). We have recently proposed techniques that jointly select SPM contents off-line, in such a way that the cost of SPM loading/unloading is hidden. Communications are fragmented to augment hiding possibilities. Experimental results show the effectiveness of the proposed techniques on streaming applications and synthetic task-graphs. The overlapping of communications with computations allows the length of generated schedules to be reduced by 4 % in average on streaming applications, and by 8 % in average (with maximum of 16 % for both test cases) for synthetic task graphs. We further show on a case study that generated schedules can be implemented with low overhead on a predictable multi-core architecture (Kalray MPPA) [23].

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

Parallel architectures are nowadays not only 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. The Argo toolchain operates on Scilab and XCoS inputs, and targets ScratchPad Memory (SPM)-based multi-cores. Data-layout and loop transformations play a key role in this flow as they improve SPM efficiency and reduce the number of accesses to shared main memory.

In our most recent work [33], we study how these transformations impact WCET estimates of sequential codes. We demonstrate that they can bring significant improvements of WCET estimates (up to 2.7$\times$) provided that the WCET analysis process is guided with automatically generated flow annotations obtained using polyhedral counting techniques.

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

7.3.2. WCET estimation and optimizing compilers

Participants: Imen Fassi, Isabelle Puaut.

Compiler optimizations, although reducing the execution times of programs, raise issues in static WCET estimation techniques and tools. Flow facts, such as loop bounds, may not be automatically found by static WCET analysis tools after aggressive code optimizations. In this work [36], we explore the use of iterative compilation (WCET-directed program optimization to explore the optimization space), with the objective to (i) allow flow facts to be automatically found and (ii) select optimizations that result in the lowest WCET estimates. We also explore to which extent code outlining helps, by allowing the selection of different optimization options for different code snippets of the application.

7.3.3. Partial WCET

Participants: Rabab Bouziane, Erven Rohou.

Computing the worst-case execution time (WCET) of tasks is important for real-time system design. The industry and research communities have developed a wealth of techniques to compute relevant WCET approximations. Traditionally, WCETs are estimated at the granularity of a function (or task). We propose an approach to estimate partial WCET (δ-WCET), i.e., the worst-case execution time between two locations in a function, such as basic blocks or instructions. Our technique [41] is derived from the well-known implicit path enumeration technique. It takes into account both the control flow graph and the architecture (pipeline and cache hierarchy). Some useful applications of such δ-WCETs are motivated in this paper.

This research is part of the ANR Continuum project.

7.4. Security

Participants: Damien Hardy, Byron Hawkins, Nicolas Kiss, Kévin Le Bon, Erven Rohou.

7.4.1. Compiler-based automation of side-channel countermeasures

Masking is a popular protection against side-channel analysis exploiting the power consumption or electromagnetic radiations. Besides the many schemes based on simple Boolean encoding, some alternative schemes such as Orthogonal Direct Sum Masking (ODSM) or Inner Product Masking (IP) aim to provide more security, reduce the entropy or combine masking with fault detection. The practical implementation of those schemes is done manually at assembly or source-code level, some of them even stay purely theoretical. We propose a compiler extension to automatically apply different masking schemes for block cipher algorithms. We introduce a generic approach to describe the schemes and we manage to insert three of them at compile-time on an AES implementation. A practical side-channel analysis as well as fault injections have been performed on an Arm microcontroller to assess the correctness of the code inserted.

The resulting compiler plugin (sigmask) is registered with APP under number IDDN.FR.001.490003.000.S.P. 2018.000.10000

This research was done within the context of the project ANR CHIST-ERA SECODE.

7.4.2. Program protection through dynamic binary rewriting

Programs written in languages such as C and C++ are prone to memory corruptions because of the manual management of the memory from the programmer. Even today, memory corruptions are among the most dangerous vulnerabilites. According to the MITRE ranking, these bugs are considered one of the top three most dangerous software vulnerabilites.

Thanks to our library Padrone, we are able to instrument the execution of a program with a minimal overhead, making it possible to move or add code in the target process during its execution. We showed that we can change the address of a function at runtime, thus presenting a moving target to an attacker, and making attacks more difficult.

Many security policies have been developed to protect programs. One of them, the Control-Flow Integrity (CFI) ensures the control-flow of the program cannot be altered, preventing the execution of malicious code. Unfortunately, implementations of precise CFI impose a consequent overhead in performance, due to the instrumentation of the execution of the program. We work on building a solution that is able to adapt its protection level to the situation. Adapting the protection level allows us to reduce even further the overhead in performance when the protection is not needed.

SUMO Project-Team

7. New Results

7.1. Analysis and Verification of Quantitative Systems

7.1.1. Verification of Concurrent Timed Systems

Participants : Éric Fabre, Loïc Hélouët, Karim Kecir

7.1.1.1. Combining Free Choice and Time in Petri Nets

Time Petri nets (TPNs) are a classical extension of Petri nets with timing constraints attached to transitions, for which most verification problems are undecidable. In [3], We consider TPNs under a strong semantics with multiple enablings of transitions. We focus on a structural subclass of unbounded TPNs, where the underlying untimed net is free choice, and show that it enjoys nice properties in the timed setting under a multi-enabling semantics. In particular, we show 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 consider 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-enabling semantics, we show decidability of robustness of firability and of termination under both guard enlargement and shrinking.

7.1.1.2. Production Systems with Concurrent Tasks

The work in [7] considers the realizability of expected schedules by production systems with concurrent tasks, bounded resources that have to be shared among tasks, and random behaviors and durations. Schedules are high level views of desired executions of systems represented as partial orders decorated with timing constraints. Production systems (production cells, train networks...) are modeled as stochastic time Petri nets STPNs with an elementary (1-bounded) semantics. We first propose a notion of time processes to give a partial order semantics to STPNs. We then consider 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 hence consider probabilistic realizability up to a time units, that holds if the probability that N realizes S with constraints enlarged by a 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.

7.1.2. Testing of Timed Systems

Participants : Léo Henry, Thierry Jéron, Nicolas Markey

Partial observability and controllability are two well-known issues in test-case synthesis for interactive systems. In [25], we address the problem of partial control in the synthesis of test cases from timed-automata specifications. Building on the tioco timed testing framework, we extend a previous game interpretation of the test-synthesis problem from the untimed to the timed setting. This extension requires a deep reworking of the models, game interpretation and test-synthesis algorithms. We exhibit strategies of a game that tries to minimize both control losses and distance to the satisfaction of a test purpose, and prove they are winning under some fairness assumptions. This entails that when turning those strategies into test cases, we get properties such as soundness and exhaustiveness of the test synthesis method.

7.1.3. Analysis of Stochastic Systems

Participants : Nathalie Bertrand

A decade ago, Abdulla, Ben Henda and Mayr introduced the elegant concept of decisiveness for denumerable Markov chains. Roughly speaking, decisiveness allows one to lift most good properties from finite Markov chains to denumerable ones, and therefore to adapt existing verification algorithms to infinite-state models. Decisive Markov chains however do not encompass stochastic real-time systems, and general stochastic transition systems (STSs for short) are needed. In [4], we provide a framework to perform both the qualitative and the quantitative analysis of STSs. First, we define various notions of decisiveness, notions of fairness and of attractors for STSs, and make explicit the relationships between them. Then, we define a notion of abstraction, together with natural concepts of soundness and completeness, and we give general transfer properties, which will be central to several verification algorithms on STSs. We further design a generic construction which will be useful for the analysis of ω-regular properties, when a finite attractor exists, either in the system (if it is denumerable), or in a sound denumerable abstraction of the system. We next provide algorithms for qualitative model-checking, and generic approximation procedures for quantitative model-checking. Finally, we instantiate our framework with stochastic timed automata (STA), generalized semi-Markov processes (GSMPs) and stochastic time Petri nets (STPNs), three models combining dense-time and probabilities. This allows us to derive decidability and approximability results for the verification of these models. Some of these results were known from the literature, but our generic approach permits to view them in a unified framework, and to obtain them with less effort. We also derive interesting new approximability results for STA, GSMPs and STPNs.

7.1.4. Opacity for Quantitative Systems

Participants : Loïc Hélouët, Hervé Marchand

7.1.4.1. Quantitative Opacity

The work in [26] considers quantitative approaches for opacity. A system satisfies opacity if its secret behaviors cannot be detected by any user of the system. Opacity of distributed systems was originally set as a boolean predicate before being quantified as measures in a probabilistic setting. This paper considers a different quantitative approach that measures the efforts that a malicious user has to make to detect a secret. This effort is measured as a distance w.r.t a regular profile specifying a normal behavior. This leads to several notions of quantitative opacity. When attackers are passive that is, when they just observe the system, quantitative opacity is brought back to a language inclusion problem, and is PSPACE-complete. When attackers are active, that is, interact with the system in order to detect secret behaviors within a finite depth observation, quantitative opacity turns out to be a two-player finite-state quantitative game of partial observation. A winning strategy for an attacker is a sequence of interactions with the system leading to a secret detection without exceeding some profile deviation measure threshold. In this active setting, the complexity of opacity is EXPTIME-complete.

7.1.4.2. Opacity with Powerful Attackers

In [27], we consider state-based opacity in a setting where attackers of a secret have additional observation capabilities allowing them to know which inputs are allowed by a system. This capability allows attackers of a system to partially disambiguate the possible set of states the system might be in, and increases the power of an attacker. We show that regular opacity (opacity of a property described by a regular language) is decidable in this setting. We then address the question of controlling a system so that it becomes opaque, and solve this question by recasting the problem in a game setting.

7.1.5. Diagnosis of Quantitative Systems

Participants : Blaise Genest, Éric Fabre, Hugo Bazille, Nicolas Markey

7.1.5.1. Diagnosis for Timed Automata

In [20], we consider the problems of efficiently diagnosing and predicting what did (or will) happen in a partially-observable one-clock timed automaton. We introduce timed sets as a formalism to keep track of the evolution of the reachable configurations over time, and build a candidate diagnoser for our timed automaton. We report on our implementation of this approach compared to the algorithm of Tripakis, *Fault diagnosis for timed automata*, 2002.

7.1.5.2. Quantitative Diagnosis for 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 in [17]. We are interested in quantifying how fast the diagnosability can be performed. For that, we give an algorithm to compute in polynomial time any moment of the distribution of the detection delay. This allows one to approximate the distribution of detection delay, and to provide lower bounds on the probability that detection takes place at most T events after the fault.

One problem with A-diagnosability is that in the worst case, a subset construction needs to be performed, leading to an exponential blow-up in the number of states. To mitigate this, we proposed in [16] different techniques that avoid this blow-up in a large number of cases.

7.2. Control of Quantitative Systems

7.2.1. Reactive Synthesis for Quantitative Systems

Participants : Hervé Marchand, Nicolas Markey

7.2.1.1. Optimal and Robust Controller Synthesis

We propose a novel framework for the synthesis of robust and optimal energy-aware controllers. The framework is based on energy timed automata, allowing for easy expression of timing-constraints and variable energy-rates. We prove decidability of the energy-constrained infinite-run problem in settings with both certainty and uncertainty of the energy-rates. We also consider the optimization problem of identifying the minimal upper bound that will permit existence of energy-constrained infinite runs. Our algorithms are based on quantifier elimination for linear real arithmetic. Using Mathematica and Mjollnir, we illustrate our framework through a real industrial example of a hydraulic oil pump. Compared with previous approaches our method is completely automated and provides improved results.

7.2.1.2. Average-Energy Games

Two-player quantitative zero-sum games provide a natural framework to synthesize controllers with performance guarantees for reactive systems within an uncontrollable environment. Classical settings include mean-payoff games, where the objective is to optimize the long-run average gain per action, and energy games, where the system has to avoid running out of energy. In [5], we study average-energy games, where the goal is to optimize the long-run average of the accumulated energy. We show that this objective arises naturally in several applications, and that it yields interesting connections with previous concepts in the literature. We prove that deciding the winner in such games is in NP∩coNP and at least as hard as solving mean-payoff games, and we establish that memoryless strategies suffice to win. We also consider the case where the system has to minimize the average-energy while maintaining the accumulated energy within predefined bounds at all times: this corresponds to operating with a finite-capacity storage for energy. We give results for one-player and two-player games, and establish complexity bounds and memory requirements.

7.2.1.3. Compositional Controller Synthesis

In [8], we present a correct-by-design method of state-dependent control synthesis for sampled switching systems. Given a target region R of the state space, our method builds a capture set S and a control that steers any element of S into R. The method works by iterated backward reachability from R. It is also used to synthesize a recurrence control that makes any state of R return to R infinitely often. We explain how the synthesis method can be performed in a compositional manner, and apply it to the synthesis of a compositional control for a concrete floor-heating system with 11 rooms and up to $2^{11} = 2048$ switching modes.

7.2.1.4. Symbolic Algorithms for Control

In [18], we put forward a new modeling technique for Dynamic Resource Management (DRM) based on discrete events control for symbolic logico-numerical systems, especially Discrete Controller Synthesis (DCS). The resulting models involve state and input variables defined on an infinite domain (Integers), thereby no exact DCS algorithm exists for safety control. We thus formally define the notion of limited lookahead, and associated best-effort control objectives targeting safety and optimization on a sliding window for a number of steps ahead. We give symbolic algorithms, illustrate our approach on an example model for DRM, and report on performance results based on an implementation in our tool ReaX.

7.2.2. Control of Stochastic Systems

Participants : Nathalie Bertrand, Blaise Genest, Nicolas Markey, Ocan Sankur

7.2.2.1. Multi-Weighted Markov Decision Processes

In [19], we study the synthesis of schedulers in double-weighted Markov decision processes, which satisfy both a percentile constraint over a weighted reachability condition, and a quantitative constraint on the expected value of a random variable defined using a weighted reachability condition. This problem is inspired by the modelization of an electric-vehicle charging problem. We study the cartography of the problem, when one parameter varies, and show how a partial cartography can be obtained via two sequences of opimization problems. We discuss completeness and feasability of the method.

7.2.2.2. Stochastic Shortest Paths and Weight-Bounded Reachability

The work in [14] deals with finite-state Markov decision processes (MDPs) with integer weights assigned to each state-action pair. New algorithms are presented to classify end components according to their limiting behavior with respect to the accumulated weights. These algorithms are used to provide solutions for two types of fundamental problems for integer-weighted MDPs. First, a polynomial-time algorithm for the classical stochastic shortest path problem is presented, generalizing known results for special classes of weighted MDPs. Second, qualitative probability constraints for weight-bounded (repeated) reachability conditions are addressed. Among others, it is shown that the problem to decide whether a disjunction of weight-bounded reachability conditions holds almost surely under some scheduler belongs to NP$\cap$coNP, is solvable in pseudo-polynomial time and is at least as hard as solving two-player mean-payoff games, while the corresponding problem for universal quantification over schedulers is solvable in polynomial time.

7.2.2.3. Distribution-based Objectives for Markov Decision Processes

In the scope of associated team EQuaVE, we have considered quantitative control of stochastic systems [10]. More precisely, the aim is to control the MDP so that the distribution over states stays inside a safe polytope. This represents a trade off between perfect information (the system is in exactly one state) and no information (we need to consider the belief distribution over states, and further the action played by the controller cannot be based on the state). Interestingly, we get an efficient polynomial time complexity to check whether there exists a distribution from which there exists a controller keeping the MDP in the safe polytope. This is surprising as the same question from a given distribution is not known to be decidable, even if the controller is fixed. Also, we have a co-NP complexity for deciding whether for every initial distribution, there is controller keeping the distribution in the safe polytope. Finally, we showed that an alternate representation of the input polytope allows us to get a polynomial time algorithm for safety from all initial distributions.

7.3. Management of Large Distributed Systems

7.3.1. Parameterized Systems

Participants : Nathalie Bertrand, Nicolas Markey

Reconfigurable broadcast networks provide a convenient formalism for modelling and reasoning about networks of mobile agents broadcasting messages to other agents following some (evolving) communication topology. The parameterized verification of such models aims at checking whether a given property holds irrespective of the initial configuration (number of agents, initial states and initial communication topology). In [15], we focus on the synchronization property, asking whether all agents converge to a set of target states after some execution. This problem is known to be decidable in polynomial time when no constraints are imposed on the evolution of the communication topology (while it is undecidable for static broadcast networks).

During the internship of A.R. Balasubramanian, we investigated how various constraints on reconfigurations affect the decidability and complexity of the synchronization problem. In particular, we show that when bounding the number of reconfigured links between two communications steps by a constant, synchronization becomes undecidable; on the other hand, synchronization remains decidable in PTIME when the bound grows with the number of agents.

7.3.2. Smart Regulation for Urban Trains

Participants : Loïc Hélouët, Karim Kecir, Flavia Palmieri

We have launched a new thread of research for efficient regulation with the M2 internship of Flavia Palmieri. The objective is to use efficient planning techniques to perform regulation in metro networks. Usually, regulation algorithms are simple reactive rules, that build decisions from local measures of train delays. These algorithms are arbitrary decisions, which efficiency is only empirically proved. On the other hand, optimality of regulation decision with respect to some quality criterion could be achieved through optimization algorithms, associating an optimal execution date to next events (arrivals and departures) while fulfilling constraints on causal dependencies, track allocations, etc. However, these algorithms are NP-complete, and do not return answers fast enough to be used online as regulation tools (use usually expects a decision within a few seconds after a train's arrival). During this internship, we have started integrating optimal planning techniques to regulation schemes. The main idea is to perform optimization online for a subset of the next occurring events. Performance of this regulation scheme is currently under evaluation.

7.3.3. Analysis of Concurrent Systems

Participants : Éric Fabre, Loïc Hélouët, Engel Lefaucheux

7.3.3.1. Generalization of Unfolding Techniques for Petri Nets

The verification of concurrent systems relies on an adequate representation of their trajectory sets, where each trajectory is a partial order of events. Several compact structures have been proposed in the past, starting with unfoldings and event structures. While unfoldings expand both time and conflicts, they generate extremely large branching constructions. To avoid expanding conflicts where they are not meaningful, more compact structures were proposed, as merged processes and trellis processes. In [23], we examine structures that would not fully unfold time as well, thus resulting in partially unfolded nets. To do so, we proposed the notion of spread nets, (safe) Petri nets equipped with vector clocks on places and with ticking functions on transitions, and such that vector clocks are consistent with the ticking of transitions. Such nets allow one to generalize previous constructions as unfoldings and merged processes, and can be fully paremeterized to display or hide some behaviors of the net, and thus facilitate its analysis.

7.3.3.2. Hyper Partial Order Logic

In [21], we define HyPOL, a local hyper logic for partial order models, expressing properties of sets of runs. These properties depict shapes of causal dependencies in sets of partially ordered executions, with similarity relations defined as isomorphisms of past observations. This type of logics is tailored to address security properties of concurrent systems. Unsurprisingly, since comparison of projections are included, satisfiability of this logic is undecidable. We then address model checking of HyPOL and show that, already for safe Petri nets, the problem is undecidable. Fortunately, sensible restrictions of observations and nets allow us to bring back model checking of HyPOL to a decidable problem, namely model checking of MSO on graphs of bounded treewidth.

7.3.3.3. Diagnosability Analysis for Concurrent Systems

Petri nets have been proposed as a fundamental model for discrete-event systems in a wide variety of applications and have been an asset to reduce the computational complexity involved in solving a series of problems, such as control, state estimation, fault diagnosis, etc. Many of those problems require an analysis of the reachability graph of the Petri net. The basis reachability graph is a condensed version of the reachability graph that was introduced to efficiently solve problems linked to partial observation. It was in particular used for diagnosis which consists in deciding whether some fault events occurred or not in the system, given partial observations on the run of the system. However this method is, with very specific exceptions, limited to bounded Petri nets. In [28], we introduce the notion of basis coverability graph to remove this requirement. We then establish the relationship between the coverability graph and the basis coverability graph. Finally, we focus on the diagnosability problem: we show how the basis coverability graph can be used to get an efficient algorithm.

7.4. Data Driven Systems

7.4.1. Modular composition of Guarded Attribute Grammars

Participants : Éric Badouel

We investigate how the role of a user in a distributed collaborative systems modelled by a Guarded Attribute Grammar can be associated with a domain specific language (DSL) encapsulating a specific domain knowledge (expertise) and defining a set of services (a language-oriented approach). These DSLs communicate through service calls (a service-oriented approach).

Language oriented programming is an approach to software composition based on domain specific languages (DSL) dedicated to specific aspects of an application domain. In order to combine such languages we embed them into a host language (namely Haskell, a strongly typed higher-order lazy functional language). A DSL is then given by an algebraic type, whose operators are the constructors of abstract syntax trees. Such a multi-sorted signature is associated to a polynomial functor. An algebra for this functor tells us how to interpret the programs. Using Bekić's Theorem we defined in [13] a modular decomposition of algebras that leads to a class of parametric multi-sorted signatures, associated with regular functors, allowing for the modular design of DSLs.

In [12] we have addressed the problem of component reuse in the context of service-oriented programming and more specifically for the design of user-centric distributed collaborative systems modelled by Guarded Attribute Grammars. Following the contract-based specification of components we developp an approach to an interface theory for the roles in a collaborative system in three stages: we define a composition of interfaces that specifies how the component behaves with respect to its environement, we introduce an implementation order on interfaces and finally a residual operation on interfaces characterizing the systems that, when composed with a given component, can complement it in order to realize a global specification.

TAMIS Project-Team

7. New Results

7.1. Results for Axis 1: Vulnerability analysis

7.1.1. Statistical Model Checking of Incomplete Stochastic Systems

Participants: Tania Richmond, Louis-Marie Traonouez, Axel Legay.

We proposed a statistical analysis of stochastic systems with incomplete information. These incomplete systems are modelled using discrete time Markov chains with unknowns (qDTMC), and the required behaviour was formalized using qBLTL logic. By doing both quantitative and qualitative analysis of such systems using statistical model checking, we also proposed refinement on the qDTMCs. These refined qDTMCs depict a decrease in the probability of unknown behaviour in the system. The algorithms for both qualitative and quantitative analysis of qDTMC were implemented in the tool Plasma Lab. We demonstrated the working of these algorithms on a case study of a network with unknown information. We plan to extend this work to analyse the behaviour of other stochastic models like Markov decision processes and abstract Markov chains, with incomplete information.

This work has been accepted and presented to a conference this year [10].

[10] We study incomplete stochastic systems that are missing some parts of their design, or are lacking information about some components. It is interesting to get early analysis results of the requirements of these systems, in order to adequately refine their design. In previous works, models for incomplete systems are analysed using model checking techniques for three-valued temporal logics. In this paper, we propose statistical model checking algorithms for these logics. We illustrate our approach on a case-study of a network system that is refined after the analysis of early designs.

7.1.2. A Language for Analyzing Security of IOT Systems

Participants: Delphine Beaulaton, Najah Ben Said, Ioana Cristescu, Axel Legay, Jean Quilbeuf.

We propose a model-based security language of Internet of Things (IoT) systems that enables users to create models of their IoT systems and to make analysis of the likelihoods of cyber-attacks to occur and succeed. The modeling language describes the interactions between different entities, that can either be humans or “Things” (i.e, hardware, sensors, software tools, ..). A malicious entity is present in the system, called the Attacker, and it carries out attacks against the system. The other IoT entities can inadvertently help the Attacker, by leaking their sensitive data. Equipped with the acquired knowledge the Attacker can then communicate with the IoT entities undetected. For instance, an attacker can launch a phishing attack via email, only if it knows the email address of the target.

Another feature of our modeling language is that security failures are modeled as a sequence of simpler steps, in the spirit of *attack trees*. As their name suggests, attacks are modeled as trees, where the leaves represent elementary steps needed for the attack, and the root represents a successful attack. The internal nodes are of two types, indicating whether all the sub-goals (an AND node) or one of the sub-goals (an OR node) must be achieved in order to accomplish the main goal. The attack tree provided with the IoT system acts as a monitor: It observes the interactions the Attacker has with the system and detects when an attack is successful.

An IoT system is analyzed using statistical model checking (SMC). The first method we use is Monte Carlo, which consists of sampling the executions of an IoT system and computing the probability of a successful attack based on the number of executions for which the attack was successful. However, the evaluation may be difficult if a successful attack is *rare*. We therefore also use a second SMC method, developed for *rare events*, called *importance splitting*.

To implement this we rely on *BIP*, a heterogeneous component-based model for which an execution engine is developed and maintained. The IoT model is translated into a BIP model and the attack tree into a BIP monitor. The two form a BIP system. The execution engine of BIP produce executions which are the input of Plasma Lab, the model checker developped in TAMIS. We have extended Plasma Lab with a plugin that interacts with the BIP execution engine.

The tools are available at http://iot-modeling.gforge.inria.fr/. This work has been published in two conference papers [20], [23]. A third paper was submitted in November [29], and is currently under review.

[20] In this paper we propose our security-based modeling language for IoT systems. The modeling language has two important features: (i) vulnerabilities are explicitly represented and (ii) interactions are allowed or denied based on the information stored on the IoT devices. An IoT system is transformed in BIP, a component-based modeling language, in which can execute the system and perform security analysis. To illustrate the features of our language, we model a use-case based on a Smart Hospital and inspired by industrial scenarios.

[23] In this paper we revisit the security-based modeling language for IoT systems. We focus here on the BIP models obtained from the original IoT systems. The BIP execution and analysis framework provides several methods to analyse a BIP model, and we discuss how these methods can be lifted on the original IoT systems. We also model a new use-case based on Amazon Smart Home.

[29] Attack trees are graphical representations of the different scenarios that can lead to a security failure. In this paper we extend our security-based framework for modeling IoT systems in two ways: (i) attack trees are defined alongside the model to detect and prevent security risks in the system and (ii) the language supports probabilistic models. A successful attack can be a *rare event* in the execution of a well designed system. When rare, such attacks are hard to detect with usual model checking techniques. Hence, we use *importance splitting* as a statistical model checking technique for rare events.

7.1.3. Verification of IKEv2 protocol

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

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. IKEv2 guarantees security properties like mutual-authentication and secrecy of exchanged key. To obtain an IKEv2 implementation as secure as possible, we use model checking to verify the properties on the protocol specification, and software formal verification tools to detect implementation flaws like buffer overflows or memory leaks.

In previous analyses, IKEv2 has been shown to possess two authentication vulnerabilities that were considered not exploitable. We analyze the protocol specification using the Spin model checker, and prove that in fact the first vulnerability does not exist. In addition, we show that the second vulnerability is exploitable by designing and implementing a novel slow Denial-of-Service attack, which we name the Deviation Attack.

We propose an expression of the time at which Denial-of-Service happens, and validate it through experiment on the strongSwan implementation of IKEv2. As a counter-measure, we propose a modification of IKEv2, and use model checking to prove that the modified version is secure.

For ethical reasons we informed our country's national security agency (ANSSI) about the existence of the Deviation Attack. The security agency gave us some technical feedback as well as its approval for publishing the attack.

We then tackle formal verification directly applied to an IKEv2 source code. We already tried to analyze strongSwan using the Angr tool. However we found that the Angr was not mature yet for a program like strongSwan. We thus try other software formal verification tools and apply them to smaller and simpler source code than strongSwan: we analyze OpenSSL asn1parse using the CBMC tool and light-weight IP using the Infer tool. We find that CBMC does not scale to a large source code and that Infer does not verify the properties we want.

We plan to explore more in-depth a formal technique and work towards the goal of verifying generic properties (absence of implementation flaws) on softwares like strongSwan.

7.1.4. Combining Software-based and Hardware-based Fault Injection Approaches

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

Software-based and hardware-based approaches have both been used to detect fault injection vulnerabilities. Software-based approaches can provide broad and rapid coverage as it was shown in the previous publications [36], [37], [38] , but may not correlate with genuine hardware vulnerabilities. Hardware-based approaches are indisputable in their results, but rely upon expensive expert knowledge and manual testing.

This work bridges software-based and hardware-based fault injection vulnerability detection by contrasting results of both approaches. To our knowledge no research where done trying to bridge the software-based and hardware-based approach to detect fault injection vulnerabilities the way it is done in this work.

Using both the software-based and hardware-based approaches showed that:

- Software-based approaches detect genuine fault injection vulnerabilities.
- Software-based approaches yield false-positive results.
- Software-based approaches did *not* yield false-negative results.
- Not all software-based vulnerabilities can be reproduced in hardware.
- Hardware-based EMP approaches do *not* have a simple fault model.
- There is a coincidence between software-based and hardware-based approaches.
- Combining software-based and hardware-based approaches yields a vastly more efficient method to detect genuine fault injection vulnerabilities.

This work implemented both the SimFI tool and the ArmL tool.

7.1.5. Side-channel analysis on post-quantum cryptography

Participants: Annelie Heuser, Tania Richmond.

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.

We started by a survey on cryptanalysis against code-based cryptography [13], that includes algebraic and side-channel attacks. Code-based cryptography reveals sensitive data mainly in the syndrome decoding. We investigate the syndrome computation from a side-channel point of view. There are different methods that can be used depending on the underlying code. We explore vulnerabilities of each one in order to propose a guideline for designers and developers. This work was presented at CryptArchi 2018 and Journées Codes et Cryptographie 2018.

[13] Nowadays public-key cryptography is based on number theory problems, such as computing the discrete logarithm on an elliptic curve or factoring big integers. Even though these problems are considered difficult to solve with the help of a classic computer, they can be solved in polynomial time on a quantum computer. Which is why the research community proposed alternative solutions that are quantum resistant. The process of finding adequate post-quantum cryptographic schemes has moved to the next level, right after NIST's announcement for post-quantum standardization.

One of the oldest quantum resistant proposition goes back to McEliece in 1978, who proposed a public-key cryptosystem based on coding theory. It benefits of really efficient algorithms as well as strong mathematical backgrounds. Nonetheless, its security has been challenged many times and several variants were cryptanalyzed. However, some versions are still unbroken.

In this paper, we propose to give a short background on coding theory in order to present some of the main flawless in the protocols. We analyze the existing side-channel attacks and give some recommendations on how to securely implement the most suitable variants. We also detail some structural attacks and potential drawback for new variants.

7.1.6. New Advances on Side-channel Distinguishers

Participants: Christophe Genevey Metat, Annelie Heuser, Tania Richmond.

[17] *On the Performance of Deep Learning for Side-channel Analysis* We answer the question whether convolutional neural networks are more suitable for SCA scenarios than some other machine learning techniques, and if yes, in what situations. Our results point that convolutional neural networks indeed outperforms machine learning in several scenarios when considering accuracy. Still, often there is no compelling reason to use such a complex technique. In fact, if comparing techniques without extra steps like preprocessing, we see an obvious advantage for convolutional neural networks only when the level of noise is small, and the number of measurements and features is high. The other tested settings show that simpler machine learning techniques, for a significantly lower computational cost, perform similar or even better. The experiments with the guessing entropy metric indicate that simpler methods like Random forest or XGBoost perform better than convolutional neural networks for the datasets we investigated. Finally, we conduct a small experiment that opens the question whether convolutional neural networks are actually the best choice in side-channel analysis context since there seems to be no advantage in preserving the topology of measurements.

[8] *The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations* We concentrate on machine learning techniques used for profiled side-channel analysis in the presence of imbalanced data. Such scenarios are realistic and often occurring, for instance in the Hamming weight or Hamming distance leakage models. In order to deal with the imbalanced data, we use various balancing techniques and we show that most of them help in mounting successful attacks when the data is highly imbalanced. Especially, the results with the SMOTE technique are encouraging, since we observe some scenarios where it reduces the number of necessary measurements more than 8 times. Next, we provide extensive results on comparison of machine learning and side-channel metrics, where we show that machine learning metrics (and especially accuracy as the most often used one) can be extremely deceptive. This finding opens a need to revisit the previous works and their results in order to properly assess the performance of machine learning in side-channel analysis.

[35] *When Theory Meets Practice: A Framework for Robust Profiled Side-channel Analysis* Profiled side-channel attacks are the most powerful attacks and they consist of two steps. The adversary first builds a leakage model, using a device similar to the target one, then it exploits this leakage model to extract the secret information from the victim's device. These attacks can be seen as a classification problem, where the adversary needs to decide to what class (corresponding to the secret key) the traces collected from the victim's devices belong to. For a number of years, the research community studied profiled attacks and proposed numerous improvements. Despite a large number of empirical works, a framework with strong theoretical foundations to address profiled side-channel attacks is still missing.

In this paper, we propose a framework capable of modeling and evaluating all profiled analysis attacks. This framework is based on the expectation estimation problem that has strong theoretical foundations. Next, we quantify the effects of perturbations injected at different points in our framework through robustness analysis where the perturbations represent sources of uncertainty associated with measurements, non-optimal classifiers, and methods. Finally, we experimentally validate our framework using publicly available traces, different classifiers, and performance metrics.

[33] *Make Some Noise: Unleashing the Power of Convolutional Neural Networks for Profiled Side-channel Analysis* Profiled side-channel attacks based on deep learning, and more precisely Convolutional Neural Networks, is a paradigm showing significant potential. The results, although scarce for now, suggest that such techniques are even able to break cryptographic implementations protected with countermeasures. In this paper, we start by proposing a new Convolutional Neural Network instance that is able to reach high performance for a number of considered datasets. Additionally, for a dataset protected with the random delay countermeasure, our neural network is able to break the implementation by using only 2 traces in the attack phase. We compare our neural network with the one designed for a particular dataset with masking countermeasure and we show how both are good designs but also how neither can be considered as a superior to the other one. Next, we address how the addition of artificial noise to the input signal can be actually beneficial to the performance of the neural network. Such noise addition is equivalent to the regularization term in the objective function. By using this technique, we are able to improve the number of measurement needed to reveal the secret key by orders of magnitude in certain scenarios for both neural networks. To strengthen our experimental results, we experiment with a number of datasets which differ in the levels of noise (and type of countermeasure) where we show the viability of our approaches.

[9] *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.

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: Olivier Decourbe, Annelie Heuser, Jean-Louis Lanet, Olivier Zendra, Cassius Puodzius, Stefano Sebastio, Lamine Nourredine, Jean Quilbeuf, Eduard Baranov, Thomas Given-Wilson, Fabrizio Biondi, Axel Legay, Alexander Zhdanov.

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: Olivier Decourbe, Lamine Nourredine, Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Jean Quilbeuf, Axel Legay, Fabrizio Biondi.

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.

- **1. Packing Detection and Classification**

  Packing is a widespread tool to prevent static malware detection and analysis. Detecting and classifying the packer used by a given malware sample is fundamental to being able to unpack and study the malware, whether manually or automatically. Existing works on packing detection and classification has focused on effectiveness, but does not consider the efficiency required to be part of a practical malware-analysis workflow. This work studies how to train packing detection and classification algorithms based on machine learning to be both highly effective and efficient. Initially, we create ground truths by labeling more than 280,000 samples with three different techniques. Then we perform feature selection considering the contribution and computation cost of features. Then we iterate over more than 1,500 combinations of features, scenarios, and algorithms to determine which algorithms are the most effective and efficient, finding that a reduction of 1-2% effectiveness can increase efficiency by 17-44 times. Then, we test how the best algorithms perform against malware collected after the training data to assess them against new packing techniques and versions, finding a large impact of the ground truth used on algorithm robustness. Finally, we perform an economic analysis and find simple algorithms with small feature sets to be more economical than complex algorithms with large feature sets based on uptime/training time ratio.

- **2. Packing clustering** A limit of supervised learning is to not be able to recognize classes that were not present in the ground truth. In the work's case above, this means that packer families for which a classifier has not been trained will not be recognized. In this work, we use unsupervised learning techniques, more particularly clustering, in order to provide information about packed malware with previously unknown packing techniques. Here, we build our own dataset of packed binaries, since in the previous work, it has been shown that the construction of the ground truth was fundamental in determining the effectiveness of the packing classification process. Choosing the right clustering algorithm with the right distance metric, dealing with different scales of features units, while being effective, efficient and robust are also majors parts of the current work.

This work is still in progress ...

- **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 and clustering

Participants: Annelie Heuser, Nisrine Jafri, Jean-Louis Lanet, Cassius Puodzius, Stefano Sebastio, Olivier Decourbe, Eduard Baranov, Jean Quilbeuf, Thomas Given-Wilson, Axel Legay, Fabrizio Biondi.

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.

Moreover, we explored the impact on the malware classification of several heuristics adoptable in the SCDGs building process and graph exploration. In particular, our purpose was to:

- identify quality characteristics and evaluation metrics of binary signatures based on SCDGs (and consequently the key properties of the execution traces), that characterize signatures able to provide high-precision malware classification
- optimize the performance of the SMT solver by designing a meta-heuristic able to select the best heuristic to tackle a specific sub-class of problem, study the impact of the configuration of the SMT solver and symbolic execution framework, and understand their interdependencies with the aim of efficiently extracting SCDGs in accordance with the identified quality metrics.

By adopting a Design of Experiments approach constituted by a full factorial experiment design and an Analysis of Variance (ANOVA) we have been able to pinpoint that, considering the graph metrics and their impact on the F-score, the litmus test for the quality of an SCDG-based classifier is represented by the presence of connected components. This could be explained considering how the graph mining algorithm (gSpan) works and the adopted similarity metric based on the number of common edges between the extracted signatures and the SCDG of the sample to classify. The results of the factorial experiments show that in our context tuning the symbolic execution is a very complex problem and that the sparsity of effect principle (stating that the system is dominated by the effect of the main factors and low-order-factor interactions) does not hold. The evaluation proved that the SMT solver is the most influential positive factor also showing an ability in reducing the impact of heuristics that may need to be enabled due to resource constraints (e.g., the max number of active paths). Results suggest that the most important factors are the disjoint union (as trace combination heuristic), and the our SMT optimization (through meta-heuristics) whereas other heuristics (such as min trace size and step timeout) have less impact on the quality of the constructed SCDGs.

Preliminary experiments show the promising results of our approach by considering the F-score in the classification of the malware families. Further investigation are needed in particular by using a larger dataset. For this purpose we established an academic collaboration with VirusTotal for helping us to build a ground truth for the family name.

One fundamental issue for supervised learning is the trustworthiness of the settled ground truth. In the scenario of malware classification, it is common to have great disagreement in the labeling of the very same malware sample (e.g. family attributed by different anti-malware vendors). Therefore, unsupervised learning on malware datasets by clustering based on the similarities of their SCDGs allows to overcome this problem.

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. Papers

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

- Efficient Extraction of Malware Signatures Through System Calls and Symbolic Execution: An Experience Report [28]

The ramping up use of network connected devices is providing hackers more incentives and opportunities to design and spread new security threats. Usually, malware analysts employ a mix of automated tools and human expertise to study the behavior of suspicious binaries and design suitable countermeasures. The analysis techniques adopted by automated tools include symbolic execution.Symbolic execution envisages the exploration of all the possible execution paths of the binary without neither concretizing the values of the variables nor dynamically executing the code (i.e., the binary is analyzed statically). Instead, all the values are represented symbolically. Progressing in the code exploration, constraints on symbolic variables are built and system calls tracked. A satisfiability-modulo-theory (SMT) checker is in charge of verifying the satisfiability of the collected symbolic constraints and thus the validity of an execution path. Unfortunately, while widely considered promising, this approach suffers from high resource consumption. Therefore, optimizing the constraint solver and tuning the features controlling symbolic execution is of fundamental importance to effectively adopting the technique. In this paper, we identify the metrics characterizing the quality of binary signatures expressed as system call dependency graphs extracted from a malware database. Then, we pinpoint some optimizations allowing to extract better binary signatures and thus to outperform the vanilla version of symbolic analysis tools in terms of malware classification and exploitation of the available resources.

7.3. Other research results

7.3.1. ContAv: a Tool to Assess Availability of Container-Based Systems

Participant: Stefano Sebastio.

This work was the result of a collaboration with former members of XRCI (Xerox Research Centre India): Rahul Ghosh, Avantika Gupta and Tridib Mukherjee.

[18] (C) The momentum gained by the microservice-oriented architecture is fostering the diffusion of operating system containers. Existing studies mainly focus on the performance of containerized services to demonstrate their low resource footprints. However, availability analysis of densely deployed container-based solutions is less visited due to difficulties in collecting failure artifacts. This is especially true when the containers are combined with virtual machines to achieve a higher security level. Inspired by Google's Kubernetes architecture, in this paper, we propose ContAv, an open-source distributed statistical model checker to assess availability of systems built on containers and virtual machines. The availability analysis is based on novel state-space and non-state-space models designed by us and that are automatically built and customized by the tool. By means of a graphical interface, ContAv allows domain experts to easily parameterize the system, to compare different configurations and to perform sensitivity analysis. Moreover, through a simple Java API, system architects can design and characterize the system behavior with a failure response and migration service.

7.3.2. (Coordination of the) TeamPlay Project, and Expression of Security Properties

Participants: Olivier Zendra, Yoann Marquer, Céline Minh, Annelie Heuser, Tania Richmond.

This work is done in the context of the TeamPlay EU project.

As mobile applications, the Internet of Things, and cyber-physical systems become more prevalent, so there is an increasing focus on energy efficiency of multicore computing applications. At the same time, traditional performance issues remain equally important. Increasingly, software designs need to find the best performance within some energy budget, often while also respecting real-time or other constraints, which may include security, data locality or system criticality, and while simultaneously optimising the usage of the available hardware resources.

While parallel multicore/manycore hardware can, in principle, ameliorate energy problems, and heterogeneous systems can help to find a good balance between execution time and energy usage, at present there are no effective analyses beyond user-guided simulations that can reliably predict energy usage for parallel systems, whether alone or in combination with timing information and security properties. In order to create energy-, time- and security- (ETS) efficient parallel software, programmers need to be actively engaged in decisions about energy usage, execution time and security properties rather than passively informed about their effects. This extends to design-time as well as to implementation-time and run-time.

In order to address this fundamental challenge, TeamPlay takes a radically new approach: by exploiting new and emerging ideas that allow non-functional properties to be deeply embedded within their programs, programmers can be empowered to directly treat energy ETS properties as first-class citizens in their parallel software. The concrete objectives of the TeamPlay project are:

1. To develop new mechanisms, along with their theoretical and practical underpinnings, that support direct language-level reasoning about energy usage, timing behaviour, security, etc.
2. To develop system-level coordination mechanisms that facilitate optimised resource usage for multicore hardware, combining system-level resource utilisation control during software development with efficient spatial and temporal scheduling at run-time.
3. To determine the fundamental inter-relationships between time, energy, security, etc. optimisations, to establish which optimisation approaches are most effective for which criteria, and to consequently develop multiobjective optimising compilers that can balance energy consumption against timing and other constraints.
4. To develop energy models for heterogeneous multicore architectures that are sufficiently accurate to enable high-level reasoning and optimisation during system development and at run-time.
5. To develop static and dynamic analyses that are capable of determining accurate time, energy usage and security information for code fragments in a way that can inform high-level programs, so achieving energy, time and security transparency at the source code level.
6. To integrate these models, analyses and tools into an analysis-based toolbox that is capable of reflecting accurate static and dynamic information on execution time and energy consumption to the programmer and that is capable of optimising time, energy, security and other required metrics at the whole system level.
7. To identify industrially-relevant metrics and requirements and to evaluate the effectiveness and potential of our research using these metrics and requirements.
8. To promote the adoption of advanced energy-, time- and security-aware software engineering techniques and tools among the relevant stake-holders.

Inria will exploit the results of the TeamPlay project in two main domains. First, they will strengthen and extend the research Inria has been carrying on low power and energy for embedded systems, especially for memory and wireless sensors networks. Second, they will complement in a very fitting way the research carried at Inria about security at a higher level (model checking, information theory).

The capability to express the energy and security properties at the developper level will be integrate in Inria own prototype tools, hence widening their applicability and the ease of experimentation. The use of energy properties wrt. evening of energy consumption to prevent information leakage, thus making side-channels attacks more difficult, is also a very promising path.

In addition, the methodological results pertaining to the development of embedded systems with a focus on low power and energy should also contribute to research lead at Inria in the domain of software engineering and advanced software engineering tools. Furthermore, security research lead at Inria will benefit from the security work undertaken by Inria and SIC in TeamPlay.

Overall, the project, with a strong industrial presence, will allow Inria to focus on matching concrete industrial requirements aiming at actual products, hence in providing more robust and validated results. In addition, the extra experience of working with industrial partners including SMEs will surely impact positively on Inria research methodology, making Inria research more attractive and influential, especially wrt. industry.

Finally, the results, both in terms of methodology and techniques, will also be integrated in the teaching Inria contributes to at Master level, in the areas of Embedded Systems and of Security.

The TeamPlay consortium agreement has been created by Inria, discussed with the various partners, and has been signed by all partners on 28 Feb. 2018. Inria has also distributed the partners initial share of the grant at the beginning of the project.

As WP7 (project management) leader and project coordinator, Inria was in charge of arranging general project meetings, including monthly meetings (tele-conferences), bi-annual physical meetings, boards meetings. During the first period, three exceptional physical meetings have been conducted, in addition to monthly project meetings: the kick-off meeting in Rennes from the 30th to the 31st of January 2018, the physical progress meeting has been conducted in Odense from the 26th to the 27th of June 2018, and the review in Brussels prepared the 19th of September 2018 and set the 17th of October 2018.

We have selected and set up utility tools for TeamPlay: shared notepads, mailing lists, shared calendars and collaborative repositories. We have ensured the timely production of the due deliverables. We set up the Project Advisory Board (PAB) with the aim of gathering external experts from both academia and industry, covering a wide range of domains addressed by TeamPlay. Finally, we ensured good working relationships (which can implicate conflict resolution when needed), monitored the overall progress of the project, and reported to the European Commission on technical matters and deliverables.

We also organized a tooling meeting in Hamburg in October the 30th, to discuss the relation between the tools from different partners, e.g. Idris from the University of St Andrews, the WCC compiler developed in the Hamburg University of Technology, or the coordination tool developed in the University of Amsterdam.

Measuring security, unlike measuring other more common non-functional properties like time or energy, is still very much in its infancy. For example, time is often measured in seconds (or divisions thereof), but security has no widely agreed, well-defined measurement. It is thus one goal of this project, especially for SIC and Inria, to design (necessarily novel) security measurements, and have them implemented as much as possible throughout the set of development tools.

Measuring security by only one value however seems impossible or may be meaningless. More precisely, if security could be defined overall by only one measurement, the latter would be a compound (i.e. an aggregation) of several more specialized measurement. Indeed, security encompasses many aspects of interest:

1. By allowing communications between different systems, security properties should be guaranteed in order to prevent low-level users from determining anything about high-level users activity, or in the case of public communication channels in a hostile environment, to evaluate vulnerability to intruders performing attacks on communications.
    1. *Confidentiality* (sometimes called *secrecy*) properties like non-interference (and many) variants can be described by using an information-flow policy (e.g. high- and low-level users) and studying traces of user inputs.
    2. *Vulnerability* captures how a system is sensible to attacks on communications (e.g. stealing or faking information on a public channel).
2. A *side-channel* is a way of transmitting informations (purposely or not) to another system out of the standard (intended) communication channels. *Side-channel attacks* rely on the relationship between information leaked through a side-channel and the secret data to obtain confidential (non-public) information.
    1. *Entropy* captures the uncertainty of the attacker about the secret key. The attacker must be able to extract information about the secret key through side-channel measurements, which is captured by the *attacker's remaining uncertainty* value, which can be computed by using heuristic techniques. The attacker must also be able to effectively recover the key from the extracted information, which is expressed by the *min-entropy leakage*, and refined by the *g-leakage* of a gain function.
    2. The power consumption of a cryptographic device can be analyzed to extract the secret key. This is done by using several techniques: visual examination of graphs of the current (*Simple Power Analysis*), by exploiting biases in varying power consumption (*Differential Power Analysis*), or by using the correlation coefficient between the power samples and hypotheses (*Correlation Power Analysis*).

3. Usual security properties guarantee only the input-output behavior of a program, and not its execution time. Closing *leakage through timing* can be done by disallowing while-loops and if-commands to depend on high security data, or by padding the branches so that the external observer cannot determine which branch was taken.
4. Finally, the correlation between the patterns of the victim's execution and the attacker's observations is formalized as a metric called the *Side-channel Vulnerability Factor*, which is refined by the *Cache Side-channel Vulnerability* for cache attacks.

3. A cryptographic scheme should be secure even if the attacker knows all details about the system, with the exception of the secret keys. In particular, the system should be secure when the attacker knows the encryption and decryption algorithms.
   1. In modern cryptography, the security level (or security strength) is given by the *work factor*, which is related to its key-length and the number of operations necessary to break a cryptographic scheme (try all possible combinations of the key). An algorithm is said to have a "security level of n bits" if the best known attack requires 2^n steps. This is a quite natural definition because symmetric algorithms with a security level of n have a key of length n bits.
   2. The relationship between cryptographic strength and security is not as straightforward in the asymmetric case. Moreover, for symmetric algorithms, a key-length of 128 bits provides an estimated long term security (i.e. several decades in the absence of quantum computer) regarding brute-force attacks. To reach an estimated long term security even with quantum computers, a key-length of 256 bits is mandatory.

Inria is implementing side-channel countermeasures (hiding) into the WCET-aware C Compiler (WCC) developed by the Hamburg University of Technology (TUHH). A research visit to TUHH was arranged with the aim at learning how to work on WCC (TUHH and WCC infrastructure, WCC developers best practices, etc.). Inria will use compiler-based techniques to prevent timing leakages and power leakages.

For instance, in a conditional branching $\texttt{if}\ b\ \texttt{then}\ P_1(x)\ \texttt{else}\ P_2(x)$, measuring the execution time or the power profile may allow to know whether the branch P_1 or P_2 have been chosen to manipulate the value x, thus to obtain the secret value b. To prevent timing leakage, P_1 and/or P_2 can be padded (i.e. dummy instructions are added) in order to obtain the worst-case execution time in both branches.

But this does not prevent information leakage from power profile. A stronger technique, from a security point of view, could be to add a dummy variable y and duplicate the code such that $y = x; \texttt{if}\ b\ \texttt{then}\ P_1(x); P_2(y)\ \texttt{else}\ P_1(Y); P_2(x)$ always performs the operations of P_1 then the operations of P_2. But the execution time is now the sum and not the worst-case of both branches, thus trading execution time to increase security.

Finally, the initialization $y = x$ can be detected, and the previous solution is still vulnerable to fault injections. Some algorithms like the Montgomery Ladder are more protected against these attacks because both variables x and y are entangled during the execution. We hope to generalize this property to a wider set of algorithms, or to automatically detect the properties required from the original code in order to transform it into a “Montgomerised" version with higher security level.

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 scheduling parameters for real-time systems modeled as synchronous data flow (SDF), cyclo-static dataflow (CSDF), and ultimately cyclo-static dataflow (UCSDF) graphs. This synthesis aims for a trade-off between throughput maximization and total buffer size minimization. The synthesizer inputs are a 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 synthesized scheduling parameters such as tasks periods, offsets, processor bindings, priorities, buffer initial markings and buffer sizes. In this section, we present new results on two aspects: (1) the improvement of ADFG's usability and tool interoperability, (2) the integration of new scheduling analysis and scheduler synthesis algorithms.

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 relevant metrics to analyze the scheduling efficiency.

Actor models and scheduling algorithms in ADFG are extended to investigate the contention-aware scheduling problem on multi/many-core architectures. The problem we tackled is that the scheduler synthesis for these platforms must account for the non-negligible delay due to shared memory accesses. We exploited the deterministic communications exposed in SDF graphs to account for the contention and further optimize the synthesized schedule. Two solutions are proposed and implemented in ADFG: contention-aware and contention-free scheduling synthesis. In other words, we either take into account the contention and synthesize a contention-aware schedule or find a one that results in no contention.

ADFG is extended to apply a transformation known as partial expansion graphs (PEG). This transformation can be applied as a pre-processing stage to improve the exploitation of data parallelism in SDF graphs on parallel platforms. In contrast to the classical approaches of transforming SDF graphs into equivalent homogeneous forms, which could lead to an exponential increase in the number of actors and excessive communication overhead, PEG-based approaches allow the designer to control the degree to which each actor is expanded. A PEG algorithm that employs cyclo-static data flow techniques is developed in ADFG. Compared to exist PEG-based approach, our solution requires neither buffer managers nor split-join actors to coordinate data production and consumption rates. This allows us to reduce the number of added actors and communication overhead in the expanded graphs.

7.2. Hardware synthesis in Polychrony

Participants: Loïc Besnard, Hafiz Muhamad Amjad.

[0] Real-Time Scheduling of Dataflow Graphs. A. Bouakaz. Ph.D. Thesis, University of Rennes 1, 2013.

[0] The Cheddar project: a GPL real-time scheduling analyzer: http://beru.univ-brest.fr/~singhoff/cheddar/

In the context of the Convex associate-project with the Chinese Academy of Science, we have this year developped code generators (VHDL, Verilog) for modeling hardware in Signal language [0]. The first scheme of the translation had been proposed by Mohammed Belhadj PhD Thesis. Independent on the HDL used, VHDL or Verilog, the translation of Signal to a HDL is quite simple, considering only the functional (executable) Signal programs and a behavioral translation. Indeed, behavioral translation is quite similar to a sequential code generator. In this case, the Signal compiler generates the clock of each SIGNAL signal and orders the execution of the equations. This control structure can be easily translated in the HDL. The generated code may contain conditionals, loops and signal assignements in a HDL process.

7.3. 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 ($d\mathcal{L}$) held our attention [0]. This formalism is built upon the Dynamic Logic of V. Pratt and augmented with the possibility of expressing Ordinary Differential Equations (ODEs). Combined with the ability of Dynamic Logic to specify and verify hybrid programs, $d\mathcal{L}$ is a particularly fit model cyber-physical systems. The proof system associated with the logic is implemented into the theorem prover KeYmaera X. Aimed toward automation, 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[6]. Commutativity provides compositionality: the possibility to compose and prove components and modules in every possible order. Associativity is mandatory to modularly design a system; it allows to construct step-by-step a system by adding new components. We have proved a theorem to automatically build contracts from the composition of components. We have exemplified our results with a cruise-controller example.

This contribution to $d\mathcal{L}$ defines a component-based approach to modularly model and prove cyber-physical systems. We have developed a working prototype in the interactive theorem prover KeYmaera X to show the feasibility of an implementation of our approach. To validate our methodology, we have case studied the example of a water-recycling plant, which rose several challenges.

The timing aspects in cyber-physical systems are a key aspect. A monitor regulating a plant, for example the water-level in a tank, must execute sufficiently often to ensure the correct behavior, for example that the water-level does not overflow. We have adapted our approach to automatically handle the compliance of execution time of monitor with the controllability of plant. We retain commutativity and associativity, thus the modularity of our approach. More importantly, we are still able to automatically build contracts from the composition of components.

We have also adapted our component-based approach to handle modes, a frequent construct in cyber-physical systems. It is also frequent to have to model causal composition between two components, for example that the sensor must execute before the monitor using the data of the sensor. Once again, we have adapted our component-based approach to take into account such design choice. It is important to emphasize that all these adaptations remain within our framework and are thus compatible.

To conclude, we have presented a methodology to tackle complexity of modeling and verification of cyber-physical systems by breaking a systems into smaller parts, the components. We have showed that it is easily adaptable to take into account new challenges. A future work would be to blend our component-based approach with refinement reasoning.

[0] Verilog Code Generation Scheme from Signal Language. Hafiz Muhammad, Amjad and Jianwei, Niu and Kai, Hu and Naveed, Akram and Loïc, Besnard. International Bhurban Conference on Applied Sciences and Technology (IBCAST). IEEE, 2019

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

7.4. Verified information flow of embedded programs

Participants: Jean-Joseph Marty, Jean-Pierre Talpin, Shravan Narayan, Deian Stefan, Rajesh Gupta.

This PhD project is about applying refinement types theory to verified programming of applications and modules of library operting systems, such as unikernels, for embedded devices (the Internet of Things (IoT). We focus on developping a model of information flow control approach using labelled input-outputs (LIO).

We are collaborating with the ProgSys group at UC San Diego in the frame of Inria associate-team Composite, which develops the LIO framework. The LIO framework allows to avoid the "label creep" problem and supports the modeling of concurrency.

Currentlymost of the properties implemented in LIO rely on Haskell properties which is not friendly for embedded devices (IoT), as Haskell requires a huge run-time compared to low resources micro-controllers with less than 32KB of memory.

Instead, we actively use the new Microsoft's verified programming language F*. This programming language is a proof assistant like language that allows us to formalize, verify (using SMT solver and tactics) and extract to clean C (without system dependency) . We succeeded in making proved programs on Arduino compatible micro-controller. Our aim is to develop a version of LIO that could be verified and then extracted to C for targeting operating systems or IoT.

At present, F* is a mix of three domain specific languages: Meta* for proof automation, Low* for system level code including memory safety and F* that glues everything. We successfully implemented a simple Low*-only LIO library allowing to use labeled values. We are now working on a formalized version that will ensure that an F* program is safe w.r.t. information flow, before code generation.

In parallel we continue to work with the ProgSys team on a second project: code-named Gluco*. The goal of this project is to strengthen the F* programming knowledge and to make a example of a safety-critical application where F* can be used [0].

7.5. Modeling cyber-physical systems: from Signal to Signal+

Participants: Thierry Gautier, Albert Benveniste.

Based, initially, on two small case studies, we started a reflection on the modeling and analysis of cyber-physical systems by extending the model of synchronous languages, and in particular that of the Signal language [15]. The principle considered here is to remain within the traditional framework, in discrete time, of synchronous languages, but to completely generalize the equational style of specifications. In the usual Signal language, clocks are defined in a relational way by constraint systems. In contrast, data are defined using functionally inspired dataflow expressions. In this new Signal+, we propose to generalize the equational style to the data: numerical quantities also become subject to constraints. This typically corresponds to the way of modeling a system that includes physical components: for example, equations respecting balance laws have to be written. A major interest is that this programming style is much more compositional than more traditional Simulink or Lustre, or even Signal-based programming. The question then arises as to whether such a program should be analyzed. There is no notion of syntactic dependence, but the incidence graph can be used to define a matching that uniquely associates an equation with each variable. A scheduling can then be synthesized, provided that the reasoning is valid, which is the case for some classes of numerical algebraic equations, for which a solver can be used. However, we can observe on our case studies that the transition to Signal+ class is a real step forward in terms of difficulty. In discrete time, at each reaction, the free variables of the system must be evaluated using what is known about states and inputs. But in some cases, there will be more variables than usable equations. By reasoning on the fact that we have systems that are invariant in time, it may then be necessary to “shift” equations, just as if we were in continuous time, we could differentiate a constraint (a program is a discrete approximation of a continuous time system). This amounts, in a way, to proposing some modifications, which are considered as legitimate, to the source program.

[0] Towards verified programming of embedded devices. J.-P. Talpin, J.-J. Marty, S. Narayan, D. Stefan, R. Gupta. Design, Automation and Test in Europe (DATE'19). IEEE, to appear 2018.

I4S Project-Team

5. New Results

5.1. System identification

5.1.1. Linear parameter varying system local model interpolation

Participant: Qinghua Zhang.

The local approach to linear parameter varying (LPV) system identification consists in interpolating a collection of linear time invariant (LTI) models, which have been estimated from data acquired at different working points of a nonlinear system. Interpolation is essential in this approach. When the local LTI models are in state-space form, as each local model can be estimated with an arbitrary state basis, it is widely acknowledged that the local models should be made coherent before their interpolation. In order to avoid the delicate task of making local state-space models coherent, a new interpolation method of local state-space models is proposed in this work, which does not require coherent local models. This method is based on the reduction of the large state-space model built by combining the local models. This work has been presented at SYSID 2018 [39].

5.1.2. State estimation for stochastic time varying systems with disturbance rejection

Participant: Qinghua Zhang.

State estimation in the presence of unknown disturbances is useful for the design of robust systems in different engineering fields. Most results available on this topic are restricted to linear time invariant (LTI) systems, whereas linear time varying (LTV) systems have been studied to a lesser extent. Existing results on LTV systems are mainly based on the minimization of the state estimation error covariance, ignoring the important issue of the stability of the state estimation error dynamics, which has been a main focus of the studies in the LTI case. The purpose of this work is to propose a numerically efficient algorithm for state estimation with disturbance rejection, in the general framework of LTV stochastic systems, including linear parameter varying (LPV) systems, with easily checkable conditions guaranteeing the stability of the algorithm. The design method is conceptually simple: disturbance is first rejected from the state equation by appropriate output injection, then the Kalman filter is applied to the resulting state-space model after the output injection. This work has been carried out in collaboration with Beijing University of Posts and Telecommunications (China) and has been presented at SYSID 2018 [40].

5.1.3. Variance estimation of modal indicators from subspace-based system identification

Participants: Michael Doehler, Laurent Mevel.

This work has been carried out in collaboration with Szymon Gres at Aalborg University and Palle Andersen at SVS.

One of the other practical modal indicators is Modal Assurance Criterion (MAC), for which uncertainty computation scheme is missing. This paper builds on the previous results using the propagation of the measurement uncertainties to estimates of MAC. The sensitivity of the MAC with respect to output covariances is derived using a first order perturbations and the uncertainties are propagated using the Delta method. The influence of the underlying mode shape scaling on both the uncertainty of mode shapes and MAC is investigated [22].

5.1.4. On damage detection system information for structural systems

5.1.4.1. On damage detection system information for structural systems

Participant: Michael Doehler.

Damage detection systems (DDSs) provide information on the integrity of structural systems in contrast to local information from inspections or non-destructive testing (NDT) techniques. In this paper, an approach is developed that utilizes DDS information to update structural system reliability and integrate this information into risk and decision analyses. For updating of the structural system reliability, an approach is developed based on Bayesian updating facilitating the use of DDS information on structural system level and thus for a structural system risk analysis. The structural system risk analysis encompasses the static, dynamic, deterioration, reliability and consequence models, which provide the basis for calculating the direct risks due to component failure and the indirect risks due to system failure[16].

5.1.4.2. The effects of deterioration models on the value of damage detection information

Participant: Michael Doehler.

This paper addresses the effects of the deterioration on the value of damage detection information. The quantification of the value of damage detection information for deteriorated structures is based on Bayesian pre-posterior decision analysis, comprising structural system performance models, consequence, benefit and costs models and damage detection information models throughout the service life of a structural system. With the developed approach, the value of damage detection information for a statically determinate Pratt truss bridge girder subjected to different deterioration models is calculated. The analysis shows the impact of the deterioration model parameters on the value of damage detection information. [28].

5.1.4.3. The effects of SHM system parameters on the value of damage detection information

Participant: Michael Doehler.

This paper addresses how the value of damage detection information depends on key parameters of the Structural Health Monitoring (SHM) system including number of sensors and sensor locations. The quantification of the value of information (VoI) is an expected utility based Bayesian decision analysis method for quantifying the difference of the expected economic benefits with and without information. The (pre-)posterior probability is computed utilizing the Bayesian updating theorem for all possible indications. Through the analysis of the value of information with different SHM system characteristics, the settings of DDS can be optimized for minimum expected costs and risks before implementation [29].

5.1.5. Filtering approaches for damage detection

5.1.5.1. Adaptive Kalman filter for actuator fault diagnosis

Participant: Qinghua Zhang.

An adaptive Kalman filter is proposed in this work for actuator fault diagnosis in discrete time stochastic time varying systems. By modeling actuator faults as parameter changes, fault diagnosis is performed through joint state-parameter estimation in the considered stochastic framework. Under the classical uniform complete observability-controllability conditions and a persistent excitation condition, the exponential stability of the proposed adaptive Kalman filter is rigorously analyzed. In addition to the minimum variance property of the combined state and parameter estimation errors, it is shown that the parameter estimation within the proposed adaptive Kalman filter is equivalent to the recursive least squares algorithm formulated for a fictive regression problem. These results have been published in [17].

5.1.5.2. Zonotopic state estimation and fault detection for systems with both time-varing and time-invariant uncertainties

Participant: Qinghua Zhang.

This paper proposes a robust guaranteed state estimation method with application to fault detection by combining H_∞ observer design with zonotopic analysis for discrete-time systems with both time-varing and time-invariant uncertainties. In order to improve the estimation accuracy, based on the H_∞ technique, the observer design is achieved by solving a linear matrix inequality. The main contribution of this paper lies in that the time invariance of some uncertainties is considered to reduce the conservatism of interval estimation. This work has been carried out in collaboration with Harbin Institute of Technology (China) and with Universitat Politècnica de Catalunya (Spain), and has been presented at SAFEPROCESS 2018 [37].

5.1.5.3. *Local adaptive observers for time-varying systems with parameter-dependent state matrices*

Participant: Qinghua Zhang.

The purpose of this work is to design an adaptive observer for linear time-varying systems whose state matrix is affine in some unknown parameters. In this case, the proposed observer generates state and parameter estimates, which exponentially converge to the plant state and the true parameters, respectively. The results are then extended to systems whose state matrix is nonlinear, instead of being affine, in the unknown parameters. This work has been carried out in collaboration with Université de Lorraine-CNRS-CRAN and has been presented at CDC 2018 [45].

5.1.5.4. *Seismic-induced damage detection through parallel force and parameter estimation using an improved interacting Particle-Kalman filter*

Standard filtering techniques for structural parameter estimation assume that the input force is either known 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. In this paper, the input force is considered to be an additional state that is estimated in parallel to the structural parameters. Two concurrent filters are employed for parameters and force respectively. For the parameters, an interacting Particle-Kalman filter is used to target systems with correlated noise. Alongside this, a second filter is used to estimate the seismic force acting on the structure [15].

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

Participants: Antoine Crinière, Laurent Mevel, Jean Dumoulin.

Existing filtering based structural health monitoring (SHM) algorithms assume constant noise environment which does not always conform to the reality as noise is hardly stationary. Thus to ensure optimal solution even with non-stationary noise processes, the assumed statistical noise models have to be updated periodically. This work incorporates a modification in the existing Interacting Particle-Kalman Filter (IPKF) to enhance its detection capability in presence of non-stationary noise processes. The Kalman filters (KF) within the IPKF have been replaced with a maximum Correntropy criterion (MCC) based KF that, unlike regular KF, takes moments beyond second order into consideration [32].

5.1.6. *Damage localization for mechanical structures*

5.1.6.1. *Damage localization using the stochastic load vectors*

Participants: Laurent Mevel, Michael Doehler.

This work was done in collaboration with BAM (Berlin) and GEM (Nantes).

In this work, a benchmark application is proposed, namely a 1/200 scale model of the Saint-Nazaire Bridge, which is a cable-stayed bridge spanning the Loire River near the river's mouth. The region of interest, the central metallic structure, measures 720 meters. The aim of the instrumentation is to assess the capability of damage assessment methods to assess a cable failure. The model is instrumented with ten accelerometers and excited by white noise. A damage localization method is applied to test the proposed setup, namely the statistical damage locating vector approach (S-SDDLV). With this method, vibration measurements from the (healthy) reference and damaged states of the structure are confronted to a finite element of the reference state. Damage indicators are provided for the different structural elements that are easy to compute, without updating the model parameters, and taking into account the intrinsic uncertainty of noisy measurements. [21].

5.1.6.2. *Asymptotic analysis of subspace-based data-driven residual for fault detection with uncertain reference*

Participants: Laurent Mevel, Michael Doehler, Eva Viefhues.

This work was in collaboration with BAM (Berlin).

The local asymptotic approach is promising for vibration-based fault diagnosis when associated to a subspace-based residual function and efficient hypothesis testing tools. In the residual function, the left null space of the observability matrix associated to a reference model is confronted to the Hankel matrix of output covariances estimated from test data. When this left null space is not perfectly known from a model, it should be replaced by an estimate from data to avoid model errors in the residual computation. In this paper, the asymptotic distribution of the resulting data-driven residual is analyzed and its covariance is estimated, which includes also the covariance related to the reference null space estimate [36].

5.1.7. Smarts roads and R5G

5.1.7.1. Multi-physics models for Energy Harvesting performance evaluation

Participants: Jean Dumoulin, Nicolas Le Touz.

We present in this paper the concept of solar hybrid road and focus on the thermal performances of such system. A finite element model is presented to couple thermal diffusion, hydraulic convection and radiative transfer. This numerical model allows to compute the temperature field for different weather conditions and also to evaluate the thermal performances of the system. Annual simulations are performed and a comparison between two surface layer solutions for different locations and climates is presented and discussed[23].

5.1.7.2. Optimal command for defreezing of solar road

Participants: Jean Dumoulin, Nicolas Le Touz.

The study presented in [26] aims to optimize the amount of energy to bring to a hybrid solar road to prevent the formation of ice on the surface. The optimal control law studied is based on a finite element multiphysics model, developed to compute the temperature field in the structure under varying environmental conditions presented in [25]. A penalization of freezing periods at the surface is introduced and the energy to be supplied to the system to preserve it is calculated from the adjoint state method [24].

5.1.8. Infrared Thermography

5.1.8.1. Sensitivity of infrared camera to environmental parameters

Participants: Laurent Mevel, Jean Dumoulin, Thibaud Toullier.

The purpose of this study is to characterize the influence of environmental parameters for long-term in-situ structure monitoring as well as projections errors due to camera view and digitization. The model used to convert 3 year gathered data to temperature is firstly presented and discussed. Then, the effect of camera resectioning on infrared measurements is commented. Finally, the effect of the environmental parameters is studied and perspectives are proposed [35].

5.1.8.2. Joint Estimation of emissivity and temeperature

Participants: Laurent Mevel, Jean Dumoulin, Thibaud Toullier.

This study deals with the simultaneous assessment of emissivity and surface temperature. of objects observed by in-situ infrared thermography. Temperature measurement by thermography infrared is hampered by the lack of knowledge of the radiative properties of the real world. The light received from a target by an infrared camera is estimated by the method of progressive radiosities implemented on a map graphic in order evaluate the sensitivity of four methods of separation of emissivity and temperature [34].

5.1.9. Sensor and hardware based research

5.1.9.1. Reflectometry

Participant: Qinghua Zhang.

5.1.9.2. De-embedding unmatched connectors for electric cable fault diagnosis

Participant: Qinghua Zhang.

In order to make accurate reflectometry measurements on electric cables for fault diagnosis, connector de-embedding is a procedure for compensating measurement distortions caused by unmatched connectors. The key step in such a procedure is the characterization of the connectors, which is realized through measurements on a pair of connectors linked by a short cable segment. The analysis for deducing the characteristics of a single connector from measurements made on an assembled pair is known as the bisection problem. In this paper, after recalling the underdetermined nature of the bisection problem, a practically effective de-embedding procedure is proposed based on a particular regularization technique. This work has been carried out in collaboration with EDF R&D and has been presented at SAFEPROCESS 2018 [38].

5.1.9.3. Active Infrared thermography by robot

Participants: Jean Dumoulin, Ludovic Gaverina.

In this paper, two Non Destructive Testing approaches by active infrared thermography mounted on a 6-axis robot are presented and studied. An automated procedure is proposed to reconstruct thermal image sequences issued from the two scanning procedure studied: Line Scan and Flying Line procedures. Defective area detection is performed by image processing and an inverse technique based on thermal quadrupole method is used to map the depth of flaws [31].

5.1.9.4. Shunting monitoring in railway track circuit receivers

Participant: Vincent Le Cam.

Track circuits play a major role in railway signaling. In some exceptional conditions, poor rail/wheel contact conditions may lead to a non-detection of the train on the zone. The paper presents new detection approaches based on signal processing on an experiment with a dedicated train running on a track equipped with a track circuit. The second objective is to present a strategy to test new detection criteria on commercial zones over a long period of time using PEGASE [30].

MINGUS Project-Team

7. New Results

7.1. Highly-oscillatory problems

7.1.1. Highly-oscillatory problems with time-dependent vanishing frequency

In the analysis of highly-oscillatory evolution problems, it is commonly assumed that a single frequency is present and that it is either constant or, at least, bounded from below by a strictly positive constant uniformly in time. Allowing for the possibility that the frequency actually depends on time and vanishes at some instants introduces additional difficulties from both the asymptotic analysis and numerical simulation points of view. This work [27] is a first step towards the resolution of these difficulties. In particular, P. Chartier, M. Lemou, F. Méhats and G. Vilmart show that it is still possible in this situation to infer the asymptotic behaviour of the solution at the price of more intricate computations and we derive a second order uniformly accurate numerical method.

7.1.2. Uniformly accurate methods for Vlasov equations with non-homogeneous strong magnetic field

In this paper [26], the authors P. Chartier, N. Crouseilles, M. Lemou, F. Méhats and X. Zhao consider the numerical solution of highly-oscillatory Vlasov and Vlasov-Poisson equations with non-homogeneous magnetic field. Designed in the spirit of recent uniformly accurate methods, our schemes remain insensitive to the stiffness of the problem, in terms of both accuracy and computational cost. The specific difficulty (and the resulting novelty of our approach) stems from the presence of a non-periodic oscillation, which necessitates a careful ad-hoc reformulation of the equations. Our results are illustrated numerically on several examples.

7.1.3. Uniformly accurate time-splitting methods for the semiclassical linear Schrödinger equation

The paper [8] is devoted to the construction of numerical methods which remain insensitive to the smallness of the semiclassical parameter for the linear Schrödinger equation in the semiclassical limit. We specifically analyse the convergence behavior of the first-order splitting. Our main result is a proof of uniform accuracy. The authors illustrate the properties of our methods with simulations. Philippe Chartier, Loïc Le Treust, Florian Méhats then illustrate the properties of the methods with simulations.

7.1.4. Numerical methods for the two-dimensional Vlasov-Poisson equation in the finite Larmor radius approximation regime

In this paper [7], the authors P. Chartier, N. Crouseilles and X. Zhao consider the numerical methods for solving the two-dimensional Vlasov-Poisson equation in the finite Larmor radius approximation regime. The model describes the behaviour of charged particles under a strong external magnetic field and the finite Larmor radius approximation. We discretise the equation under Particle-in-Cell method, where the characteristics equations are highly oscillatory system in the limit regime. We apply popular numerical integrators including splitting methods, multi-revolution composition methods, two-scale formulation method and limit solver to integrate the characteristics. Dissuasions are made to highlight the strength and drawback of each method. Numerical experiments are done, and comparisons on the accuracy, efficiency and long-time behaviour of the methods are made, aiming to suggest the method with the best performance for the problem.

7.1.5. A new class of uniformly accurate numerical schemes for highly oscillatory evolution equations

In [9], we introduce a new methodology to design uniformly accurate methods for oscillatory evolution equations. The targeted models are envisaged in a wide spectrum of regimes, from non stiff to highly oscillatory. Thanks to an averaging transformation, the stiffness of the problem is softened, allowing for standard schemes to retain their usual orders of convergence. Overall, high order numerical approximations are obtained with errors and at a cost independent of the regime.

7.1.6. Uniformly accurate exponential-type integrators for Klein-Gordon equations with asymptotic convergence to classical splitting schemes in the nonlinear Schrödinger limit

In [2], we introduce efficient and robust exponential-type integrators for Klein-Gordon equations which resolve the solution in the relativistic regime as well as in the highly-oscillatory non-relativistic regime without any step-size restriction, and under the same regularity assumptions on the initial data required for the integration of the corresponding limit system. In contrast to previous works we do not employ any asymptotic/multiscale expansion of the solution. This allows us derive uniform convergent schemes under far weaker regularity assumptions on the exact solution. In particular, the newly derived exponential-type integrators of first-, respectively, second-order converge in the non-relativistic limit to the classical Lie, respectively, Strang splitting in the nonlinear Schrödinger limit.

7.1.7. A micro-macro method for a kinetic graphene model in one-space dimension

In [29], for the one space dimensional semiclassical kinetic graphene model recently introduced in the literature, we propose a micro-macro decomposition based numerical approach, which reduces the computational dimension of the nonlinear geometric optics method based numerical method for highly oscillatory transport equation developed in a previous work. The method solves the highly oscillatory model in the original coordinate, yet can capture numerically the oscillatory space-time quantum solution pointwisely even without numerically resolving the frequency. We prove that the underlying micro-macro equations have smooth (up to certain order of derivatives) solutions with respect to the frequency, and then prove the uniform accuracy of the numerical discretization for a scalar model equation exhibiting the same oscillatory behavior. Numerical experiments verify the theory.

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

In [12], 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.

7.1.9. Symmetric high order Gautschi-type exponential wave integrators pseudospectral method for the nonlinear Klein-Gordon equation in the nonrelativistic limit regime

In [19], a group of high order Gautschi-type exponential wave integrators (EWIs) Fourier pseudospectral method are proposed and analyzed for solving the nonlinear Klein-Gordon equation (KGE) in the nonrelativistic limit regime, where a parameter which is inversely proportional to the speed of light, makes the solution propagate waves with wavelength in time and in space. With the Fourier pseudospectral method to discretize the KGE in space, we propose a group of EWIs with designed Gautschi's type quadratures for the temporal integrations, which can offer any intended even order of accuracy provided that the solution is smooth enough, while all the current existing EWIs offer at most second order accuracy. The scheme is explicit, time symmetric and rigorous error estimates show the meshing strategy of the proposed method is time step and mesh size as, which is optimal among all classical numerical methods towards solving the KGE directly in the limit regime, and which also distinguish our methods from other high order approaches such as Runge-Kutta methods which require . Numerical experiments with comparisons are done to confirm the error bound and show the superiority of the proposed methods over existing classical numerical methods.

7.1.10. On the rotating nonlinear Klein-Gordon equation: non-relativistic limit and numerical methods

In [32], we consider both numerics and asymptotics aspects for the rotating nonlinear Klein Gordon (RKG) equation, an important PDE in relativistic quantum physics that can model a rotating galaxy in Minkowski metric and serves also as a model e.g. for a "cosmic superfluid". Firstly, we formally show that in the non-relativistic limit RKG converges to coupled rotating nonlinear Schrödinger equations (RNLS), which is used to describe the particle-antiparticle pair dynamics. Investigations of the vortex state of RNLS are carried out. Secondly, we propose three different numerical methods to solve RKG from relativistic regimes to non-relativistic regimes in polar and Cartesian coordinates. In relativistic regimes, a semi-implicit finite difference Fourier spectral method is proposed in polar coordinates where both rotation terms are diagonalized simultaneously. While in non relativistic regimes, to overcome the fast temporal oscillations, we adopt the rotating Lagrangian coordinates and introduce two efficient multiscale methods with uniform accuracy, i.e., the multi revolution composition method and the exponential integrator. Various numerical results confirm (uniform) accuracy of our methods. Simulations of vortices dynamics are presented.

7.2. Numerical schemes for Hamiltonian PDEs

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

In [14], we consider time discretizations of the Vlasov-HMF (Hamiltonian Mean-Field) equation based on splitting methods between the linear and non-linear parts. We consider solutions starting in a small Sobolev neighborhood of a spatially homogeneous state satisfying a linearized stability criterion (Penrose criterion). We prove that the numerical solutions exhibit a scattering behavior to a modified state, which implies a nonlinear Landau damping effect with polynomial rate of damping. Moreover, we prove that the modified state is close to the continuous one and provide error estimates with respect to the time stepsize.

7.2.2. 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 [17], The focus of this paper is 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 be at $O(h^2 + \tau^2)$ 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.

7.2.3. Modulation equations approach for solving vortex and radiation in nonlinear Schrödinger equation

In [16], we apply the modulation theory to study the vortex and radiation solution in the 2D nonlinear Schrödinger equation. The full modulation equations which describe the dynamics of the vortex and radiation separately are derived. A general algorithm is proposed to efficiently and accurately find vortices with prescribed values of energy and spin index. The modulation equations are solved by accurate numerical method. Numerical tests and simulations of radiation are given.

7.2.4. Unconditional L^∞-convergence of two compact conservative finite difference schemes for the nonlinear Schrödinger equation in multi-dimensions

In [18], we are concerned with the unconditional and optimal L^∞-error estimates of two fourth-order (in space) compact conservative finite difference time domain schemes for solving the nonlinear Schrödinger equation in two or three space dimensions. The fact of high space dimension and the approximation via compact finite difference discretization bring difficulties in the convergence analysis. The two proposed schemes preserve the total mass and energy in the discrete sense. To establish the optimal convergence results without any constraint on the time step, besides the standard energy method, the cut-off function technique as well as a lifting technique are introduced. On the contrast, previous works in the literature often require certain restriction on the time step. The convergence rate of the proposed schemes are proved to be of $O(h^4 + \tau^2)$ with time step τ and mesh size h in the discrete L^∞-norm. The analysis method can be directly extended to other finite difference schemes for solving the nonlinear Schrödinger-type equations. Numerical results are reported to support our theoretical analysis, and investigate the effect of the nonlinear term and initial data on the blow-up solution.

7.2.5. Verification of $2D$x$2D$ and two-species Vlasov-Poisson solvers

Recently $1D$x$1D$ two-species Vlasov-Poisson simulations have been performed by the semi-Lagrangian method. Thanks to a classical first order dispersion analysis, we are able to check in [1] the validity of their simulations; the extension to second order is performed and shown to be relevant for explaining further details. In order to validate multi-dimensional effects, we propose a $2D$x$2D$ single species test problem that has true $2D$ effects coming from the sole second order dispersion analysis. Finally, we perform, in the same code, full $2D$x$2D$ non linear two-species simulations with mass ratio $\sqrt{0.01}$, and consider the mixing of semi-Lagrangian and Particle-in-Cell methods.

7.2.6. An exponential integrator for the drift-kinetic model

In [11], 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.

7.2.7. Convergence of a normalized gradient algorithm for computing ground states

In [15], we consider the approximation of the ground state of the one-dimensional cubic nonlinear Schrödinger equation by a normalized gradient algorithm combined with linearly implicit time integrator, and finite difference space approximation. We show that this method, also called imaginary time evolution method in the physics literature, is con-vergent, and we provide error estimates: the algorithm converges exponentially towards a modified solitons that is a space discretization of the exact soliton, with error estimates depending on the discretization parameters.

7.3. Analysis of PDE

7.3.1. Bounds on the growth of high discrete Sobolev norms for the cubic discrete nonlinear Schrödinger equations on $h\mathbb{Z}$

In [22], we consider the discrete nonlinear Schrödinger equations on a one dimensional lattice of mesh h, with a cubic focusing or defocusing nonlinearity. We prove a polynomial bound on the growth of the discrete Sobolev norms, uniformly with respect to the stepsize of the grid. This bound is based on a construction of higher modified energies.

7.3.2. Existence and stability of traveling waves for discrete nonlinear Schrödinger equations over long times

In [23], we consider the problem of existence and stability of solitary traveling waves for the one dimensional discrete non linear Schrödinger equation (DNLS) with cubic nonlinearity, near the continuous limit. We construct a family of solutions close to the continuous traveling waves and prove their stability over long times. Applying a modulation method, we also show that we can describe the dynamics near these discrete traveling waves over long times.

7.3.3. Smoothing properties of fractional Ornstein-Uhlenbeck semigroups and null-controllability

In [20], we study fractional hypoelliptic Ornstein-Uhlenbeck operators acting on $L^2(\mathbb{R}^n)$ satisfying the Kalman rank condition. We prove that the semigroups generated by these operators enjoy Gevrey regularizing effects. Two byproducts are derived from this smoothing property. On the one hand, we prove the null-controllability in any positive time from thick control subsets of the associated parabolic equations posed on the whole space. On the other hand, by using the interpolation theory, we get global L^2 subelliptic estimates for the these operators.

7.3.4. Stable ground states for the HMF Poisson model

In [31], 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.

7.4. Dissipative problems

7.4.1. A formal series approach to the center manifold theorem

In [4], the author considers 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.

7.4.2. Analysis of an asymptotic preserving scheme for stochastic linear kinetic equations in the diffusion limit

In [21], we present an asymptotic preserving scheme based on a micro-macro decomposition for stochastic linear transport equations in kinetic and diffusive regimes. We perform a mathematical analysis and prove that the scheme is uniformly stable with respect to the mean free path of the particles in the simple telegraph model and in the general case. We present several numerical tests which validate our scheme.

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

In [28], 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 of the authors 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.

7.4.4. Time diminishing schemes (TDS) for kinetic equations in the diffusive scaling

In [28], we develop a new class of numerical schemes for collisional kinetic equations in the diffusive regime. The first step consists in reformulating the problem by decomposing the solution in the time evolution of an equilibrium state plus a perturbation. Then, the scheme combines a Monte Carlo solver for the perturbation with a Eulerian method for the equilibrium part, and is designed in such a way to be uniformly stable with respect to the diffusive scaling and to be consistent with the asymptotic diffusion equation. Moreover, since particles are only used to describe the perturbation part of the solution, the scheme becomes computationally less expensive - and is thus time diminishing (TDS) - as the solution approaches the equilibrium state due to the fact that the number of particles diminishes accordingly. This contrasts with standard methods for kinetic equations where the computational cost increases (or at least does not decrease) with the number of interactions. At the same time, the statistical error due to the Monte Carlo part of the solution decreases as the system approaches the equilibrium state: the method automatically degenerates to a solution of the macroscopic diffusion equation in the limit of infinite number of interactions. After a detailed description of the method, we perform several numerical tests and compare this new approach with classical numerical methods on various problems up to the full three dimensional case.

7.5. Stochastic PDE

7.5.1. Linearized wave turbulence convergence results for three-wave systems

In [30], E. Faou considers stochastic and deterministic three-wave semi-linear systems with bounded and almost continuous set of frequencies. Such systems can be obtained by considering nonlinear lattice dynamics or truncated partial differential equations on large periodic domains. We assume that the nonlinearity is small and that the noise is small or void and acting only in the angles of the Fourier modes (random phase forcing). We consider random initial data and assume that these systems possess natural invariant distributions corresponding to some Rayleigh-Jeans stationary solutions of the wave kinetic equation appearing in wave turbulence theory. We consider random initial modes drawn with probability laws that are perturbations of theses invariant distributions. In the stochastic case, we prove that in the asymptotic limit (small nonlinearity, continuous set of frequency and small noise), the renormalized fluctuations of the amplitudes of the Fourier modes converge in a weak sense towards the solution of the linearized wave kinetic equation around these Rayleigh-Jeans spectra. Moreover, we show that in absence of noise, the deterministic equation with the same random initial condition satisfies a generic Birkhoff reduction in a probabilistic sense, without kinetic description at least in some regime of parameters.

7.5.2. Large deviations for the dynamic Φ^{2n}_d model

In [5], 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 $0 < \rho, \delta(\rho) \ll 1$. 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.

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

In [3], 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 previously. The tools we develop are very general and can be used to study many other examples of applications.

7.5.4. Large deviations for the two-dimensional stochastic Navier-Stokes equation with vanishing noise correlation

In [6], we are dealing with the validity of a large deviation principle for the two-dimensional Navier-Stokes equation, with periodic boundary conditions, 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 ε and $\delta(\varepsilon)$, respectively, with $0 < \varepsilon, \delta(\varepsilon) \ll 1$. Depending on the relationship between ε and $\delta(\varepsilon)$ we will prove the validity of the large deviation principle in different functional spaces.

7.5.5. The Schrödinger equation with spatial white noise potential

In [13], we consider the linear and nonlinear Schrödinger equation with a spatial white noise as a potential in dimension 2. We prove existence and uniqueness of solutions thanks to a change of unknown originally used in a paper by Hairer and Labbé (2015) and conserved quantities.

SIMSMART Team

6. New Results

6.1. Objective 1 – Rare events simulation

In [16], we present a short historical perpective of the importance splitting approach to simulate and estimate rare events, with a detailed description of several variants. We then give an account of recent theoretical results on these algorithms, including a central limit theorem for Adaptive Multilevel Splitting (AMS). Considering the asymptotic variance in the latter, the choice of the importance function, called the reaction coordinate in molecular dynamics, is also discussed. Finally, we briefly mention some worthwhile applications of AMS in various domains.

Adaptive Multilevel Splitting (AMS for short) is a generic Monte Carlo method for Markov processes that simulates rare events and estimates associated probabilities. Despite its practical efficiency, there are almost no theoretical results on the convergence of this algorithm. In [15], we prove both consistency and asymptotic normality results in a general setting. This is done by associating to the original Markov process a level-indexed process, also called a stochastic wave, and by showing that AMS can then be seen as a Fleming-Viot type particle system. This being done, we can finally apply general results on Fleming-Viot particle systems that we have recently obtained.

Probability measures supported on submanifolds can be sampled by adding an extra momentum variable to the state of the system, and discretizing the associated Hamiltonian dynamics with some stochastic perturbation in the extra variable. In order to avoid biases in the invariant probability measures sampled by discretizations of these stochastically perturbed Hamiltonian dynamics, a Metropolis rejection procedure can be considered. The so-obtained scheme belongs to the class of generalized Hybrid Monte Carlo (GHMC) algorithms. In [21], we show here how to generalize to GHMC a procedure suggested by Goodman, Holmes-Cerfon and Zappa for Metropolis random walks on submanifolds, where a reverse projection check is performed to enforce the reversibility of the algorithm for large timesteps and hence avoid biases in the invariant measure. We also provide a full mathematical analysis of such procedures, as well as numerical experiments demonstrating the importance of the reverse projection check on simple toy examples.

Feynman-Kac semigroups appear in various areas of mathematics: non-linear filtering, large deviations theory, spectral analysis of Schrodinger operators among others. Their long time behavior provides important information, for example in terms of ground state energy of Schrodinger operators, or scaled cumulant generating function in large deviations theory. In [17], we propose a simple and natural extension of the stability of Markov chains for these non-linear evolutions. As other classical ergodicity results, it relies on two assumptions: a Lyapunov condition that induces some compactness, and a minorization condition ensuring some mixing. We show that these conditions are satisfied in a variety of situations. We also show that our technique provides uniform in the time step convergence estimates for discretizations of stochastic differential equations.

6.2. Objective 2 – High dimensional and advanced particle filtering

Existing filtering based structural health monitoring (SHM) algorithms assume constant noise environment which does not always conform to the reality as noise is hardly stationary. Thus to ensure optimal solution even with non-stationary noise processes, the assumed statistical noise models have to be updated periodically. [8] incorporates a modification in the existing Interacting Particle-Kalman Filter (IPKF) to enhance its detection capability in presence of non-stationary noise processes. To achieve noise adaptability, the proposed algorithm recursively estimates and updates the current noise statistics using the post-IPKF residual uncertainty in prediction as a measurement which in turn enhances the optimality in the solution as well. Further, this algorithm also attempts to mitigate the ill effects of abrupt change in noise statistics which most often deteriorates/ diverges the estimation. For this, the Kalman filters (KF) within the IPKF have been replaced with a maximum Correntropy criterion (MCC) based KF that, unlike regular KF, takes moments beyond second order into consideration. A Gaussian kernel for MCC criterion is employed to define a correntropy index that controls the update in state and noise estimates in each recursive steps. Numerical experiments on an eight degrees-of-freedom system establish the potential of this algorithm in real field applications.

Standard filtering techniques for structural parameter estimation assume that the input force is either known 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, force must therefore also be estimated. In [5], the input force is considered to be an additional state that is estimated in parallel to the structural parameters. Two concurrent filters are employed for parameters and force respectively. For the parameters, an interacting Particle-Kalman filter is used to target systems with correlated noise. Alongside this, a second filter is used to estimate the seismic force acting on the structure. In the proposed algorithm, the parameters and the inputs are estimated as being conditional on each other, thus ensuring stability in the estimation. The proposed algorithm is numerically validated on a sixteen degrees-of-freedom mass-spring-damper system and a five-story building structure. The stability of the proposed filter is also tested by subjecting it to a sufficiently long measurement time history. The estimation results confirm the applicability of the proposed algorithm.

6.3. Objective 3 – Non-parametric inference

The forecasting and reconstruction of ocean and atmosphere dynamics from satellite observation time series are key challenges. While model-driven representations remain the classic approaches, data-driven representations become more and more appealing to benefit from available large-scale observation and simulation datasets. In [12], [13] and [4], we investigate the relevance of recently introduced neural network representations for the forecasting and assimilation of geophysical fields from satellite-derived remote sensing data. As a case-study, we consider satellite-derived Sea Surface Temperature time series off South Africa, which involves intense and complex upper ocean dynamics. Our numerical experiments report significant improvements in terms of reconstruction performance compared with operational and state-of-the-art schemes.

Data assimilation methods aim at estimating the state of a system by combining observations with a physical model. When sequential data assimilation is considered, the joint distribution of the latent state and the observations is described mathematically using a state-space model, and filtering or smoothing algorithms are used to approximate the conditional distribution of the state given the observations. The most popular algorithms in the data assimilation community are based on the Ensemble Kalman Filter and Smoother (EnKF/EnKS) and its extensions. In [14], we investigate an alternative approach where a Conditional Particle Filter (CPF) is combined with Backward Simulation (BS). This allows to explore efficiently the latent space and simulate quickly relevant trajectories of the state conditionally to the observations. We also tackle the difficult problem of parameter estimation. Indeed, the models generally involve statistical parameters in the physical models and/or in the stochastic models for the errors. These parameters strongly impact the results of the data assimilation algorithm and there is a need for an efficient method to estimate them. Expectation-Maximization (EM) is the most classical algorithm in the statistical literature to estimate the parameters in models with latent variables. It consists in updating sequentially the parameters by maximizing a likelihood function where the state is approximated using a smoothing algorithm. In this paper, we propose an original Stochastic Expectation-Maximization (SEM) algorithm combined to the CPF-BS smoother to estimate the statistical parameters. We show on several toy models that this algorithm provides, with reasonable computational cost, accurate estimations of the statistical parameters and the state in highly nonlinear state-space models, where the application of EM algorithms using EnKS is limited. We also provide a Python source code of the algorithm.

6.4. Objective 4 – Model reduction

In [19], [7], we propose new methodologies to decrease the computational cost of safe screening tests for LASSO. We first introduce a new screening strategy, dubbed "joint screening test", which allows the rejection of a set of atoms by performing one single test. Our approach enables to find good compromises between complexity of implementation and effectiveness of screening. Second, we propose two new methods to decrease the computational cost inherent to the construction of the (so-called) "safe region". Our numerical experiments show that the proposed procedures lead to significant computational gains as compared to standard methodologies.

Model-order reduction methods tackle the following general approximation problem: find an "easily-computable" but accurate approximation of some target solution h. In order to achieve this goal, standard methodologies combine two main ingredients: i) a set of problem-specific constraints; ii) some "simple" prior model on the set of target solutions. The most common prior model encountered in the literature assume that the target solution h is "close" to some low-dimensional subspace. Recently, several contributions have shown that refined prior models (based on a set of embedded approximation subspaces) may lead to enhanced approximation performance. Unfortunately, to date, no theoretical results have been derived to support the good empirical performance observed in these contributions. The goal of [18] is to fill this gap. More specifically, we provide a mathematical characterization of the approximation performance achievable by some particular "multi-space" decoder and emphasize that, in some specific setups, this "multi-space" decoder has provably better recovery guarantees than its standard counterpart based on a single approximation subspace.

In [20], we deal with the estimation of rare event probabilities using importance sampling (IS), where an *optimal* proposal distribution is computed with the cross-entropy (CE) method. Although, IS optimized with the CE method leads to an efficient reduction of the estimator variance, this approach remains unaffordable for problems where the repeated evaluation of the score function represents a too intensive computational effort. This is often the case for score functions related to the solution of parametric partial differential equations (PPDE) with random inputs. This work proposes to alleviate computation by adapting a score function approximation along the CE optimization process. The score function approximation is obtained by selecting the surrogate of lowest dimensionality, whose accuracy guarantees to pass the current CE optimization stage. The adaptation of the surrogate relies on certified upper bounds on the error norm. An asymptotic analysis provides some theoretical guarantees on the efficiency and convergence of the proposed algorithm. Numerical results demonstrate the gain brought by the adaptive method in the context of pollution alerts and a system modelled by a PPDE.

In [2], we deal with model order reduction of PPDE. We consider the specific setup where the solutions of the PPDE are only observed through a partial observation operator and address the task of finding a good approximation subspace of the solution manifold. We provide and study several tools to tackle this problem. We first identify the best worst-case performance achievable in this setup and propose simple procedures to approximate this optimal solution. 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.

In [3], 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; *(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 e.g. Proper Orthogonal Decomposition or Dynamic Mode Decomposition, can be revisited and recast within the probabilistic framework considered in this paper. We illustrate the performance of the proposed approach by numerical results obtained for a standard geophysical model.

6.5. Miscellaneous

In [22], we devise methods of variance reduction for the Monte Carlo estimation of an expectation of the type $\mathbb{E}[\phi(X,Y)]$, when the distribution of X is exactly known. The key general idea is to give each individual of a sample a weight, so that the resulting weighted empirical distribution has a marginal with respect to the variable X as close as possible to its target. We prove several theoretical results on the method, identifying settings where the variance reduction is guaranteed. We perform numerical tests comparing the methods and demonstrating their efficiency.

DYLISS Project-Team

7. New Results

7.1. Scalable methods to query data heterogenity

Participants: Guillaume Alviset, Olivier Dameron, Xavier Garnier, Vijay Ingalalli, Marine Louarn, Yann Rivault, Anne Siegel, Denis Tagu.

Ontology design and integration [*O. Dameron, Y. Rivault*] We have contributed to several technics improving data integration in onthologies

- The ATOL ontology [link to ontology] supports the annotation of phenotype traits in livestock. It was extended with health-related traits. For each organism, livestock diseases are organized according to their type (infectious, genetic, metabolic,...), their transmission and their symptoms. [32]
- queryMed is an R package [url] that provides both high-level and low-level functions for facilitating the integration of reference ontologies and datasets represented in RDF as Linked Data. It currently focuses on drugs indications, interactions and contra-indications by integrating the Drug Indication Database (DID) and the Drug Interaction Knowledge Base (DIKB). Typical applications concern public health and pharmaco-epidemiology. [27], [26]

Using AskOmics to integrate heterogeneous data [*O. Dameron, A. Siegel*]

- We contributed to the conversion of an Alzheimer's disease map into a heavyweight ontology, the Alzheimer's Disease Map Ontology (ADMO, [url]), an ontological upper model based on systems biology terms. It provides the ontological formalization for the existing disease map AlzPathway that gives a detailed and broad account of Alzheimer's Disease pathophysiology [25], [20].
- We also contributed to decipher the role of small non-coding RNAs in the regulation of animal reproduction, especially the role of miR-202 in female fecundity by regulating medaka oncogenesis [16].

Graph compression and analysis [*L. Bourneuf*]. We introduced a general approach combining procedural and logical languages to specify graph objects. This is a generalization of previous work [37], using the reconstruction of Formal Concept Analysis framework example to target the AI community [23].

7.2. Metabolism: from enzyme sequences to systems ecology

Participants: Meziane Aite, Arnaud Belcour, Marie Chevallier, Mael Conan, François Coste, Olivier Dameron, Clémence Frioux, Jeanne Got, Jacques Nicolas, Anne Siegel, Hugo Talibart.

Efficient identification of substitutable context-free grammars by reduction [*F. Coste, J. Nicolas*] To study more formally the approach by reduction initiated by ReGLiS [40], we introduced a formal characterization of the grammars in reduced normal form (RNF) which can be learned by this approach. Modifying the core of ReGLiS to ensure polynomial running time, we show that local substitutable languages represented by RNF context-free grammars are identifiable in polynomial time and thick data (IPTtD) from positive examples by reduction [19].

Learning grammars capturing 3D structural features of proteins [*F. Coste, H. Talibart*] With the team of Witold Dyrka in Polland, we investigated the problem of learning context-free grammars modeling well protein sequences with respect to their 3D structures.

- A preliminary step is to be able to quantify the relevance of a grammar with respect to a structure. In [21], we introduced and assessed quantitative measures for comparing the topology of the parse tree of a protein sequence analyzed by a context-free grammar with the topology of the protein structure.
- In [24], we established a new framework for learning probabilistic context-free grammars for protein sequences using predicted or experimentally assessed amino acid 3D contacts. We relied on maximum-likelihood and contrastive estimators of parameters in this setting and an implementation for simple yet practical grammars. Tested on samples of protein motifs, grammars developed within the framework showed improved precision in recognition and higher fidelity to protein structures.

Metabolic pathway inference from non genomic data [*A. Belcour, M. Aite, J. Nicolas, A. Siegel, N. Théret, V. Dellannée, M. Conan*] We designed methods for the identification of metabolic pathways for which enzyme information is not precise enough.

- Heterocyclic Aromatic Amines (HAAs) are environmental and food contaminants classified as probable carcinogens. Our approach based on a refinement of molecular predictions with enzyme activity scores allowed us to accurately predict HAAs biotransformation and their potential DNA reactive compounds [13].
- We designed a prototype (*Pathmodel*) implementing inference methods to reconstruct biochemical reactions and metabolite structures to cope with metabolic pathway drift mechanisms. Using known metabolic pathways and metabolomics data, the tool infers alternative pathways compatible with the species known metabolites [29].

Large-scale eukaryotic metabolic network reconstruction [*A. Siegel, M. Chevallier, C. Frioux, M. Aite, J. Cambefort*] Metabolic network reconstruction has attained high standards but is still challenging for complex organisms such as eukaryotes.

- In this direction, we developed AuReMe for a flexible and reproducible reconstruction of these models. Together with a convenient mean for exploration through a local wiki, AuReMe is well-suited for the study of non-model organisms [12].
- In addition, a new gap-filling method satisfying the two main semantics of activation in metabolism is available. It enables to refine the models by pinpointing reactions such that metabolic objectives are met [15].

Systems ecology: design of microbial consortia [*C. Frioux, A. Siegel*]. Finding key elements among hundreds or thousands in microbiota to explain metabolic behaviours or prepare biological experimentations is a highly combinatorial problem. We introduced a two-step approach, MiSCoTo to screen the metabolic capabilities of microbiotas and exhaustively select members of interest by solving optimization problems with logic programming. We applied these methods to data from the Human Microbiome Project and a system composed of the Human metabolic network and 773 models for gut bacteria [14], [11].

7.3. Regulation and signaling: detecting complex and discriminant signatures of phenotypes

Participants: Catherine Belleannée, Samuel Blanquart, Célia Biane-Fourati, Nicolas Guillaudeux, Marine Louarn, Maxime Folschette, François Moreews, Anne Siegel, Nathalie Théret, Pierre Vignet, Méline Wery.

Comparative-genomics based prediction of non-model transcriptomes [*C. Belleannée, S. Blanquart, N. Guillaudeux*] In order to annotate the transcriptome of a non-model species, *Canis lupus familiaris*, we developed a method to predict whether or not a transcript known in a given species/gene could be expressed in an other species/gene. Exploiting knowledge in human, mouse and dog , we predicted a total of 7201 unknown yet transcripts and interpreted the evolutionary dynamics of gene's isoform sets. [30]

Signaling network identification [*M. Folschette, A. Siegel*] [22], [17]

- We introduced a new method to learn an interaction graph from the knowledge of its state space, without assumption on the semantics that was used to produce it. Proofs and characterizations are given for the synchronous, asynchronous and generalized semantics.
- We also used the caspo time-series software to integrate large-scale time series phosphoproteomic data (HPN-DREAM Breast Cancer challenge) into protein signaling networks and infer a family of Boolean Networks. The method highlights commonalities and discrepancies between the four cell lines.

Static analysis of ruled-based models [*P. Vignet, N. Théret*] We used a model of TGF-β to illustrate the main features of Kasa, a static analyzer for Kappa models. Kappa is a rule based language that describes systems of mechanistic interactions between proteins by the means of site-graph rewriting rules. The cornerstone of KaSa is a fix-point engine which detects some patterns that may never occur whatever the evolution of the system may be [18].

FLUMINANCE Project-Team

6. New Results

6.1. Fluid motion estimation

6.1.1. Stochastic uncertainty models for motion estimation

Participants: Musaab Khalid Osman Mohammed, Etienne Mémin.

This work is concerned with the design of motion estimation technique for image-based river velocimetry. The method proposed is based on an advection diffusion equation associated to the transport of large-scale quantity with a model of the unresolved small-scale contributions. Additionally, since there is no ground truth data for such type of image sequences, a new evaluation method to assess the results has been developed. It is based on trajectory reconstruction of few Lagrangian particles of interest and a direct comparison against their manually-reconstructed trajectories. The new motion estimation technique outperformed traditional optical flow and PIV-based methods used in hydrology [24]. This study has been performed within the PhD thesis of Musaab Khalid [18] and through a collaboration with the Irstea Lyon hydrology research group (HHLY).

6.1.2. 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. The technique was validated on a mixing layer in a wind tunnel for HFSB and smoke tracers [40] and applied on a laboratory fume-hood [39]. This study demonstrated the feasibility of conducting on-site large-scale image-based measurements for indoor airflows characterization.

6.1.3. 3D flows reconstruction from image data

Participants: Dominique Heitz, Etienne Mémin.

Our work focuses on the design of new tools for the estimation of 3D turbulent flow motion in the experimental setup of Tomo-PIV. This task includes both the study of physically-sound models on the observations and the fluid motion, and the design of low-complexity and accurate estimation algorithms. This year, we continued our investigation on the problem of efficient volume reconstruction. We have developed an 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 » (STB). The study developed within an Irstea post-doctoral funding introduced a novel approach originated from the data assimilation technique comprising a sampling-based optimal estimation algorithm, namely a group of ensemble-based filtering variational schemes. We found that employing such an ensemble-based optimal estimation method helped tackling the problems associated with STB: the inaccurate predictor and/or the robustness of the optimization procedure. The proposed method (ENS) was quantitatively evaluated with synthetic particle image data built by transporting virtual particles in a turbulent cylinder wake-flow at Reynolds number equal to 3900 [43]. The study will be continued within an Irstea doctoral funding.

6.2. Tracking, Data assimilation and model-data coupling

6.2.1. 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, Gilles Tissot.

In this axis of work, we explore 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 planes 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 [45]. We extended 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 is 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.

6.2.2. Ensemble variational data assimilation of large-scale dynamics with uncertainty

Participant: Etienne Mémin.

Estimating the parameters of geophysical dynamic models is an important task in Data Assimilation (DA) technique used for forecast initialization and reanalysis. In the past, most parameter estimation strategies were derived by state augmentation, yielding algorithms that are easy to implement but may exhibit convergence difficulties. The Expectation-Maximization (EM) algorithm is considered advantageous because it employs two iterative steps to estimate the model state and the model parameter separately. In this work, we propose a novel ensemble formulation of the Maximization step in EM that allows a direct optimal estimation of physical parameters using iterative methods for linear systems. This departs from current EM formulations that are only capable of dealing with additive model error structures. This contribution shows how the EM technique can be used for dynamics identification problem with a model error parameterized as arbitrary complex form. The proposed technique is used for the identification of stochastic subgrid terms that account for processes unresolved by a geophysical fluid model. This method, along with the augmented state technique, has been evaluated to estimate such subgrid terms through high resolution data. Compared to the augmented state technique, our method is shown to yield considerably more accurate parameters. In addition, in terms of prediction capacity, it leads to smaller generalization error as caused by the overfitting of the trained model on presented data and eventually better forecasts [27].

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

Participants: Dominique Heitz, Etienne Mémin, Gilles Tissot.

During the PhD thesis of Valentin Resseguier we have 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 as 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. 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 [15]. Note that these reduced order models can be extended to a full system of stochastic differential equations 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 as 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.

6.2.4. Estimation and control of amplifier flows

Participant: Gilles Tissot.

Estimation and control of fluid systems is an extremely hard problem. The use of models in combination with data is central to take advantage of all information we have on the system. Unfortunately all flows do not present the same physical and mathematical behaviour, thus using models and methodologies specialised to the flow physics is necessary to reach high performances.

A class of flows, denoted "oscillator flows", are characterised by unstable modes of the linearised operator. A consequence is the dominance of relatively regular oscillations associated with a nonlinear saturation. Despite the non-linear behaviour, associated structures and dynamical evolution are relatively easy to predict. Canonical configurations are the cylinder wake flow or the flow over an open cavity.

By opposition to that, "amplifier flows" are linearly stable with regard to the linearised operator. However, due to their convective nature, a wide range of perturbations are amplified in time and convected away such that it vanishes at long time. The consequence is the high sensitivity to perturbations and the broad band response that forbid a low rank representation. Jets and mixing layers show this behaviour and a wide range of industrial applications are affected by these broad band perturbations. It constitutes then a class of problems that are worth to treat separately since it is one of the scientific locks that render hard the transfer of methodologies existing in flow control and estimation to industrial applications.

There exists a type of models, that we will denote as "parabolised", that are able to efficiently represent amplifier flows. These models, such as parabolised stability equations and one-way Navier-Stokes propagate, in the frequency domain, hydrodynamic instability waves over a given turbulent mean flow. We can note that these models, by their structure, give access to a natural experimental implementation. They are an ingredient adapted to represent the system, but have a mathematical structure strongly different from the dynamical models classically used in control and data assimilation. It is then important to develop new methodologies of control, estimation and data assimilation with these models to reach our objectives. Moreover, inventing new models by introducing the modelling under location uncertainties in these parabolised models will be perfectly adapted to represent the evolution and the variability of an instability propagating within a turbulent flow. It will be consistent with actual postprocessing of experimental data performed in similar flow configurations.

6.3. Analysis and modeling of turbulent flows and geophysical flows

6.3.1. Geophysical flows modeling under location uncertainty

Participants: Werner Bauer, Long Li, Etienne Mémin.

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 [12], [13], [14]. 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 [22]. 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, we continue this effort. The goal is to propose relevant techniques to define or calibrate the noise term from data. In that prospect, we intend to explore statistical learning techniques.

6.3.2. 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 a flow over a circular cylinder at Re $\sim$ 3900 [21] and wall-bounded flows [45]. 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.

6.3.3. Variational principles for structure-preserving discretizations in stochastic fluid dynamics

Participants: Werner Bauer, Long Li, Etienne Mémin.

The overarching goal of this interdisciplinary project is to use variational principles to derive deterministic and stochastic models and corresponding accurate and efficient structure preserving discretizations and to use these schemes to obtain a deeper understanding of the principle conservation laws of stochastic fluid dynamics. The newly developed systematic discretization framework is based on discrete variational principles whose highly structured procedures shall be exploited to develop a general software framework that applies automatic code generation. This project, will first provide new stochastic fluid models and suitable approximations, with potential future applications in climate science using the developed methods to perform accurate long term simulations while quantifying the solutions' uncertainties. The generality of our approach addresses also other research areas such as electrodynamics (EDyn), magnetohydrodynamics (MHD), and plasma physics.

6.3.4. 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,carried 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).

Within a collaboration with L. Berselli (Univ. Pisa, Italy) we have achieved a study on the well known Bardina's turbulent model [20]. In this problem, we considered the Helmholtz filter usually used within the framework of Large Eddy Simulation. We carried out a similar analysis, by showing in particular that no singularity occurs for Bardina's model.

Within the same collaboration, we considered the three dimensional incompressible Navier-Stokes equations with non stationary source terms chosen in a suitable space. We proved the existence of Leray-Hopf weak solutions and that it is possible to characterize (up to sub-sequences) their long-time averages, which satisfy the Reynolds averaged equations, involving a Reynolds stress. Moreover, we showed that the turbulent dissipation is bounded by the sum of the Reynolds stress work and of the external turbulent fluxes, without any additional assumption, than that of dealing with Leray-Hopf weak solutions. Finally, we considered ensemble averages of solutions, associated with a set of different forces and we proved that the fluctuations continue to have a dissipative effect on the mean flow.

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 have tested the performances of an incompressible turbulence Reynolds-Averaged Navier-Stokes one-closure equation model in a boundary layer, which requires the determination of the mixing length ℓ. A series of direct numerical simulation have been performed, with flat and non trivial topographies, to obtain by interpolation a generic formula $\ell = \ell(Re_{\star}, z)$, $Re_{\star}$ being the frictional Reynolds number, and z the distance to the wall. Numerical simulations have been carried out at high Reynolds numbers with this turbulence model, in order to discuss its ability to properly reproduce the standard profiles observed in neutral boundary layers, and to assess its advantages, its disadvantages and its limits. We also proceeded to a mathematical analysis of the model.

6.3.5. Stochastic flow model to predict the mean velocity in wall bounded flows

Participants: Roger Lewandowski, Etienne Mémin, Benoit Pinier.

To date no satisfying model exists to explain the mean velocity profile within the whole turbulent layer of canonical wall bounded flows. We propose a modification of the velocity profile expression that ensues from the stochastic representation of fluid flows dynamics proposed recently in the group and referred to as "modeling under location uncertainty". This framework introduces in a rigorous way a subgrid term generalizing the eddy-viscosity assumption and an eddy-induced advection term resulting from turbulence inhomogeneity. This latter term gives rise to a theoretically well-grounded model for the transitional zone between the viscous sublayer and the turbulent sublayer. An expression of the small-scale velocity component is also provided in the viscous zone. Numerical assessment of the results have been performed for turbulent boundary layer flows, pipe flows and channel flows at various Reynolds numbers [49].

6.3.6. 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 carry 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 have been be used within a 4DVar assimilation technique, to reconstruct three-dimensional turbulent flows from data [45]. The third challenge was to carry out a time resolved tomoPIV experiments in a turbulent wake flow.

6.3.7. Fast 3D flow reconstruction from 2D cross-plane observations

Participants: Pranav Chandramouli, Dominique Heitz, Etienne Mémin.

We proposed a computationally efficient flow reconstruction technique, exploiting homogeneity in a given direction, to recreate three dimensional instantaneous turbulent velocity fields from snapshots of two dimension planar fields. This methodology, termed as 'snapshot optimisation' or SO, can help provide 3D data-sets for studies which are currently restricted by the limitations of experimental measurement techniques. The SO method aims at optimising the error between an inlet plane with a homogeneous direction and snap-shots, obtained over a sufficient period of time, on the observation plane. The observations are carried out on a plane perpendicular to the inlet plane with a shared edge normal to the homogeneity direction. The method is applicable to all flows which display a direction of homogeneity such as cylinder wake flows, channel flow, mixing layer, and jet (axi-symmetric). The ability of the method is assessed with two synthetic data-sets, and three experimental PIV data-sets. A good reconstruction of the large-scale structures are observed for all cases. This study has been recently submitted to the journal "Experiments in Fluids".

6.4. Visual servoing approach for fluid flow control

6.4.1. 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.

6.5. Coupled models in hydrogeology

6.5.1. Characterizations of Solutions in Geochemistry

Participants: Jocelyne Erhel, Tangi Migot.

We study the properties of a geochemical model involving aqueous and precipitation-dissolution reactions at a local equilibrium in a diluted solution. This model can be derived from the minimization of the free Gibbs energy subject to linear constraints. By using logarithmic variables, we define another minimization problem subject to different linear constraints with reduced size. The new objective function is strictly convex, so that uniqueness is straightforward. Moreover, existence conditions are directly related to the totals, which are the parameters in the mass balance equation. These results allow us to define a partition of the totals into mineral states, where a given subset of minerals are present. This precipitation diagram is inspired from thermodynamic diagrams where a phase depends on physical parameters. 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, supported by ANDRA, was presented at a workshop [29] and will be published in the journal Computational Geosciences [23].

6.5.2. Reactive transport in multiphase flow

Participants: Jocelyne Erhel, Bastien Hamlat.

Reactive transport modeling is a critical issue for energy industry, with application to carbon capture and storage (CCS) and geothermy. In this work, we focus on mathematical modeling and numerical solution of reactive transport in porous media. We study reaction kinetics leading to the appearance or disappearance of pure phases and present a mathematical model of PDEs with discontinuous right-hand sides. We propose a regularization process to obtain a PDE system with continuous right-hand sides which can be numerically solved. Using a simple academic case, we illustrate the benefits of our approach.

This work, supported by IFPEN, was presented at a workshop [30] and an international conference [35].

6.6. Sparse Linear solvers

6.6.1. Parallel GMRES

Participant: Jocelyne Erhel.

Sparse linear systems $Ax = b$ arise very often in computational science and engineering. Krylov methods are very efficient iterative methods, and restarted GMRES is a reference algorithm for non-symmetric systems. A first issue is to ensure a fast convergence, by preconditioning the system with a matrix M. Preconditioning must reduce the number of iterations, and be easy to solve. A second issue is to achieve high performance computing. The most time-consuming part in GMRES is to build an orthonormal basis V. With the Arnoldi process, many scalar products involve global communications. In order to avoid them, s-step methods have been designed to find a tradeoff between parallel performance and stability. Also, solving a system with the matrix M and multiplying a vector by the matrix A should be efficient. A domain decomposition approach involves mainly local communications and is frequently used. A coarse grid correction, based on deflation for example, improves convergence. These techniques can be combined to provide fast convergence and fast parallel algorithms. Numerical results illustrate various issues and achievements.

This work was presented at an international conference (invited talk) [28].

GENSCALE Project-Team

7. New Results

7.1. Data Structure

7.1.1. Quasi-dictionary data structure

Participants: Camille Marchet, Lolita Lecompte, 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. This work proposes a probabilistic data structure based on a minimal perfect hash function for indexing large sets of keys. Our structure out-compete 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 provide two applications based on similarity computation between collections of sequences, and for which this calculation is an expensive but required operation. In particular, we show a practical case in which other bioinformatics tools fail to scale up the tested data set or provide lower recall quality results.

The quasi-dictionary, is freely available at https://github.com/pierrepeterlongo/quasi_dictionary. The associated paper has been published in Discrete Applied Mathematics [17].

7.2. Algorithms & Methods

7.2.1. Genome assembly of targeted organisms in metagenomic data

Participants: Wesley Delage, Cervin Guyomar, Fabrice Legeai, Claire Lemaitre.

In this work, we propose a two-step reference-guided assembly method tailored for metagenomic data. First, a subset of the reads belonging to the species of interest are recruited by mapping and assembled *de novo* into backbone contigs using a classical assembler. Then an all-versus-all contig gap-filling is performed using a modified version of MindTheGap with the whole metagenomic dataset. The originality and success of the approach lie in this second step, that enables to assemble the missing regions between the backbone contigs, which may be regions absent or too divergent from the reference genome. The result of the method is a genome assembly graph in gfa format, accounting for the potential structural variations identified within the sample. We showed that this method is able to assemble the *Buchnera aphidicola* genome in a single contig in pea aphid metagenomic samples, even when using a divergent reference genome, it runs at least 5 times faster than classical *de novo* metagenomics assemblers and it is able to recover large structural variations co-existing in a sample. The modified version of MindTheGap is freely available at http://github.com/GATB/MindTheGap (version $>$ 2.1.0) [31].

7.2.2. De Novo Clustering of Long Reads by Gene from Transcriptomics Data

Participants: Camille Marchet, Lolita Lecompte, Jacques Nicolas, Pierre Peterlongo.

Long-read sequencing currently provides sequences of a few thousand base pairs. It is therefore possible to obtain complete transcripts, offering an unprecedented vision of the cellular transcriptome. However the literature lacks tools for *de novo* clustering of such data, in particular for Oxford Nanopore Technologies reads, because of the inherent high error rate compared to short reads. Our goal is to process reads from whole transcriptome sequencing data accurately and without a reference genome in order to reliably group reads coming from the same gene. This *de novo* approach is therefore particularly suitable for non-model species, but can also serve as a useful pre-processing step to improve read mapping. Our contribution both proposes a new algorithm adapted to clustering of reads by gene and a practical and free access tool that allows to scale the complete processing of eukaryotic transcriptomes. We sequenced a mouse RNA sample using the MinION device. This dataset is used to compare our solution to other algorithms used in the context of biological clustering. We demonstrate that it is the best approach for transcriptomics long reads. When a reference is available to enable mapping, we show that it stands as an alternative method that predicts complementary clusters.

The tool, called CARNAC-LR, is freely available at https://github.com/kamimrcht/CARNAC-LR. This work has been published in Nucleic Acids Research journal [16] and presented in several conferences [33], [28].

7.2.3. 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 multiple equivalent matches. 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 [11].

7.2.4. Short read correction

Participant: 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 kmer-spectrum techniques, while being scalable, making it possible to apply it to human-size genomic datasets and beyond [27].

7.2.5. Long read splitting of heterozygous genomes

Participants: Dominique Lavenier, Maxime Bridoux.

This study aims to directly split long reads of highly heterozygous genomes to help assembly. Long read technologies provide very noisy sequences with many short indel errors. Standard assembly software do not really make difference between heterozygosity and sequencing errors. For highly heterozygous genomes this confusion may lead to misassembly. To separate long reads accordingly to their haplotype, we developed a new k-mer based method. After an alignment step to group similar reads, we build slices of 1 kbp along the multiple alignment containing a representative number of reads. The splitting is done by focusing on k-mers that are absent in one group and not in another one. This in an ongoing work started by the internship of M. Bridoux [29] in the framework of the France Genomique ALPAGA project.

7.3. Optimisation

7.3.1. Distance-Constrained Elementary Path Problem

Participants: Sebastien François, Rumen Andonov.

Given a directed graph $G = (V, E, l)$ with weights $l_e \geq 0$ associated with arcs $e \in E$ and a set of vertex pairs with distances between them (called *distance constraints*), the problem is to find an elementary path in G that satisfies a maximum number of distance constraints. We call it *Distance-Constrained Elementary Path (DCEP)* problem. This problem is motivated by applications in genome assembly. We describe three Mixed Integer Programming (MIP) formulations for this problem and discuss their advantages [25].

7.3.2. Complete Assembly of Circular Genomes Based on Global Optimization

Participants: Sebastien François, Rumen Andonov, Dominique Lavenier.

The goal here is to develop a new methodology and tools based on strong mathematical foundations and novel optimization techniques for solving the genome assembly problem. During the current year we focused on the last two stages of genome assembly, namely scaffolding and gap-filling, and showed that they can be solved as part of a single optimization problem. We obtained this by modeling genome assembly as a problem of finding a simple path in a specific graph that satisfies as many as possible of the distance constraints encoding the insert-size information. We formulated it as a mixed-integer linear programming problem and applied an optimization solver to find the exact solution on a benchmark of chloroplasts. Our tool is called GAT (Genscale Assembly Tool) and we tested it on a set of 33 chloroplast genome data. Comparisons with some of the most popular recent assemblers show that our tool produces assemblies of significantly higher quality than these heuristics [26]. These results fully justify the efforts for designing exact approaches for genome assembly.

7.4. Parallelism

7.4.1. Variant detection using processing-in-memory technology

Participants: Dominique Lavenier, Mohamed Moselhy.

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 parallelized the detection of small mutations on the human genome. Traditionally, 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. In 2018, we define an ad-hoc data structure allowing the variant calling step to be performed simultaneously on the host processor. Basically, the two steps are overlapped in such a way that reads are mapped by packet. When a packet is mapped, the mapping results of the previous one dynamically feed the variant calling data structure. Performance evaluation on the FPGA UPMEM memory prototype indicates a very high speed-up (two orders of magnitude) compared with state-of-the-art software (specifically GATK).

7.5. Benchmarks and Reviews

7.5.1. Evaluation of error correction tools for long Reads

Participants: Lolita Lecompte, Camille Marchet, Pierre Peterlongo.

Long read technologies, Pacific Biosciences and Oxford Nanopore, have high error rates (from 9% to 30%). Hence, numerous error correction methods have been recently proposed, each based on different approaches and, thus, providing different results. As this is important to assess the correction stage for downstream analyses, we designed the ELECTOR software, providing evaluation of long read correction methods. This software generates additional quality metrics compared to previous existing tools. It also scales to very long reads and large datasets and is compatible with a wide range of state-of-the-art error correction tools.

ELECTOR is freely available at https://github.com/kamimrcht/ELECTOR. It as been presented during the Jobim2018 conference [32]

7.5.2. Computational pan-genomics: status, promises and challenges

Participant: Pierre Peterlongo.

We took part to the redaction of the review paper that proposes a state of the art of the pan-genomics current status, methods and future orientations. This paper has been published in *Briefings in Bioinformatics* [18].

7.6. Bioinformatics Analysis

7.6.1. Metagenomic analysis of pea aphid symbiotic communities

Participants: Cervin Guyomar, Fabrice Legeai, Claire Lemaitre.

We worked on a methodological 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, and within individual aphids. Thorough phylogenomic analyses highlighted that the evolutionary dynamics of symbiotic associations strongly varied depending on the symbiont, reflecting different evolutionary histories and possible constraints [14].

7.6.2. Analysis of pea aphid genomic polymorphism

Participants: Fabrice Legeai, Claire Lemaitre.

We participated in the analyses of a large re-sequencing dataset of pea aphid individuals and populations. We performed the data cleaning, mapping to the reference genome and variant calling steps. The resulting polymorphism data shed light on two novel findings regarding the pea aphid genome evolution.

First, we showed that relaxed selection is likely to be the greatest contributor to the faster evolution of the X chromosome compared to autosomes [15]. Secondly, we looked for genomic bases of adaptation to novel environments, and identified 392 genomic hotspot regions of differentiation spanning 47.3 Mb and 2,484 genes. Interestingly, these hotspots were significantly enriched for candidate gene categories that are related to host–plant selection and use. These genes represent promising candidates for the genetic basis of host–plant specialization and ecological isolation in the pea aphid complex [21].

7.6.3. A de novo approach to disentangle partner identity and function in holobiont systems

Participants: Camille Marchet, Pierre Peterlongo.

Study of meta-transcriptomic datasets involving non-model organisms represents bioinformatic challenges that affect the study of holobiont meta-transcriptomes. Hence, we proposed an innovative bioinformatic approach and tested it on marine models as a proof of concept.

We considered three holobiont models, of which two transcriptomes were previously published and a yet unpublished transcriptome, to analyze and sort their raw reads using Short Read Connector (see section 7.1.1). Before assembly, we thus defined four distinct categories for each holobiont meta-transcriptome: host reads, symbiont reads, shared reads, and unassigned reads. Afterwards, we observed that independent *de novo* assemblies for each category led to a diminution of the number of chimeras compared to classical assembly methods. Moreover, the separation of each partner's transcriptome offered the independent and comparative exploration of their functional diversity in the holobiont. Finally, our strategy allowed to propose new functional annotations for two well-studied holobionts (a Cnidaria-Dinophyta, a Porifera-Bacteria) and a first meta-transcriptome from a planktonic Radiolaria-Dinophyta system forming widespread symbiotic association for which our knowledge is considerably limited [19].

7.6.4. Whole genome detection of micro-satellites

Participant: Dominique Lavenier.

This study has been done in cooperation with the federal university of de São João del-Rei, Brazil. The objective was to locate tens of thousands of micro-satellite loci for an endangered piracema (i.e. migratory) South American fish, *Brycon orbignyanus*. Together with the Brazil group we designed a specific pipeline that first assembles short paired-end reads into contigs and then performs micro-satellite oriented scaffolding processing [23].

7.6.5. Analysis of the genes and genomes involved in plant and insects interactions

Participant: Fabrice Legeai.

These study has been done in cooperation with various laboratories. In particular, we characterized the effectors (secreted proteins suppressing plant defense) of the pea aphid fed on different plants, by firstly identifying these genes in the pea aphid genome, then studying and comparing their expression between different conditions, and then finally by observing their evolution among a broad set of phytophagous insects [10]. We also identified microRNAs from smallRNA datasets from *Spodoptera frugiperda* strains fed on different host-plants [20]. Finally, we predicted the transposable elements in the genome of *Cephus cinctus*, an important insect pest [22].

7.6.6. Analysis of the expression and identification of the targets of mir202 during the medaka oogenesis

Participant: Fabrice Legeai.

This study has been done in cooperation with the INRA LPGP laboratory (Rennes). Its goal was to identify the role of small non-coding RNAs in the regulation of the reproduction of the fish model *Oryzias latipes* (medaka). We predicted the putative targets of the microRNA miR202, already observed as being specifically expressed in gonads. In the second part of the work, we identified important genes and functions targeted by miR202 and differentially expressed in the gonads when the microRNA was artificially repressed [13].

SERPICO Project-Team

7. New Results

7.1. A Monte-Carlo approach for missing wedge restoration in cryo-electron tomography

Participants: Emmanuel Moebel, Charles Kervrann.

We investigated a Monte-Carlo approach 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 computational method tackles the restoration problem by filling up the MW by iterating the two 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. Also, specific constraint is added in the spectral domain by imposing the known Fourier coefficients to be unchanged through iterations. We justified this approach in the Monte-Carlo simulation and Bayesian framework. In practice, different denoising algorithms (BM3D, NL-Bayes, NL-means...) can be applied. In our experiments, several transforms have been tested in order to apply the constraint (Fourier transform, Cosine transform, pseudo-polar Fourier transform). Convincing results have been achieved (see Fig. 3) using the Fourier Shell Correlation (FSC) as an evaluation metric.

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

7.2. Algorithms for deconvolving and denoising fluorescence 2D-3D microsocopy images

Participants: Hoai-Nam Nguyen, Charles Kervrann.

In this work, we proposed a restoration method for 2D and 3D fluorescence imaging using a novel family of convex regularizers. The proposed regularization functionals are based on the concept of sparse variation, that consists in penalizing jointly the image intensity and gradient at each pixel to favor the co-localization of non-zero intensities and gradients, by considering eventually higher-order differentiation operators. By construction, these regularizers possess interesting mathematical properties, namely convexity, invariance to scale, rotation, and translation as the well-known total variation regularization approach. It enables therefore to design efficient algorithms to solve the underlying deconvolution problem, which is in general large-scale in the context of fluorescence microscopy. We reformulated denoising or deconvolution (given the point spread function) as a minimization problem of a convex energy function composed of a quadratic data fidelity term and a sparse-variation-based regularity term under the constraint of positivity and maximum intensity value. In order to minimize this energy, we considered a primal-dual (proximal) algorithm based on the full splitting technique, which only involves first-order operators to cope with the large-scale nature of the problem. Experimental results on both 2D and 3D synthetic and real fluorescence images demonstrated that our method was able to produce very competitive deconvolution results, when compared to several competing methods such as the Schatten norm of the Hessian matrix, in terms of quantitative performance as well as visual quality and computational time. The method is able to process a 512×512 image in 250 ms (in Matlab) with a non optimized implementation and can process 3D images in a few minutes (with no code optimization technique and multithreading).

Collaborators: Cyril Cauchois (Innopsys company).

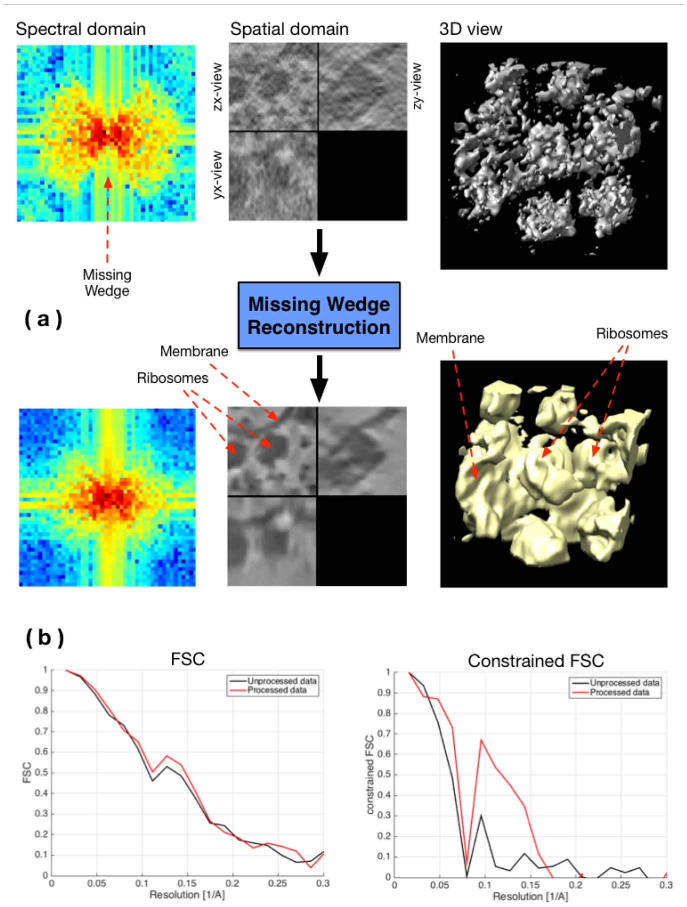


Figure 3. Experimental sub-tomogram containing ribosomes attached to a membrane. (a) Top row: original data in the spectral (left) and spatial (middle) domains and 3D view of the thresholded data (right). Bottom row: denoised data shown as above. (b) FSC and constrained FSC measures of the method input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

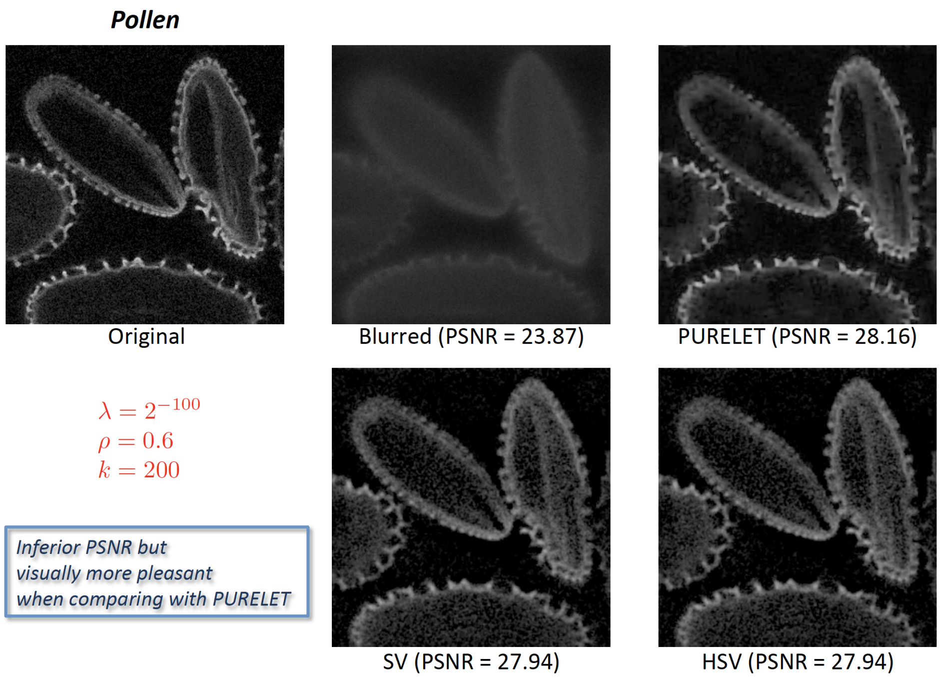


Figure 4. Comparison (2D view) of sparse deconvolution (SparsEvolution) method with Purelet method [42] on the artificial blurred and noisy (Gaussian-Poisson noise) 3D pollen image ($256 \times 256 \times 32$) described in [42].

7.3. Colocalization and co-orientation in fluorescence imaging

Participant: Charles Kervrann.

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. It enables to quantify the protein-protein interactions inside the cell, just at the resolution limit of the microscope. 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 or volume) between the two (2D or 3D) binary segmented images. In the context of super-localization imaging, we exploit the localization uncertainty of molecules to generate two input binary images. A the end, GcoPS handles 2D and 3D images, variable signal-to-noise ratios and any fluorescence image pair acquired with conventional or super-resolution microscopy. 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, we adapted the GcoPS framework to analyze the spatiotemporal molecular interactions from a set of 3D computed trajectories or motion vector fields (e.g., co-alignment), and then to fully quantify specific molecular machineries.

Collaborators: Frédéric Lavancier (University of Nantes, Laboratoire de Mathématiques Jean Leray),
Thierry Pécot (Hollings Cancer Center at the Medical University of South Carolina),
Jean Salamero and Liu Zengzhen (UMR 144 CNRS-Institut Curie).

7.4. Detection of transitions between diffusion models along biomolecule trajectories

Participants: Antoine Salomon, Charles Kervrann.

Recent advances in molecular biology and fluorescence microscopy imaging have made possible the inference of the dynamics of single molecules in living cells. When we observe a long trajectory (more than 100 points), it is possible that the particle switches mode of motion over time. Then, a goal is to estimate the temporal change-points, that is the distances at which a change of dynamics occurs. To address this issue, we proposed a non-parametric procedure based on test statistics [16], computed on local windows along the trajectory, to detect the change-points. This algorithm controls the number of false change-point detections in the case where the trajectory is fully Brownian. Our algorithm is user-friendly as there is only one parameter to tune, namely the sliding window size. A Monte Carlo study is proposed to demonstrate the performances of the method and also to compare the procedure to two competitive algorithms. Our method is much faster than previous methods which is an advantage when dealing with a large numbers of trajectories. With this computational approach, we analyzed real data depicting neuronal mRNPs (mRNAs in complex with mRNA-binding), and another very complex biological example, Gal-3 trafficking from the plasma membrane to different cellular compartments (acquired with Lattice Light Sheet microscopy). The analysis of multiple Gal-3 trajectories demonstrates nicely that there is not one typical signature. Biological trafficking events are very multifaceted. The algorithm was capable of identifying and characterizing the multistep biological movement, switching several times between subdiffusive, superdiffusive and Brownian motion.

Collaborators: Vincent Briane (UNSW Sydney, School of Medical Sciences, Australia),
Myriam Vimond (CREST ENSAI Rennes),
C.A. Valades Cruz and C. Wunder, (Institut Curie, PSL Research University, Cellular and Chemical Biology, U1143 INSERM / UMR 3666 CNRS).

7.5. Intracellular drift and diffusion coefficient estimation: a trajectory label-based approach

Participants: Antoine Salomon, Charles Kervrann.

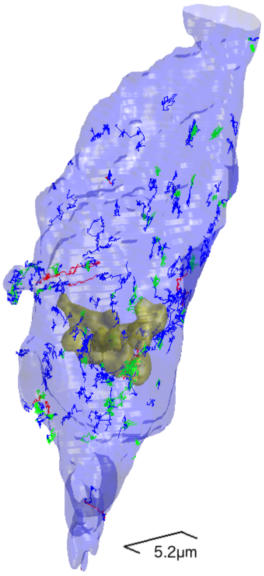


Figure 5. Change point detection over a set of three-dimensional trajectories of Galectin-3 proteins observed in HeLa cells with a Lattice Light Sheet microscope. The blue parts correspond to Brownian portions of the trajectory, red parts to superdiffusive portions, green parts to the subdiffusive portion. In light blue we plot the cell membrane and in yellow the Golgi apparatus for structural orientation.

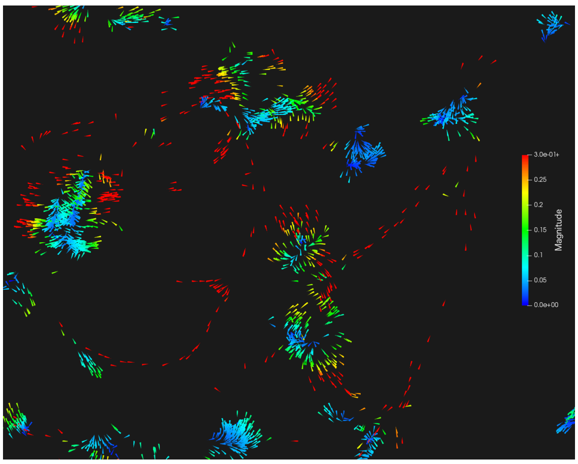


Figure 6. Estimation of a 2D drift field in 2D-TIRF microscopy of exocytosis biomolecules.

Dedicated computational methods for intracellular drift and diffusion map estimation rely on a scanning approach, using a sliding window (or cube, in the 3D case) in which all elementary particle movements taken from the trajectories inside the window/box are averaged [41]. This approach lacks precision, and a huge amount of trajectory data is needed to obtain significative results. In our new approach, we currently investigate several high-level features to obtain more satisfying results even from a small amount of data. First, we exploit classification of sub-trajectories (Brownian motion, superdiffusion, subdiffusion) obtained in the Lagrangian setting [16]. This classification is used to select trajectories of the same type in a local region and to guide weighting averaging. It allows us to calculate the drift separately on each type of movement, which avoids confusion between Brownian, confined and directed motions (see Fig. 6). Furthermore, the calculation is efficiently performed on trajectory-sliding kernels instead of scanning windows to save computing time in 2D and 3D. We considered square-box (sliding window), circle-box, cone-shaped, and Gaussian-shaped kernels.

Collaborators: Vincent Briane (UNSW Sydney, School of Medical Sciences, Australia),
Myriam Vimond (CREST ENSAI Rennes),
C.A. Valades Cruz and C. Wunder, (Institut Curie, PSL Research University, Cellular and Chemical Biology, U1143 INSERM / UMR 3666 CNRS).

7.6. 3D flow estimation in 3D fluorescence microscopy image sequences

Participants: Sandeep Manandhar, Patrick Bouthemy, Charles Kervrann.

Three-dimensional (3D) motion estimation for light-sheet microscopy is challenged by the heterogeneous scales and nature of intracellular dynamics. As typical examples in cell imaging, blebbing of a cell has small motion magnitude, while cell migration may show large displacement between frames. To tackle this problem, we have designed a two-stage 3D optical flow method. The first stage involves an extension of two-dimensional PatchMatch paradigm to 3D data, which in addition operates in a coarse-to-fine manner. We exploit multiple spatial scales to explore the possible range of intracellular motions. Our findings show that the metric based on Census transform is more robust to noise and to intensity variation between time steps. Only discrete displacements are estimated in this stage. Then, in a second stage, a 3D variational method enables to recover a sub-voxel dense 3D flow map. The variational approach still involves a data fidelity term based on the Census transform. The combination of the 3D PatchMatch and the 3D variational method is able to capture both large and small displacements. We assessed the performance of our method on data acquired with two different light sheet microscopes and compared it with a couple of other methods. The dataset depicts blebbing and migration of MV3 melanoma cells, and collagen network displacement induced by cell motility. As seen in Fig. 7 , our method is able to successfully estimate various range of motion during cell migration and blebbing. A straightforward way to visualize the resulting 3D flow field is to use 3D glyphs (arrows), which represent vector direction and magnitude. However, it may not lead to easy understanding for visualization in 3D and over time. Consequently, we propose to visualize 2D projections of 3D flow field in 3 orthogonal planes (see Figure 7 a.1.2 and b.1.2). Using the standard Middlebury-style color coding for 2D optic flow, the motion field becomes easier to understand. This work is carried out in collaboration with UTSW Dallas int the frame of the Inria associated team CytoDI.

Collaborators: Philippe Roudot and Erik Welf (UTSW, Dallas, USA).

7.7. Connecting trajectories in TIRF images of rod-shaped bacteria

Participants: Yunjiao Lu, Charles Kervrann.

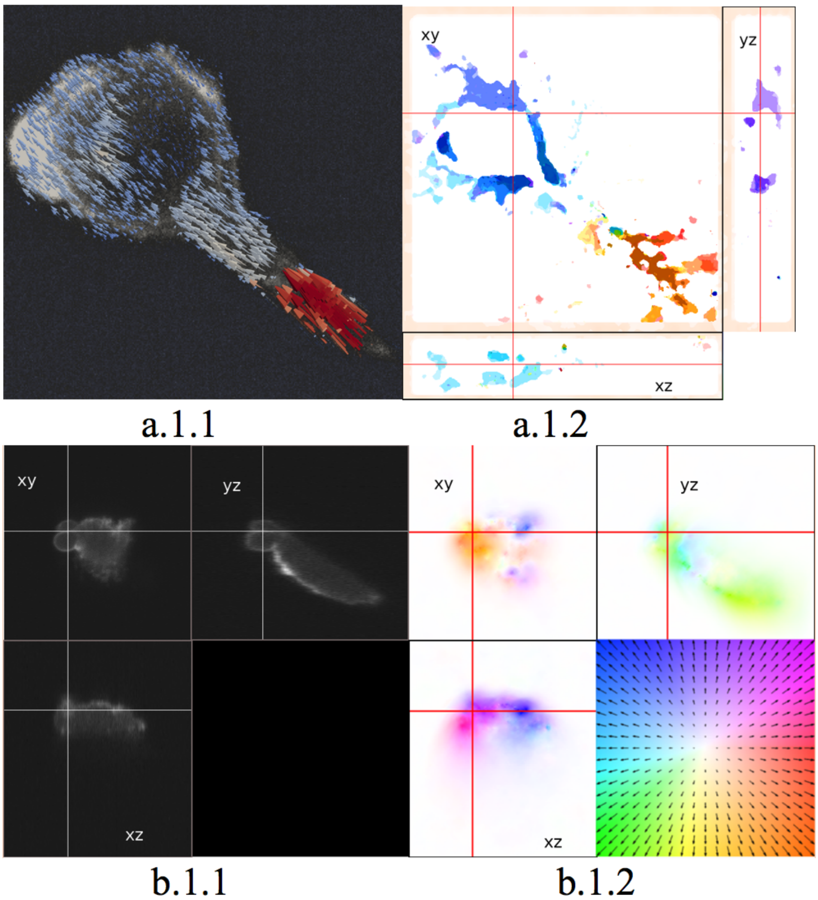


Figure 7. (a) Migration of MV3 melanoma cell in collagen; (a.1.1) 3D flow field in a slice of cell data represented with glyphs. Larger motion magnitude is coded in warm colors and smaller one in cold colors. (a.1.2) Motion map of collagen channel in 3 orthogonal planes (see b.1.2 for color code). (b) Blebbing of MV3 cell in a cover slip; (b.1.1) 3 orthogonal planes of cell data. (b.1.2) 3D computed flow field projected in 3 orthogonal planes. The motion map is color-coded as shown in lower-right corner of b.1.2. (input images by courtesy of Danuser lab, UTSW Dallas, USA).

In this work, we investigate the role of MreB protein involved in the cell wall construction in rod-shaped bacteria Bacillus. Previous work concerning the quantification of dynamics of MreB is performed from 2D TIRFM image sequences, ignoring the curvature of the rod-shaped cell wall and the thickness of TIRFM plane. To evaluate the effect of these approximations, we have developed a simulator to generate the trajectories on the surface of the cylinder. We assume that trajectories are modeled as:

$$dX = b(X)dt + \sqrt{2}B(X)dW,$$

where $b(X)$ is the drift tangent to the surface, W is a white noise, and $\sigma(X) = \frac{1}{2}B(X)B(X)^T$ is the diffusion tensor assumed to be isotropic. In our approach, the drift and diffusion tensor on the surface and on the projected plane are estimated respectively. Moreover, we propose to extrapolate the drift and diffusion tensor to the hidden region, to reconnect the segments we observe in the visible region, and then, to recover the entire trajectory on the surface of the cylinder in 3D (see Fig. 8).

Collaborators: Alain Trubuil and Pierre Hodara (INRA UR MAIAGE, Jouy-en-Josas),
Rut Carbllido-López and Cyrille Billaudeau (INRA, UR MICALIS, Jouy-en-Josas).

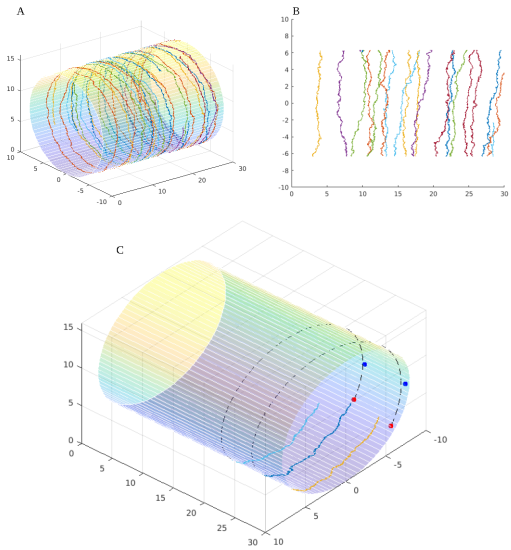


Figure 8. (A): Trajectories generated on the surface of the cylinder and (B): the view of TIRF microscope which is the projection of the dynamics on the cylinder near the support surface. The unity in the two figures is in pixel and in our TIRF images 1 pixel $\approx$ 64 nm. As the theoretical thickness of TIRF is 200 nm, the trajectories whose z coordinate is under 3.125 pixels are projected onto $x - y$ plan. Colors for different trajectories are random. (C): The connection result through the hidden region of three segments (light blue, dark blue and yellow in order of passage). Red points represent the extrapolated points using drift and diffusion tensor of segments on the surface, while blue points are the extrapolated points using projected drift and diffusion tensor on 2D plan.

7.8. 3D registration for correlative light-electron microscopy

Participants: Bertha Mayela Toledo Acosta, Patrick Bouthemy.

Correlative light and electron microscopy (CLEM) enables the study of cells and subcellular elements in complementary ways by combining information on the dynamics and on the structure of the cell, provided a reliable registration between light microscopy (LM) and electron microscopy (EM) images is efficiently achieved. We have developed a general automatic registration method. Due to large discrepancies in appearance, field-of-view, resolution and position, a pre-alignment stage is required before any 3D fine registration stage. 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 candidates, we estimate, using 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 carried out experimental results on different real datasets of 3D correlative microscopy, demonstrating computational efficiency and overlay accuracy.

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

7.9. 3D Convolutional Neural Networks for macromolecule localization in cryo-electron tomograms of intact cells

Participants: Emmanuel Moebel, Charles Kervrann.

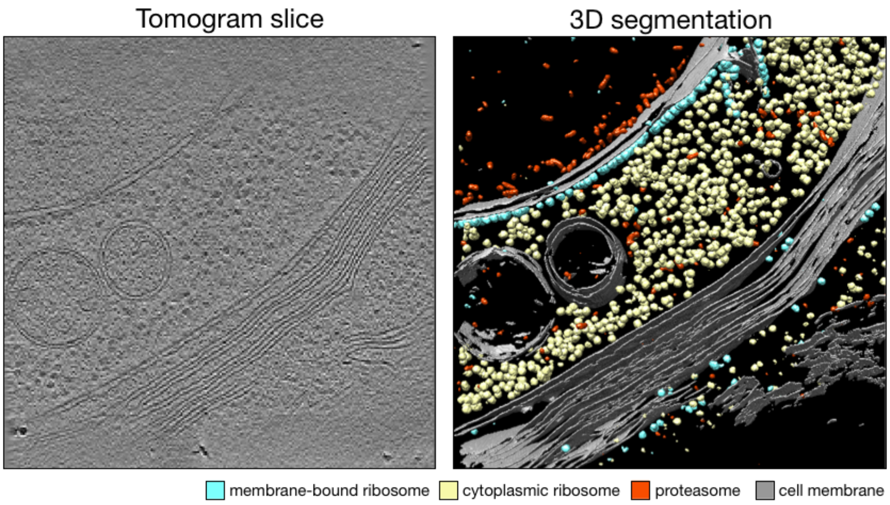


Figure 9. Experimental tomogram of a Chlamydomonas Reinhardtii cell (central slice) and the 3D segmentation obtained by our convolutional neural network. The network is able to identify 3 macromolecule classes (membrane-bound ribosome, cytoplasmic ribosome and proteasome) as well as the cell membrane (tomogram by courtesy of Max-Planck Institute, Martinsried, Germany).

In this study, we focus on macromolecule localization and classification in cryo-electron tomography images. Cryo-electron tomography (cryo-ET) allows one to capture 3D images of cells in a close to native state, at sub-nanometer resolution. However, noise and artifact levels are such that heavy computational processing is needed to access the image content. We propose a deep learning framework to accurately and jointly localize multiple types and states of macromolecules in cellular cryo-electron tomograms. We compare this framework to the commonly-used template matching method on both synthetic and experimental data. On synthetic image data, we show that our framework is very fast and produces superior detection results. On experimental data, the detection results obtained by our method correspond to an overlap rate of 86% with the expert annotations, and comparable resolution is achieved when applying subtomogram averaging. In addition, we show that our method can be combined to template matching procedures to reliably increase the number of expected detections. In our experiments, this strategy was able to find additional 20.5% membrane-bound ribosomes that were missed or discarded during manual annotation.

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

7.10. Data assimilation and modeling of cell division mechanism

Participants: Ancageorgiana Caranfil, Charles Kervrann.

In this work, we focus on the dynamics of the spindle during cell division mechanism. 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. 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 behavior is led by the spindle poles behavior. We thus proposed a new biophysical model for the spindle pole functioning during anaphase, that explains the oscillatory behavior with a minimum number of parameters. By mathematically analyzing our model, we confirmed some previous findings, such as the existence of a threshold number of active force-generator motors required for the onset of oscillations. We also confirmed that the monotonic increase of motor activity accounts for their build-up and die-down. We determined boundaries for the motor activity-related parameters for these oscillations to happen. This also allowed us to describe the influence of the number of motors, as well as physical parameters related to viscosity or string-like forces, on features such as the amplitude and frequency of oscillations. Lastly, by using a Bayesian approach to confront our model to experimental data, we were able to estimate distributions for our biological and biophysical parameters. A statistical reduction model approach was preliminary applied to select the most influential model parameters. These results give us insights on variations in spindle behavior during anaphase in asymmetric division, and provide means of prediction for phenotypes related to misguided asymmetric division. Data assimilation will be further used to properly combine the information given by our model and the experimental data.

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

7.11. Motion saliency in videos

Participants: Léo Maczyta, Patrick Bouthemy.

The problem we have addressed appertains to the domain of motion saliency in videos. More specifically, we aim to extract the temporal segments of the video where motion saliency is present. It is a prerequisite for computing motion saliency maps in relevant images. It turns out to be a frame classification problem. A frame is classified as dynamically salient if it contains local motion departing from its context. Temporal motion saliency detection is relevant for applications where one needs to trigger alerts or to monitor dynamic behaviors from videos. The proposed approach handles situations with a mobile camera, and involves a deep learning classification framework after camera motion compensation. We have designed and compared two methods respectively based on image warping, and on residual flow. A baseline that relies on a two-stream network to process temporal and spatial information, but that does not use camera motion compensation, was also defined. Experiments on real videos demonstrate that we can obtain an accurate classification in highly challenging situations, and get significant improvement over the baseline. In particular, we showed that the compensation of the camera motion produces a better classification. We also showed that for the limited training data available, providing the residual flow as input to the classification network produces better results than providing the warped images.

Collaborators: Olivier Le Meur (Percept team, Irisa).

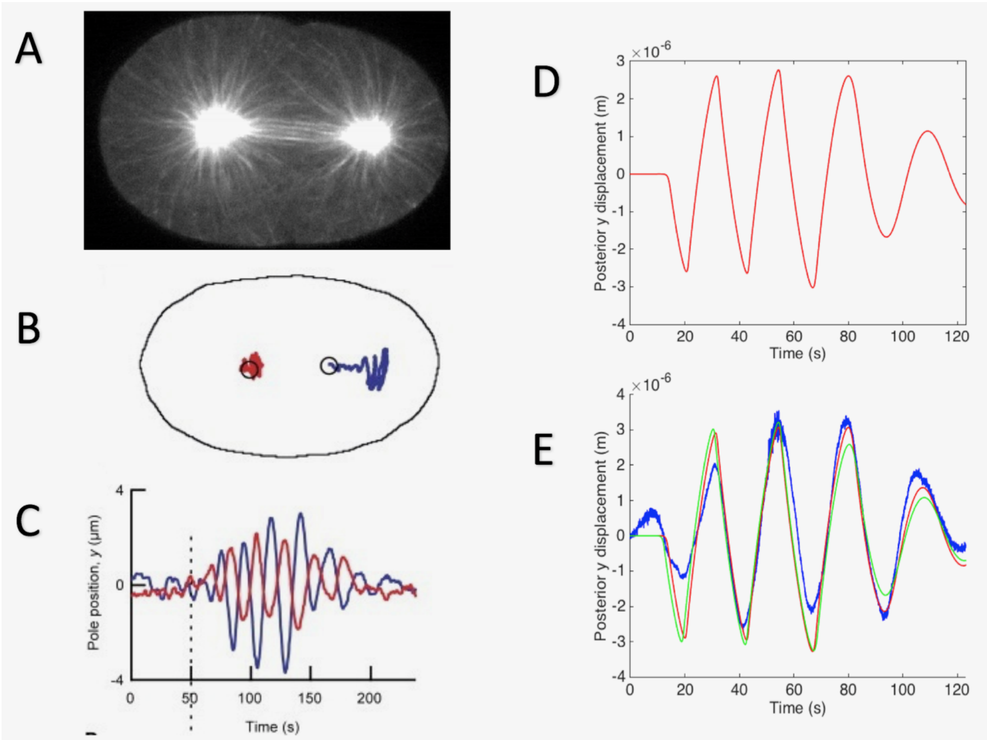


Figure 10. A. Cell division observed by fluorescence microscopy. B. and C. Tracking of the two spindle pole (anterior in red and posterior in blue). Oscillations of the poles during metaphase and anaphase. D. Simulation of oscillations using our model. E. Fitting of experimental data (in blue), and the two estimated curves with our method (minimum estimator in red, and mean estimator in green).

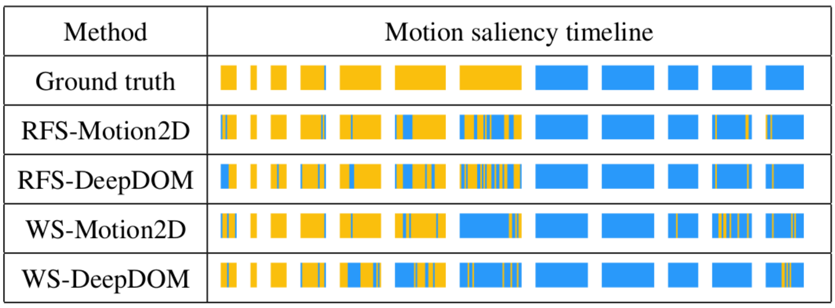


Figure 11. Timeline of the classification results supplied by the methods, for 12 videos of the testing set. Orange denotes salient frames, and blue non salient frames. Videos are separated by horizontal spaces. WS and RFS denote respectively the use of image warping and or residual flow in order to predict saliency. For dominant parametric motion estimation, the robust classical method Motion2D and the neural network DeepDOM were used.

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. Optimal selection of diffusion-weighting gradient waveforms using compressed sensing and dictionary learning

Participants: Raphaël Truffet, Emmanuel Caruyer.

Acquisition sequences in diffusion MRI rely on the use of time-dependent magnetic field gradients. Each gradient waveform encodes a diffusion weighted measure; a large number of such measurements are necessary for the in vivo reconstruction of microstructure parameters. We proposed here a method to select only a subset of the measurements while being able to predict the unseen data using compressed sensing. We learnt a dictionary using a training dataset generated with Monte-Carlo simulations; we then compare two different heuristics to select the measures to use for the prediction. We found that an undersampling strategy limiting the redundancy of the measures allows for a more accurate reconstruction when compared with random undersampling with similar sampling rate [57].

7.1.1.2. A Bayes Hilbert Space for Compartment Model Computing in Diffusion MRI

Participant: O. Commowick.

The single diffusion tensor model for mapping the brain white matter microstructure has long been criticized as providing sensitive yet non-specific clinical biomarkers for neurodegenerative diseases because (i) voxels in diffusion images actually contain more than one homogeneous tissue population and (ii) diffusion in a single homogeneous tissue can be non-Gaussian. Analytic models for compartmental diffusion signals have thus naturally emerged but there is surprisingly little for processing such images (estimation, smoothing, registration, atlas-ing, statistical analysis). We propose to embed these signals into a Bayes Hilbert space that we properly define and motivate. This provides a unified framework for compartment diffusion image computing. Experiments show that (i) interpolation in Bayes space features improved robustness to noise compared to the widely used log-Euclidean space for tensors and (ii) it is possible to trace complex key pathways such as the pyramidal tract using basic deterministic tractography thanks to the combined use of Bayes interpolation and multi-compartment diffusion models [26]

This work was done in collaboration with A. Stamm, A. Menafoglio and S.K. Warfield.

7.1.2. Arterial Spin Labeling

7.1.2.1. Patch-Based Super-Resolution of Arterial Spin Labeling Magnetic Resonance Images

Participants: Cédric Meurée, Pierre Maurel, Jean-Christophe Ferré, Christian Barillot.

Arterial spin labeling is a magnetic resonance perfusion imaging technique that, while providing results comparable to methods currently considered as more standard concerning the quantification of the cerebral blood flow, is subject to limitations related to its low signal-to-noise ratio and low resolution. In this work, we investigated the relevance of using a non-local patch-based super-resolution method driven by a high-resolution structural image to increase the level of details in arterial spin labeling images. This method was evaluated by comparison with other image dimension increasing techniques on a simulated dataset, on images of healthy subjects and on images of subjects diagnosed with brain tumors, who had a dynamic susceptibility contrast acquisition. The influence of an increase of ASL images resolution on partial volume effects was also investigated in this work. [56]

The development of this super-resolution algorithm in the context of a thesis financed by Siemens Healthineers conducted to a stay of one month of the PhD candidate in Erlangen, during summer 2018. This immersion into the neuro-development team allowed to integrate the proposed solution with tools in use within this team. Part of the work also consisted in reducing the calculation time, a factor of 5 being achieved at the end of these four weeks.

7.1.2.2. Resting-state ASL : Toward an optimal sequence duration

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

Resting-state functional Arterial Spin Labeling (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 leads to significant clinical subject scaled application as CBF can be considered as a biomarker in common neuropathology. Our work here focused on the link between overall quality of rs-fASL and duration of acquisition. To this end, we consider subject self-Default Mode Network (DMN), and assess DMN quality depletion compared to a gold standard DMN depending on the duration of acquisition [46].

7.1.3. Quantitative imaging

7.1.3.1. Identification of Gadolinium contrast enhanced regions in MS lesions using brain tissue microstructure information obtained from diffusion and T2 relaxometry MRI

Participants: S. Chatterjee, O. Commowick, C. Barillot.

A multiple sclerosis (MS) lesion at an early stage undergoes active blood brain barrier (BBB) breakdown. Identifying MS lesions in a patient which are undergoing active BBB breakdown is of critical importance for MS burden evaluation and treatment planning. However in non-contrast enhanced structural magnetic resonance imaging (MRI) the regions of the lesion undergoing active BBB breakdown cannot be distinguished from the other parts of the lesion. Hence gadolinium (Gd) contrast enhanced T1-weighted MR images are used for this task. However some side effects of Gd injection into patients have been increasingly reported recently. The BBB breakdown is reflected by the condition of tissue microstructure such as increased inflammation, presence of higher extra-cellular matter and debris. We have thus proposed a framework to predict enhancing regions in MS lesions using tissue microstructure information derived from T2 relaxometry and diffusion MRI (dMRI) multicompartment models. We showed that combination of the dMRI and T2 relaxometry microstructure information can distinguish the Gd enhancing lesion regions from the other regions in MS lesions [23].

7.1.3.2. A three year follow-up study of gadolinium enhanced and non-enhanced regions in multiple sclerosis lesions using a multi-compartment T_2 relaxometry model

Participants: S. Chatterjee, O. Commowick, B. Combes, C. Barillot.

Demyelination, axonal damage and inflammation are critical indicators of the onset and progress of neurodegenerative diseases such as multiple sclerosis (MS) in patients. Due to physical limitations of imaging such as acquisition time and imaging resolution, a voxel in a MR image is heterogeneous in terms of tissue microstructure such as myelin, axons, intra and extra cellular fluids and free water. We present a multi-compartment tissue model which estimates the water fraction (WF) of tissues with short, medium and high T_2 relaxation times in a T_2 relaxometry MRI voxel. The proposed method is validated on test-retest data of healthy controls. This model was then used to study longitudinal trends of the tissue microstructures for two sub-regions of the lesions: gadolinium enhanced (E+) and non-enhanced (L-) regions of MS lesions in 10 MS patients over a period of three years. The water fraction values in E+ and L- regions were found to be significantly different ($p < 0.05$) over the period of first three months. The results of this study also showed that the estimates of the proposed T_2 relaxometry model on brain tissue microstructures have potential to distinguish between regions undergoing active blood brain barrier breakdown from the other regions of the lesion [49].

This work was done in collaboration with Onur Afacan and Simon K. Warfield from Harvard Medical School.

7.1.3.3. Multi-Compartment Model of Brain Tissues from T2 Relaxometry MRI Using Gamma Distribution

Participants: S. Chatterjee, O. Commowick, C. Barillot.

The brain microstructure, especially myelinated axons and free fluids, may provide useful insight into brain neurodegenerative diseases such as multiple sclerosis (MS). These may be distinguished based on their transverse relaxation times which can be measured using T_2 relaxometry MRI. However, due to physical limitations on achievable resolution, each voxel contains a combination of these tissues, rendering the estimation complex. We presented a novel multi-compartment T_2 (MCT2) estimation based on variable projection, applicable to any MCT2 microstructure model. We derived this estimation for a three-gamma distribution model. We validated our framework on synthetic data and illustrated its potential on healthy volunteer and MS patient data [32].

This work was done in collaboration with Onur Afacn and Simon K. Warfield.

7.1.3.4. A 3-year follow-up study of enhancing and non-enhancing multiple sclerosis (MS) lesions in MS patients demonstrating clinically isolated syndrome (CIS) using a multi-compartment T_2 relaxometry (MCT2) model

Participants: S. Chatterjee, O. Commowick, B. Combes, A. Kerbrat, C. Barillot.

Obtaining information on condition of tissue microstructures (such as myelin, intra/extra cellular cells, free water) can provide important insights into MS lesion. However, MRI voxels are heterogeneous in terms of tissue microstructure due to the limited imaging resolution owing to existing physical limitations of MRI scanners. Here we evaluated a multi-compartment T_2 relaxometry model and then used it to study the evolution of enhancing (USPIO and gadolinium positive) and non-enhancing lesions in 6 MS patients with CIS characteristics over a period 3 years with 7 follow-up scans post baseline [31].

This work was done in collaboration with Onur Afacn and Simon K. Warfield.

7.1.4. Atlases

7.1.4.1. Anisotropic similarity, a constrained affine transformation: Application to brain development analysis

Participants: A. Legouhy, O. Commowick, C. Barillot.

The study of brain development provides insights in the normal trend of brain evolution and enables early detection of abnormalities. We proposed a method to quantify brain growth in three arbitrary orthogonal directions of the brain through linear registration. We introduced a 9 degrees of freedom transformation that gives the opportunity to extract scaling factors describing brain growth along those directions by registering a database of subjects in a common basis. We applied this framework to create longitudinal curves of scaling ratios along fixed orthogonal directions from 0 to 16 years highlighting anisotropic brain development [39].

This work was done in collaboration with François Rousseau under the ANR MAIA project.

7.1.5. Simultaneous EEG/fMRI

7.1.5.1. Automated Electrodes Detection during simultaneous EEG/fMRI

Participants: M. Fleury, C. Barillot, E. Bannier, P. Maurel.

The coupling of Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) enables the measure of brain activity at high spatial and temporal resolution. The localisation of EEG sources depends on several parameters including the knowledge of the position of the electrodes on the scalp. An accurate knowledge about this information is important for source reconstruction. Currently, when acquiring EEG and fMRI together, the position of the electrodes is generally estimated according to fiducial points by using a template. In the context of simultaneous EEG/fMRI acquisition, a natural idea is to use magnetic resonance (MR) images to localise EEG electrodes. However, most MR compatible electrodes are built to be almost invisible on MR Images. Taking advantage of a recently proposed Ultra short Echo Time (UTE) sequence, we introduce a fully automatic method to detect and label those electrodes in MR images. Our method was tested on 8 subjects wearing a 64-channel EEG cap. This automated method showed an average detection accuracy of 94% and the average position error was 3.1 mm. These results suggest that the proposed method has potential for determining the position of the electrodes during simultaneous EEG/fMRI acquisition with a very light cost procedure" [11], [35].

This work was done in collaboration with Marsel Mano from Biotrial.

7.1.6. Inference in neuroimaging

7.1.6.1. Validity of summary statistics-based mixed-effects group fMRI

Participant: Camille Maumet.

Statistical analysis of multi-subject functional Magnetic Resonance Imaging (fMRI) data is traditionally done using either: 1) a mixed-effects GLM (MFX GLM) where within-subject variance estimates are used and incorporated into per-subject weights or 2) a random-effects General linear model (GLM) (RFX GLM) where within-subject variance estimates are not used. Both approaches are implemented and available in major neuroimaging software packages including: SPM (MFX analysis; 2nd-Level statistics), FSL (FLAME; OLS) and AFNI (3dMEMA; 3dttest++). While MFX GLM provides the most efficient statistical estimate, its properties are only guaranteed in large samples, and it has been shown that RFX GLM is a valid alternative for one-sample group analyses in fMRI [1]. We recently showed that MFX GLM for image-based meta-analysis could lead to invalid results in small-samples. We investigated whether this issue also affects group fMRI [42].

This work was done in collaboration with Prof. Thomas Nichols from the Oxford Big Data Institute, UK.

7.1.6.2. Choosing a practical and valid Image-Based Meta-Analysis

Participant: Camille Maumet.

Meta-analysis provides a quantitative approach to summarise the rich functional Magnetic Resonance Imaging (fMRI) literature. When image data is available for each study, the optimal approach is to perform an Image-Based Meta-Analysis (IBMA) [1]. A number of IBMA approaches have been proposed including combination of standardised statistics (Z's), just effect estimates (E's) or both effect estimates and their standard errors (SE's). While using both E's & SE's and estimating between-study variance should be optimal, the methods are not guaranteed to work for small number of studies. Also, often only standardised estimates are shared, reducing the possible meta-analytic approaches. Finally, because the BOLD signal is non-quantitative care has to be taken in order to insure that E's are expressed in the same units [2,3], especially when combining data from different software packages. Given the growing interest in data sharing in the neuroimaging community there is a need to identify what is the minimal data to be shared in order to allow for future IBMAs. We investigated the validity of 8 IBMA approaches [41].

This work was done in collaboration with Prof. Thomas Nichols from the Oxford Big Data Institute, UK.

7.1.7. Machine learning

7.1.7.1. Learning sparse predictor from hybrid EEG-fMRI neurofeedback

Participants: C. Cury, C. Barillot, P. Maurel.

Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) both allow measurement of brain activity for neuro-feedback (NF), respectively with high temporal resolution for EEG and high spatial resolution for fMRI. Using simultaneously fMRI and EEG for NF training is very promising to devise brain rehabilitation protocols, however performing NF-fMRI is costly, exhausting and time consuming, and cannot be repeated too many times for the same subject. The original contribution of this work concerns the prediction of NF scores from EEG recordings only, using a training phase where both EEG and fMRI NF are available. We have proposed a model able to predict NF scores from coupling EEG-fMRI (NF-EEG-fMRI) of 17 subjects in motor imagery task. The prediction of NF-EEG-fMRI scores was found as satisfactory with a significant improved performance with respect to what EEG can provide alone, when adding to NF-EEG, the prediction of NF-fMRI from EEG signals. The prediction of NF-fMRI significantly adds information and increases the quality of the estimated NF-EEG-fMRI scores.

This work was done in collaboration with Remi Gribonval from the Inria/IRISA PANAMA team.

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

7.2.1. *Bimodal EEG-fMRI Neurofeedback for Stroke Rehabilitation: a Case Report*

Participants: Giulia Lioi, Mathis Fleury, Simon Butet, Christian Barillot, Isabelle Bonan.

Neurofeedback (NF) consists on training self-regulation of brain activity by providing real-time information about the participant brain function. Few works have shown the potential of NF for stroke rehabilitation however its effectiveness has not been investigated yet. NF approaches are usually based on real-time monitoring of brain activity using a single imaging technique. Recent studies have revealed the potential of combining EEG and fMRI to achieve a more efficient and specific self-regulation. In a case report, we tested the feasibility of applying bimodal EEG-MRI NF on two stroke patients. [54]

This work was done in collaboration with Anatole Lecuyer from the Inria/IRISA HYBRID team.

7.2.2. *Refining understanding of working memory buffers through the construct of binding: Evidence from a single case informs theory and clinical practise*

Participants: Pierre-Yves Jonin, Quentin Duché.

Binding operations carried out in working memory enable the integration of information from different sources during online performance. While available evidence suggests that working memory may involve distinct binding functions, whether or not they all involve the episodic buffer as a cognitive substrate remains unclear. Similarly, knowledge about the neural underpinnings of working memory buffers is limited, more specifically regarding the involvement of medial temporal lobe structures. In the present study, we report on the case of patient KA, with developmental amnesia and selective damage to the whole hippocampal system. We found that KA was unable to hold shape-colours associations (relational binding) in working memory. In contrast, he could hold integrated coloured shapes (conjunctive binding) in two different tasks. Otherwise, and as expected, KA was impaired on three relational memory tasks thought to depend on the hippocampus that are widely used in the early detection of Alzheimer's disease. Our results emphasized a dissociation between two binding processes within working memory, suggesting that the visuo-spatial sketchpad could support conjunctive binding, and may rely upon a large cortical network including sub-hippocampal structures. By contrast, we found evidence for a selective impairment of relational binding in working memory when the hippocampal system is compromised, suggesting that the long-term memory deficit observed in amnesic patients may be related to impaired short-term relational binding at encoding. Finally, these findings may inform research on the early detection of Alzheimer's disease as the preservation of conjunctive binding in KA is in sharp contrast with the impaired performance demonstrated very early in this disease [15].

This work was done in collaboration with Mario Alfredo Parra and Clara Calia, from Herriot Wyatt University, Edinburgh, UK; with Emmanuel Barbeau and Sophie Muratot from the CNRS 5549 CerCo unit in Toulouse, France; and with Serge Belliard, from CHU de Rennes, Service de Neurologie, Rennes, France.

7.2.3. *Superior explicit memory despite severe developmental amnesia: In-depth case study and neural correlates*

Participants: Pierre-Yves Jonin, Christian Barillot.

The acquisition of new semantic memories is sometimes preserved in patients with hippocampal amnesia. Robust evidence for this comes from case reports of developmental amnesia suggesting that low-to-normal levels of semantic knowledge can be achieved despite compromised epi-sodic learning. However, it is unclear whether this relative preservation of semantic memory results from normal acquisition and retrieval or from residual episodic memory, combined with effortful repetition. Furthermore, lesion studies have mainly focused on the hippocampus itself, and have seldom reported the state of structures in the extended hippocampal system. Preserved components of this system may therefore mediate residual episodic abilities, contributing to the apparent semantic preservation. We recently reported an in-depth study of Patient KA, a 27-year-old man who had severe hypoxia at birth, in which we carefully explored his residual episodic learning abilities. We used novel speeded recognition paradigms to assess whether KA could explicitly acquire and retrieve new context-free memories. Despite a pattern of very severe amnesia, with a 44-point discrepancy between his intelligence and memory quotients, KA exhibited normal-to-superior levels of knowledge, even under strict time constraints. He also exhibited normal-to-superior recognition memory for new material, again under strict time constraints. Multimodal neuroimaging revealed an unusual pattern of selective atrophy within each component of the extended hippocampal system, contrasting with the preservation of anterior subhippocampal cortices. A cortical thickness analysis yielded a pattern of thinner but also thicker regional cortices, pointing toward specific temporal lobe reorganization following early injury. We thus report the first case of superior explicit learning and memory in a severe case of amnesia, raising important questions about how such knowledge can be acquired [14].

This work was done in collaboration with Emmanuel Barbeau and Gabriel Besson from the CNRS 5549 CerCo unit in Toulouse, France; with Renaud La Joie; with Jérémie Pariente from the Inserm UMR 1214 Tonic unit in Toulouse, France; and with Serge Belliard, from CHU de Rennes, Service de Neurologie, Rennes, France.

7.2.4. *Retrieval practice based on recognition memory: testing the retrieval effort hypothesis*

Participants: Pierre-Yves Jonin, Christian Barillot.

We tested a core prediction of the retrieval effort hypothesis as an account for the testing effect (TE). Retrieval effort predicts that automatic, effortful retrieval should not lead to TE. Experiment 1 (N=76) showed that despite an encoding duration of three times less, object pictures retention is better after repeated testing under Old/New recognition conditions, than following repeated study. Experiment 2 (N=30) used a speeded and accuracy boosting procedure to rule out the contribution of recollection to retrieval. Retention of object pictures at 25 minutes and 6 months was similar after repeated testing or after repeated studying. These results call for a revision of the retrieval effort hypothesis as a mechanistic account of TE [53].

This work was done in collaboration with Audrey Noël, from Université de Rennes 2, Rennes, France; with Gabriel Besson from the CNRS 5549 CerCo unit in Toulouse, France; with Emmanuel Barbeau and Sophie Muratot from the CNRS 5549 CerCo unit in Toulouse, France; and with Serge Belliard, from CHU de Rennes, Service de Neurologie, Rennes, France.

7.2.5. *Voxel-wise Comparison with a-contrario Analysis for Automated Segmentation of Multiple Sclerosis Lesions from Multimodal MRI*

Participants: Francesca Galassi, Olivier Commowick, Christian Barillot.

We have introduced a new framework for the automated and un-supervised segmentation of Multiple Sclerosis lesions from multimodal Magnetic Resonance images. It relies on a voxel-wise approach to detect local white matter abnormalities, with an a-contrario analysis, which takes into account local information. First, a voxel-wise comparison of multimodal patient images to a set of controls is performed. Then, region-based probabilities are estimated using an a-contrario approach. Finally, correction for multiple testing is performed. Validation was undertaken on a multi-site clinical dataset of 53 MS patients with various number and volume of lesions. We showed that the proposed framework outperforms the widely used FDR-correction for this type of analysis, particularly for low lesion loads [37].

This work was done in collaboration with Emmanuel Vallée from Orange Labs, Lannion, France.

7.2.6. Integration of Probabilistic Atlas and Graph Cuts for Automated Segmentation of Multiple Sclerosis lesions

Participants: Francesca Galassi, Olivier Commowick, Christian Barillot.

We have proposed a framework for automated segmentation of Multiple Sclerosis (MS) lesions from MR brain images. It integrates a priori tissues and MS lesions information into a Graph-Cuts algorithm for improved segmentation results. [36].

7.2.7. Multiple sclerosis

7.2.7.1. Spinal Cord

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 publication [6]. Based on this work, a second work investigated the sensitivity of magnetization transfer to assess diffuse and focal burden in MS patients and was published in Multiple Sclerosis [5].

7.2.7.2. Reproducibility and Evolution of Diffusion MRI measurements within the Cervical Spinal Cord in Multiple Scleros

Participants: Haykel Snoussi, Anne Kerbrat, Benoit Combès, Olivier Commowick, Élise Bannier, Emmanuel Caruyer, Christian Barillot.

In Multiple Sclerosis (MS), there is a large discrepancy between the clinical observations and how the pathology is exhibited on brain images, this is known as the clinical-radiological paradox (CRP). One of the hypotheses is that the clinical deficit may be more related to the spinal cord damage than the number or location of lesions in the brain. Therefore, investigating how the spinal cord is damaged becomes an acute challenge to better understand and overcome the CRP. Diffusion MRI is known to provide quantitative

figures of neuronal degeneration and axonal loss, in the brain as well as in the spinal cord. In this work, we have proposed to investigate how diffusion MRI metrics vary in the different cervical regions with the progression of the disease. We first study the reproducibility of diffusion MRI on healthy volunteers with a test-retest procedure using both standard diffusion tensor imaging (DTI) and multi-compartment Ball-and-Stick models. Then, based on the test re-test quantitative calibration, we provided quantitative figures of pathology evolution between M0 and M12 in the cervical spine on a set of 31 MS patients, exhibiting how the pathology damage spans in the cervical spinal cord.

7.2.8. Epilepsy

Participants: Élise Bannier, Jean-Christophe Ferré.

Accurate localization of the thalamic subregions is of paramount importance for Deep Brain Stimulation (DBS) planning. Current MRI protocols use T2 and Gadolinium-enhanced T1 images, to visualize both the basal ganglia and the vessels, in order to define the electrode trajectory and target. A study showing the usefulness of Fluid and White Matter Suppression, i.e. FLAWS imaging, in eleven drug-resistant epileptic patients for preoperative Deep Brain Stimulation planning and anterior thalamic nucleus targeting was presented as a Power Pitch at the ISMRM Meeeting in Paris [27].

This work was done in collaboration with Giulio Gambarota, Anca Nica and Claire Haegelen from the LTSI and Tobias Kober from Lausanne.

7.2.9. Arterial Spin Labeling in pediatric populations

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

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. As part of the PhD of Maia Proisy, we have been working on processing and analysing MR perfusion images using arterial spin labeling in neonates and children for several purposes:

- Investigation of brain perfusion evolution between 6 month and 15 years using ASL sequence in order to provide reference values in this age range [4],
- Evaluation of the evolution of the cerebral blood flow changes betwren day-of-life 3 and day-of-life 10 in a population of neonates with hypoxic-ischemic encephalopathy [45], [*"Changes in brain perfusion in successive arterial spin labelling MRI scans in neonates with hypoxic-ischaemic encephalopathy"*, article in revision in Neuroimage: Clinical].

7.2.10. Diffusion MRI in depression

7.2.10.1. Diffusion MRI as an imaging marker of depression from a large and homogenous population study

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

Despite the extensive therapy options available for depression, up to 80% of patients will suffer from a relapse. Consequently, understanding the neural correlates underlying the depression will optimize the diagnosis and treatment of individual depressed patients. In an experimental study, we investigated alterations of white matter integrity in a large cohort of patients suffering from depression using diffusion tensor imaging. 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 depression pathogenesis. [34].

This work was done in collaboration with Dominique Drapier from Academic Psychiatry Department, Centre Hospitalier Guillaume Régnier, Rennes, France, EA 4712 Behavior and Basal Ganglia, CHU Rennes, Rennes 1 University, Rennes, France.

7.2.10.2. Diffusion MRI as a descriptive imaging marker of the pathogenesis of treatment-resistant depression

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

Despite the extensive therapy options available for depression, treatment-resistant depression (TRD) occurs in 20-30% of depressed patients. Consequently, identification of neural changes in TRD could support to better understand the mechanism of resistance and to improve the treatment of individual depressed patients. We aimed to investigate the white-matter microstructure in a sample of depressed patients in which response to treatment was subsequently evaluated 6 months after. Our findings suggest the abnormalities of the white-matter integrity in multiple white matter tracts, such as anterior limb of internal capsule and genu of corpus may play a role in the pathogenesis of treatment-resistant depression [33].

Depressive disorder is characterized by a profound dysregulation of affect and mood as well as additional abnormalities including cognitive dysfunction, insomnia, fatigue and appetite disturbance. This disease is the most prevalent mental illness, with an estimated lifetime prevalence reported to range from 10% to 15% worldwide. Despite the extensive therapy options available for depression, up to 80% of patients will suffer from a relapse. Consequently, understanding the neural correlates underlying the depression is critical for improving the specificity and efficacy of diagnostic and treatment strategies.Previous studies of structural and functional magnetic resonance imaging have reported several microstructural abnormalities in the prefrontal cortex, anterior cingulate cortex, hippocampus and thalamus. These observations suggest a dysfunction of the circuits connecting frontal and subcortical brain regions, leading to a "disconnection syndrome". Using graph theory-based analysis, we examined white matter changes in the organization of networks in patients suffering from depression. Our diffusion imaging data showed white matter alteration in patients suffering from depression is occurring in the anterior thalamic radiation and in the cingulate bundle. Our findings suggest decreased fiber density in circuits connecting subcortical brain regions with the frontal and parietal cortex, supporting the theory of limbic-frontal circuit dysfunction [50]. *We were awarded for this workby the French Institute of Psychiatry for our communication at its annual Forum.*

This work was done in collaboration with Dominique Drapier from Academic Psychiatry Department, Centre Hospitalier Guillaume Régnier, Rennes, France, EA 4712 Behavior and Basal Ganglia, CHU Rennes, Rennes 1 University, Rennes, France.

7.3. Research axis 3: Management of Information in Neuroimaging

7.3.1. Large-scale data analyses

7.3.1.1. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure

Participants: O. Commowick, M. Kain, F. Leray, M. Simon, J.-C. Ferré, A. Kerbrat, G. Edan, C. Barillot.

In collaboration with OFSEP and France Life Imaging, we have proposed a study of multiple sclerosis segmentation algorithms conducted at the international MICCAI 2016 challenge. This challenge was operated using France Life Imaging (FLI-IAM), a new open-science computing infrastructure. This allowed for the automatic and independent evaluation of a large range of algorithms in a fair and completely automatic manner. This computing infrastructure was used to evaluate thirteen methods of MS lesions segmentation, exploring a broad range of state-of-the-art algorithms, against a high-quality database of 53 MS cases coming from four centers following a common definition of the acquisition protocol. Each case was annotated manually by an unprecedented number of seven different experts. Results of the challenge highlighted that automatic algorithms, including the recent machine learning methods (random forests, deep learning,...), are still trailing human expertise on both detection and delineation criteria. In addition, we demonstrated that computing a statistically robust consensus of the algorithms performs closer to human expertise on one score (segmentation) although still trailing on detection scores [7]

This work was done in collaboration with A. Istace, B. Laurent, S. C. Pop, P. Girard, R. Ameli, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. Mckinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M. M. Santos, W. P. Santos, A. G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F. J. Vera-Olmos, N. Malpica, C. R. G. Guttmann, S. Vukusic, M. Dojat, M. Styner, S. K. Warfield and F. Cotton.

7.3.1.2. Same Data - Different Software - Different Results? Analytic Variability of Group fMRI Results

Participant: Camille Maumet.

A wealth of analysis tools are available to fMRI researchers in order to extract patterns of task variation and, ultimately, understand cognitive function. However, this 'methodological plurality' comes with a drawback. While conceptually similar, two different analysis pipelines applied on the same dataset may not produce the same scientific results. Differences in methods, implementations across software packages, and even operating systems or software versions all contribute to this variability. Consequently, attention in the field has recently been directed to reproducibility and data sharing. Neuroimaging is currently experiencing a surge in initiatives to improve research practices and ensure that all conclusions inferred from an fMRI study are replicable. In this work, our goal was to understand how choice of software package impacts on analysis results. We used publically shared data from three published task fMRI neuroimaging studies, reanalyzing each study using the three main neuroimaging software packages, AFNI, FSL and SPM, using parametric and nonparametric inference. We obtained all information on how to process, analyze, and model each dataset from the publications. We made quantitative and qualitative comparisons between our replications to gauge the scale of variability in our results and assess the fundamental differences between each software package. While qualitatively we found broad similarities between packages, we also discovered marked differences, such as Dice similarity coefficients ranging from 0.000-0.743 in comparisons of thresholded statistic maps between software [28], [48].

This work was done in collaboration with Alexander Bowring and Prof. Thomas Nichols from the Oxford Big Data Institute in the UK.

7.3.1.3. Detecting and Interpreting Heterogeneity and Publication Bias in Image-Based Meta-Analyses

Participant: Camille Maumet.

With the increase of data sharing, meta-analyses are becoming increasingly important in the neuroimaging community. They provide a quantitative summary of published results and heightened confidence due to higher statistical power. The gold standard approach to combine results from neuroimaging studies is an Image-Based Meta-Analysis (IBMA) [1] in which group-level maps from different studies are combined. Recently, we have introduced the IBMA toolbox, an extension for SPM that provides methods for combining image maps from multiple studies [2]. However, the current toolbox lacks diagnostic tools used to assess critical assumptions of meta-analysis, in particular whether there is inter-study variation requiring random-effects IBMA, and whether publication bias is present. We have proposed two new tools added to the IBMA toolbox to detect heterogeneity and to assess evidence of publication bias [40].

This work was done in collaboration with Thomas Maullin-Saper and Prof. Thomas Nichols from the Oxford Big Data Institute in the UK.

7.3.2. Infrastructures

7.3.2.1. Open Science for the Neuroinformatics community

Participants: Camille Maumet, Xavier Rolland, Michael Kain, Christian Barillot.

The Neuroinformatics community in OpenAire-Connect is represented by members of the France Life Imaging (FLI) collaboration. In this context, we aim at leveraging OpenAire-Connect services and give our community members the possibility to easily publish and exchange research artefacts from FLI platforms, such as VIP and Shanoir. This will enable open and reproducible science, since literature, data, and methods can be linked, retrieved, and replayed by all the members of the community [30].

This work was done in collaboration with Sorina Pop, Axel Bonnet and Tristan Glatard.

7.3.3. Standardisation and interoperability

7.3.3.1. Interoperability with Boutiques and CARMIN

Participants: Camille Maumet, Michael Kain, Christian Barillot.

A growing number of platforms and tools have lately been developed to meet the needs of various scientific communities. Most of these solutions are optimized to specific requirements from different user groups, leading to technological fragmentation and lack of interoperability. In our quest of open and reproducible science, we proposed two complementary tools, Boutiques and CARMIN, providing cross-platform interoperability for scientific applications, data sharing and processing [29].

This work was done in collaboration with Sorina Pop, Axel Bonnet and Tristan Glatard.

7.3.3.2. A standardised representation for non-parametric fMRI results

Participant: Camille Maumet.

Reuse of data collected and analysed at another site is becoming more prevalent in the neuroimaging community but this process usually relies on intensive data and metadata curation. Given the ever-increasing number of research datasets produced and shared, it is desirable to rely on standards that will enable automatic data and metadata retrieval for large-scale analyses. We recently introduced NIDM-Results, a data model to represent and publish data and metadata created as part of a mass univariate neuroimaging study (typically functional magnetic resonance imaging). In this work, we have proposed to extend this model to allow for the representation of non-parametric analyses and we introduce a JSON API that will facilitate export into NIDM-Results [25].

This work was done as part of an international collaboration with Guillaume Flandin, Martin Perez-Guevara, Jean-Baptiste Poline, Justin Rajendra, Richard Reynolds, Betrand Thirion, Thomas Maullin-Sapey and Thomas Nichols.

7.3.3.3. Development of an Ontology for the INCF Neuroimaging Data Model (NIDM)

Participant: C. Maumet.

The successful reuse of shared data relies on the existence of easily-available well-described metadata. The metadata, as a rich description of the data, must capture information on how the data was acquired, processed and analyzed. The terms used to describe the data should be chosen with a logical, consistent framework in mind and include definitions to avoid ambiguity. In addition, a lexicon or ontology should reuse terms from existing efforts as much as possible [38].

This work was done as part of an international collaboration with K.G. Helmer, K.B. Keator, T. Auer, S. Ghosh, T.E. Nichols, P. Smruti and J.B. Poline.

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation

Participants: Hamza Ben Ammar, Yann Busnel, Pierre L'Ecuyer, Gerardo Rubino, Yassine Hadjadj-Aoul, Sofiène Jelassi, Patrick Maillé, Yves Mocquard, Bruno Sericola.

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. We proposed 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.

We consider recently an application to continuous queries, which are processed by a stream processing engine (SPE) to generate timely results given the ephemeral input data. Variations of input data streams, in terms of both volume and distribution of values, have a large impact on computational resource requirements. Dynamic and Automatic Balanced Scaling for Storm (DABS-Storm) [17] is an original solution for handling dynamic adaptation of continuous queries processing according to evolution of input stream properties, while controlling the system stability. Both fluctuations in data volume and distribution of values within data streams are handled by DABS-Storm to adjust the resources usage that best meets processing needs. To achieve this goal, the DABS-Storm holistic approach combines a proactive auto-parallelization algorithm with a latency-aware load balancing strategy.

Sampling techniques is a classical method for detection in large-scale data streams. We have proposed a new algorithm that detects on the fly the k most frequent items in the sliding window model [37]. This algorithm is distributed among the nodes of the system. It is inspired by a recent and innovative approach, which consists in associating a stochastic value correlated with the item's frequency instead of trying to estimate its number of occurrences. This stochastic value corresponds to the number of consecutive heads in coin flipping until the first tail occurs. The original approach was to retain just the maximum of consecutive heads obtained by an item, since an item that often occurs will have a higher probability of having a high value. While effective for very skewed data distributions, the correlation is not tight enough to robustly distinguish items with comparable frequencies. To address this important issue, we propose to combine the stochastic approach with a deterministic counting of items. Specifically, in place of keeping the maximum number of consecutive heads obtained by an item, we count the number of times the coin flipping process of an item has exceeded a given threshold. This threshold is defined by combining theoretical results in leader election and coupon collector problems. Results on simulated data show how impressive is the detection of the top-k items in a large range of distributions [38].

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 switch among these based on measurements collected during the connection. Knowing the achievable data rate before the connection establishment should definitely help content providers to choose the most appropriate representation from the very beginning. We have conducted several large measurement campaigns involving a panel of 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 establish evidence that it is indeed possible to exploit these measurements to predict, with an acceptable accuracy, the achievable data rate. We thus introduce cooperation strategies between the mobile user, the network operator and the content provider to implement such anticipatory solution [23].

Call Centers. In emergency call centers (for police, firemen, ambulances) a single event can sometimes trigger many incoming calls in a short period of time. Several people may call to report the same fire or the same accident, for example. Such a sudden burst of incoming traffic can have a significant impact on the responsiveness of the call center for other events in the same period of time. We examine in [53] data from the SOS Alarm center in Sweden. We also build a stochastic model for the bursts. We show how to estimate the model parameters for each burst by maximum likelihood, how to model the multivariate distribution of those parameters using copulas, and how to simulate the burst process from this model. In our model, certain events trigger an arrival process of calls with a random time-varying rate over a finite period of time of random length.

Performance of CDNs. In order to track the users who illegally re-stream live video streams, one solution is to embed identified watermark sequences in the video segments to distinguish the users. However, since all types of watermarked segments should be prepared, the existing solutions require an extra cost of bandwidth for delivery (at least multiplying by two the required bandwidth). In [66], we study how to reduce the inner delivery (traffic) cost of a Content Delivery Network (CDN). We propose a mechanism that reduces the number of watermarked segments that need to be encoded and delivered. We calculate the best- and worst-case traffics for two different cases: multicast and unicast. The results illustrate that even in the worst cases, the traffic with our approach is much lower than without reducing. Moreover, the watermarked sequences can still maintain uniqueness for each user. Experiments based on a real database are carried out, and illustrate that our mechanism significantly reduces traffic with respect to the current CDN practice.

Beyond CDNs, Network Operators (NOs) are developing caching capabilities within their own network infrastructure, in order to face the rise in data consumption and to avoid the potential congestion at peering links. These factors explain the enthusiasm of industry and academics around the Information-Centric Networking (ICN) concept and its in-network caching feature. Many contributions focused these last years on improving the caching performance of ICN. In [41], we propose a very versatile model capable of modeling the most efficient caching strategies. We first start by representing a single generic cache node. We then extend our model for the case of a network of caches. The obtained results are used to derive, in particular, the cache hit probability of a content in such caching systems. Using a discrete event simulator, we show the accuracy of the proposed model under different network configurations.

Probabilistic analysis of population protocols. In [20], we studied the well-known problem of 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. This model 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 the 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 to each node to determine the instants at which interactions occur. The rumor spreading time is thus more realistic since it is the real 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 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.

The context of [57] is the two-choice paradigm which is deeply used in balanced online resource allocation, priority scheduling, load balancing and more recently in population protocols. The model governing the evolution of these systems consists in throwing balls one by one and independently of each others into n bins, which represent the number of agents in the system. At each discrete instant, a ball is placed in the least filled bin among two bins randomly chosen among the n ones. A natural question is the evaluation of the difference between the number of balls in the most loaded and the one in the least loaded bin. At time t, this difference is denoted by Gap(t). A lot of work has been devoted to the derivation of asymptotic approximations of this gap for large values of n. In this paper we go a step further by showing that for all $t \geq 0$, $n \geq 2$ and $\sigma \geq 0$, the variable Gap(t) is less than $a(1+\sigma)\ln(n)+b$ with probability greater than $1-1/n^{\sigma}$, where the constants a and b, which are independent of t, σ and n, are optimized and given explicitly, which to the best of our knowledge has never been done before.

The work described in [58] focuses on pairwise interaction-based protocols, and proposes an universal mechanism that allows each agent to locally detect that the system has converged to the sought configuration with high probability. To illustrate our mechanism, we use it to detect the instant at which the proportion problem is solved. Specifically, let n_A (resp. n_B) be the number of agents that initially started in state A (resp. B) and $\gamma_A = n_A/n$, where n is the total number of agents. Our protocol guarantees, with a given precision $\varepsilon > 0$ and any high probability $1-\delta$, that after $O(n\ln(n/\delta))$ interactions, any queried agent that has set the detection flag will output the correct value of the proportion γ_A of agents which started in state A, by maintaining no more than $O(\ln(n)/\varepsilon)$ integers. We are not aware of any such results. Simulation results illustrate our theoretical analysis.

All these works are part of the thesis [11].

Fluid Queues. Stochastic fluid flow models and in particular those driven by Markov chains have been intensively studied in the last two decades. Not only they have been proven to be efficient tools to mimic Internet traffic flow at a macroscopic level but they are useful tools in many areas of applications such as manufacturing systems or in actuarial sciences to cite but a few. We propose in a forthcoming book, entitled Advanced Trends in Queueing Theory, edited by V. Anisimov and N. Limnios in the Mathematics and Statistics Series, Sciences, Iste & J. Wiley, a chapter which focus on such a model in the context of performance analysis of a potentially congested system. The latter is modeled by means of a finite-capacity system whose content is described by a Markov driven stable fluid flow. We step-by-step describe a methodology to compute exactly the loss probability of the system. Our approach is based on the computation of hitting probabilities jointly with the peak level reached during a busy period, both in the infinite and finite buffer case. Accordingly we end up with differential Riccati equations that can be solved numerically. Moreover we are able to characterize the complete distribution of both the duration of congestion and of the total information lost during such a busy period.

Organizing both transactions and blocks in a distributed ledger. We propose in [39] a new way to organize both transactions and blocks in a distributed ledger to address the performance issues of permissionless ledgers. In contrast to most of the existing solutions in which the ledger is a chain of blocks extracted from a tree or a graph of chains, we present a distributed ledger whose structure is a balanced directed acyclic graph of blocks.

We call this specific graph a SYC-DAG. We show that a SYC-DAG allows us to keep all the remarkable properties of the Bitcoin blockchain in terms of security, immutability, and transparency, while enjoying higher throughput and self-adaptivity to transactions demand. To the best of our knowledge, such a design has never been proposed.

Additional intermittent server. We analyzed the performance of a system consisting of a queue with one regular single server supported by an additional intermittent server who, in order to decrease the mean response time, i) leaves the back office to join the first server when the number of customers reaches the threshold K, and ii) leaves the front office when it has no more customers to serve. This study produced a closed-form solution for the steady state probability distribution and for different metrics such as expected response times for customers or expectation of busy periods. Then, for a given value of K, the influence of the intermittent server on the response time is exhibited. The consequences on the primary task of the intermittent server are investigated through metrics such as mean working and pseudo-idle periods. Finally, a cost function is proposed from which an optimal value of the threshold K is obtained. The results were the subject of the book chapter [71].

Operational availability prediction. The evaluation of the operational availability of a fleet of systems on an operational site is far from trivial when the size of the state space of a faithful Markovian model makes this issue unrealistic for many large models. The main difficulty comes from the existence on the site of "on line replaceable units" (LRU) that may be unavailable from time to time when a breakdown occurs. To be more precise, let us say that the intrinsic availability is an upper limit of the operational availability because the considered unavailability corresponds only to the time necessary to proceed with the exchange of the defective element. This assumes that the repairer and the spare part are always immediately available on the operational site. To reduce the intervention time at an operational site, system repair consists of replacing the defective subset with an identical one in good condition. These exchangeable subsets are called LRUs. Restoring a piece of system by exchanging an LRU makes it possible to obtain a rapid return to service of the system and does not require the presence on the operational site of several specialists. Thus, we can change an aircraft engine in a few hours thanks to stakeholders who are knowledgeable about the procedures for intervention on fasteners and connections but who are not required to know the techniques to repair a broken engine. The operational availability (the one that is of interest to the user) will generally be lower than the intrinsic availability because the latter integrates the unavailability of repairer or LRU, if any. And if the waiting time for a repairer is generally measured in hours, the unavailability of a LRU can be measured in days, in weeks, even in months! It is therefore this last point which a potential customer should worry about in priority. Moreover, the unavailability of the LRU is more difficult to model and evaluate than that of the repairer.

In a first study we considered a virtual system with only one type of LRU in order to understand the influence of various parameters on the operational availability both of a faithful Markovian model (for which we are able to get the exact answer) and of a proposed approximate method. By doing so, we were able to compute the relative errors induced when using this latter solution. The approximate method showed a quite good accuracy [56]. The generalization to systems with multiple types of line replaceable units was conducted in a second study. The main idea is to consider a non product-form queuing network and to aggregate subsets of it as if they were parts of a product-form queuing network. Note that if the spare LRUs were always available on the operational site (which is never the case) we could fairly model the behavior of the support system on the site by means of a product-form network. Also in this second study, the potential lack of repairer is taken into account. The low relative complexity of the new recurrent approximate method allows it to be used for applications encountered in the field of maintenance. Although approximate, the method provides the result with a small relative error, ranging from 10^{-2} to 10^{-8} for examples which can be compared with our reference Markovian model. For now, the method only concerns systems consisting of a series of LRUs [64]).

Modeling loss processes in voice traffic. Markov models of loss incidents happening during packet voice communications are needful for many engineering tasks, namely network dimensioning and automatic quality assessment. Two very simple ones are Bernoulli and 2-state Markov models, but they carry limited information about incurred loss incidents. On the other hand, a general Markov loss model with 2^k states, where k is the window length used for observing the voice packet arrival process, leads to heavy computations and an excessive lookahead delay. Moreover, legacy Markov loss models concentrate mostly on capturing some physical characteristics of loss incidents, rather than their perceived effects.

In [16], we propose a comprehensive and detailed Markov loss model considering the distinguished perceived effects caused by different loss incidents. Specifically, it explicitly differentiates between (1) isolated 20 msec loss incidents which are inaudible by the human ears, (2) highly and lowly frequent short loss incidents (20-80 msec) that are perceived by humans as bubbles and (3) long loss incidents ($\geq$ 80 msec) inducing interruptions that dramatically decrease speech intelligibility. Our numerical analysis show that our Markov loss model captures subtle characteristics of loss incidents observed in empirical traces sampled over representative network paths.

Transient analysis of Markovian models. Continuing with a research line that we started years ago with colleagues in California, where we addressed the transient state distributions of Markovian queuing models using the concept of pseudo-dual proposed by Anderson, we discovered this year a new way to attack these problems, leading to an approach with a much wider application range. Moreover, this methodology clarifies some phenomena that appeared when Anderson's tools were used. The result is now two new concepts we propose, related in general to arbitrary square matrices (possibly infinite), the *power-dual* and the *exponential-dual*, and the way we can apply them to the analysis of linear systems of difference or differential equations. The first elements of this new theory were discussed in [26].

7.2. Machine learning

Participants: Imad Alawe, Yassine Hadjadj-Aoul, Corentin Hardy, Gerardo Rubino, Bruno Sericola, César Viho.

Distributed deep learning on edge-devices. A recently celebrated type of deep neural network is the Generative Adversarial Network (GAN). GANs are generators of samples from a distribution that has been learned; they are up to now centrally trained from local data on a single location. We question in [49] and in [74] the performance of training GANs using a spread dataset over a set of distributed machines, using a gossip approach shown to work on standard neural networks. This performance is compared to the federated learning distributed method, that has the drawback of sending model data to a server. We also propose a gossip variant, where GAN components are gossiped independently. Experiments are conducted with Tensorflow with up to 100 emulated machines, on the canonical MNIST dataset. The position of these papers is to provide a first evidence that gossip performances for GAN training are close to the ones of federated learning, while operating in a fully decentralized setup. Second, to highlight that for GANs, the distribution of data on machines is critical (i.e., i.i.d. or not). Third, to illustrate that the gossip variant, despite proposing data diversity to the learning phase, brings only marginal improvements over the classic gossip approach.

Machine learning acceleration. The number of connected devices is increasing with the emergence of new services and trends. This phenomenon is leading to a traffic growth over both the control and the data planes of the mobile core network. Therefore the 3GPP group has rethought the architecture of the New Generation Core (NGC) by defining its components as Virtualized Network Functions (VNF). However, scalability techniques should be envisioned in order to answer the needs, in term of resource provisioning, without degrading the Quality Of Service (QoS) already offered by hardware based core networks. Neural networks, and in particular deep learning, having shown their effectiveness in predicting time series [13], could be good candidates for predicting traffic evolution.

In [35], we proposed a novel solution to generalize neural networks while accelerating the learning process by using K-mean clustering, and a Monte-Carlo method. We benchmarked multiple types of deep neural networks using real operator's data in order to compare their efficiency in predicting the upcoming network load for dynamic and proactive resource provisioning. The proposed solution allows obtaining very good predictions of the traffic evolution while reducing by 50% the time needed for the learning phase.

Machine Learning in Quality of Experience assessment. In a series of presentations we have disseminated the main ideas behind a new generation of Quality of Esperience assessing tools in preparation in the team. In the meetings [70] and [69], and also in the plenary [32], we described some of the key features of the tools we used in our PSQA project, the Random Neural Network of Erol Gelenbe, and the ideas we are following for extending some of their capabilities. The goal is to allow the user to evaluate with little additional cost, the sensitivities of the Quality of Experience with respect to specific metrics of interest, having in mind design applications, or improvement of existing systems. Another example is to invert the PSQA function providing a measure of the Quality of Experience as a function of several QoS and channel-based metrics, in order to define subsets of their joint state space where quality has a given property of interest (for instance, being good enough). In the plenary talk [31], we described other properties of these tools, and other directions being explored, such as the replacement of the subjective testing sessions leading to fully automatic tools, as well as to big data problems. In the keynote talk [82] we showed how to use our PSQA technology for classic performance evaluation works. The idea is that instead if targeting classic performance metrics such as a mean response time, or a loss rate (or dependability ones, the approach is the same), we can develop models that target the "ultimate goal", the Quality of Experience itself. That is, instead of, say, providing a formula allowing to relate the loss rate of a system to the input data, we can obtain a (more complex) formula giving a numerical measure of the Quality of Experience as a function of the same data.

7.3. Network Economics

Participants: Bruno Tuffin, Patrick Maillé.

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 [83].

Reliability/security. In an ad hoc network, accessing a point depends on the participation of other, intermediate, nodes. Each node behaving selfishly, we end up with a non-cooperative game where each node incurs a cost for providing a reliable connection but whose success depends not only on its own reliability investment but also on the investment of nodes which can be on a path to the access point. Our purpose in [76] is to formally define and analyze such a game: existence of an equilibrium output, comparison with the optimal cooperative case, etc.

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 discuss in [79] the impact consequences of such a measure.

Community networks. Community networks have emerged as an alternative to licensed-band systems (WiMAX, 4G, etc.), providing an access to the Internet with Wi-Fi technology while covering large areas. A community network is easy and cheap to deploy, as the network is using members' access points in order to cover the area. We study in [80] the competition between a community operator and a traditional operator (using a licensed-band system) through a game-theoretic model, while considering the mobility of each user in the area.

Spectrum sharing & cognitive networks. Licensed shared access (LSA) is a new approach that allows Mobile Network Operators to use a portion of the spectrum initially licensed to another incumbent user, by obtaining a license from the regulator via an auction mechanism. In this context, different truthful auction mechanisms have been proposed, and differ in terms of allocation (who gets the spectrum) but also on revenue. Since those mechanisms could generate an extremely low revenue, we extend them by introducing a reserve price per bidder which represents the minimum amount that each winning bidder should pay. Since this may be at the expense of the allocation fairness, for each mechanism we find in [44] by simulation the reserve price that optimizes a trade-off between expected fairness and expected revenue. For each mechanism, we analytically express the expected revenue when valuations of operators for the spectrum are independent and identically distributed from a uniform distribution. We also propose in [46] PAM: Proportional Allocation Mechanism, which is a truthful auction mechanism offering a good compromise between fairness and efficiency and can generate the highest revenue to the regulator compared to other truthful mechanisms proposed in the literature.

Selfish primary user emulation (PUE) is a serious security problem in cognitive radio networks. By emitting emulated incumbent signals, a PUE attacker can selfishly occupy more channels. Consequently, a PUE attacker can prevent other secondary users from accessing radio resources and interfere with nearby primary users. To mitigate the selfish PUE, a surveillance process on occupied channels could be performed. Determining surveillance strategies, particularly in multi-channel context, is necessary for ensuring network operation fairness. Since a rational attacker can learn to adapt to the surveillance strategy, the question is how to formulate an appropriate modeling of the strategic interaction between a defender and an attacker. In [24], we study the commitment model in which the network manager takes the leadership role by committing to its surveillance strategy and forces the attacker to follow the committed strategy. The relevant strategy is analyzed through the Strong Stackelberg Equilibrium (SSE). Analytical and numerical results suggest that, by playing the SSE strategy, the network manager significantly improves its utility with respect to playing a Nash equilibrium (NE) strategy, hence obtains a better protection against selfish PUEs. Moreover, the computational effort to compute the SSE strategy is lower than to find a NE strategy.

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. We gave a tutorial on this topic [33], with a video available at https://www.youtube.com/watch?v=EaKtzxPHluU .

In [78], we place and discuss with a net neutrality context the conflict in early 2018 between Orange and TV channel TF1 to prevent some content to be distributed. The related issue of big CPs puhsing ISPs to improve their (own) QoS is further analyzed [77][54]. Indeed, there is a trend for big content providers such as Netflix and YouTube to give grades to ISPs, to incentivize those ISPs to improve at least the quality offered to their service. We design a model analyzing ISPs' optimal allocation strategies in a competitive context and in front of quality-sensitive users. We show that the optimal strategy is non-neutral, that is, it does not allocate bandwidth proportionally to the traffic share of content providers. On the other hand, we show that non-neutrality does not benefit ISPs but is surprisingly favorable to users' perceived quality.

Another current important issue in the current net neutrality debate is that of 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 [55] 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 depending on the chosen advertisement level. We also discuss the impact of zero-rating where an ISP offers free data to a CP to attract more customers, and vertical integration where a CP and an ISP are the same company.

Search engines. Different search engines provide different outputs for the same keyword. This may be due to different definitions of relevance, and/or to different knowledge/anticipation of users' preferences, but rankings are also suspected to be biased towards own content, which may prejudicial to other content providers. In [75], we make some initial steps toward a rigorous comparison and analysis of search engines, by proposing a definition for a consensual relevance of a page with respect to a keyword, from a set of search engines. More specifically, we look at the results of several search engines for a sample of keywords, and define for each keyword the visibility of a page based on its ranking over all search engines. This allows to define a score of the search engine for a keyword, and then its average score over all keywords. Based on the pages visibility, we can also define the consensus search engine as the one showing the most visible results for each keyword. We have implemented this model and present in [75] an analysis of the results.

7.4. Monte Carlo

Participants: Bruno Tuffin, Ajit Rai, 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.

MCMC. The current popular method for approximate simulation from the posterior distribution of the linear Bayesian LASSO is a Gibbs sampler. It is well-known that the output analysis of an MCMC sampler is difficult due to the complex dependence among the states of the underlying Markov chain. Practitioners can usually only assess the convergence of MCMC samplers using heuristics. In [42] we construct a method that yields an independent and identically distributed (iid) draws from the LASSO posterior. The advantage of such exact sampling over the MCMC sampling is that there are no difficulties with the output analysis of the exact sampler, because all the simulated states are independent. The proposed sampler works well when the dimension of the parameter space is not too large, and when it is too large to permit exact sampling, the proposed method can still be used to construct an approximate MCMC sampler

Rare event simulation. We develop in [48] simulation estimators of measures associated with the tail distribution of the hitting time to a rarely visited set of states of a regenerative process. In various settings, the distribution of the hitting time divided by its expectation converges weakly to an exponential as the rare set becomes rarer. This motivates approximating the hitting-time distribution by an exponential whose mean is the expected hitting time. As the mean is unknown, we estimate it via simulation. We then obtain estimators of a quantile and conditional tail expectation of the hitting time by computing these values for the exponential approximation calibrated with the estimated mean. Similarly, the distribution of the sum of lengths of cycles before the one hitting the rare set is often well-approximated by an exponential, and we analogously exploit this to estimate tail measures of the hitting time. Numerical results demonstrate the effectiveness of our estimators.

In rare event simulation, the two main approaches are Splitting and Importance Sampling.

Concerning Splitting, we study in [52] the behavior of a method for sampling from a given distribution conditional on the occurrence of a rare event. The method returns a random-sized sample of points such that unconditionally on the sample size, each point is distributed exactly according to the original distribution conditional on the rare event. For a cost function which is nonzero only when the rare event occurs, the method provides an unbiased estimator of the expected cost, but if we select at random one of the returned points, its distribution differs in general from the exact conditional distribution given the rare event. However, we prove that if we repeat the algorithm, the distribution of the selected point converges to the exact one in total variation.

Splitting and another technique based on a conditional Monte Carlo approach have been applied and compared in [73] for the reliability estimation for networks whose links have random capacities and in which a certain target amount of flow must be carried from some source nodes to some destination nodes. Each destination node has a fixed demand that must be satisfied and each source node has a given supply. We want to estimate the unreliability of the network, defined as the probability that the network cannot carry the required amount of flow to meet the demand at all destination nodes. When this unreliability is very small, which is our main interest in this paper, standard Monte Carlo estimators become useless because failure to meet the demand is a rare event. We find that the conditional Monte Carlo technique is more effective when the network is highly reliable and not too large, whereas for a larger network and/or moderate reliability, the splitting approach is more effective. In [65] we presented the main ideas behind another approach for the same kind of problem, where we generalize the Creation Process idea to a multi-level setting, on top of which we explore the behavior of the Splitting method, with very good results when the system is highly reliable.

Importance sampling (IS) is the main used other technique, but it requires a fine tuning of parameters. This has been applied in [60] to urban passenger rail systems, that are large scale systems comprising highly reliable redundant structures and logistics (e.g., spares or repair personnel availability, inspection protocols, etc). To meet the strict contractual obligations, steady state unavailability of such systems needs to be accurately estimated as a measure of a solution's life cycle costs. We use Markovian Stochastic Petri Nets models to conveniently represent the systems. We propose a multi-level Cross-Entropy optimization scheme, where we exploit the regenerative structure in the underlying continuous time Markov chain and to determine optimal

(IS) rates in the case of rare events. The CE scheme is used in a pre-simulation and applied to failure transitions of the Markovian SPN models only. The proposed method divides a rare problem into a series of less rare problems by considering increasingly rare component failures. In the first stage a standard regenerative simulation is used for non-rare system failures. At each subsequent stage, the rarity is progressively increased (by decreasing the failure rates of components) and the IS rates of transitions obtained from the previous problem are used at the current stage. The final pre-simulation stage provides a vector of IS rates that are optimized and are used in the main simulation. The experimental results showed bounded relative error property as the rarity of the original problem increases, and as a consequence a considerable variance reduction and gain (in terms of work normalized variance).

In [68] and [29] we introduced the idea that the problem with the standard estimator in the case of rare events is not the estimator itself but its usual implementation, and we describe an efficient way of implementing it in order to be able to perform estimations that, otherwise, are out of reach following the crude approach as it is usually coded. The idea is to reduce the time needed to sample the standard estimator and not the variance. The interest in taking this viewpoint is also discussed.

In [30] we gave a tutorial on Monte Carlo techniques for rare event analysis, where the basic Splitting and Importance Sampling families of estimators are presented, together with the Zero Variance subfamily of the first class of techniques, plus other methods such as the Recursive Variance Reduction approach.

Taking dependence into account. The Marshall-Olkin copula model has emerged as the standard tool for capturing dependency between components in failure analysis in reliability. In this model shocks arise at exponential random times, that affect one or several components inducing a natural correlation in the failure process. However, the method presents the classic "curse of dimensionality" problem. Marshall-Olkin models are usually intended to be applied to design a network before its construction, therefore it is natural to assume that only partial information about failure behavior can be gathered, mostly from similar existing networks. To construct such a MO model, we propose in [19] an optimization approach in order to define the shock's parameters in the copula, with the goal of matching marginal failures probabilities and correlations between these failures. To deal with the exponential number of parameters of this problem, we use a column-generation technique. We also discuss additional criteria that can be incorporated to obtain a suitable model. Our computational experiments show that the resulting approach produces a close estimation of the network reliability, especially when the correlation between component failures is significant.

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. We study in [18] the lattice structure of random number generators of the 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. Its matrix entries were also selected in a special way to allow a fast implementation and this has an impact on the lattice structure. 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.

7.5. Wireless Networks

Participants: Imad Alawe, Gerardo Rubino, Yassine Hadjadj-Aoul, Patrick Maillé.

Congestion control. The explosive growth of connected objects is certainly one of the most important challenges facing operators' network infrastructures. Although it has been foreseen for a very long time, it is still not clear how to support such huge number of devices efficiently.

A smarter planning of dedicated access slots would certainly limit the burden, at the access network, but remains insufficient since some equipments react to events which cannot be timed. Moreover, barring some Internet of Things (IoT) devices from accessing the network is very efficient; nevertheless, efficiency is generally linked to precise knowledge of the number of contending devices. Though, before connection establishment, the terminals are invisible to access points and, therefore, it is very difficult to estimate their number. A lower bound of backlogged devices can be determined. However, underestimating this number may lead to a congestion collapse whereas an overestimation implies underutilization of resources. In [14], we propose a lightweight change to the standard to accurately reveal the state of network congestion by overloading connections' requests with the number of access attempts (number of times the device has been barred as well as number of attempts). Using such information, we propose an accurate recursive estimator of the number of devices. The obtained results demonstrated that the proposed solution not only makes it possible to estimate the number of equipments much better than existing techniques, but also allows determining precisely the number of blocked equipments.

Even if the support of IoT objects represents a real challenge to the access, as mentioned above, it is nevertheless important to support them effectively in the core network, this is one of the requirements of 5G networks. 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 [34] 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 via computer system indicate the superiority of our algorithm in ensuring fair load balancing while scaling dynamically with the traffic load.

Energy efficiency. The use of Low Power Wide Area Networks (LPWANs) is growing due to their advantages in terms of low cost, energy efficiency and range. Although LPWANs attract the interest of industry and network operators, it faces certain constraints related to energy consumption, network coverage and quality of service. We demonstrate in [51] the possibility to optimize the performance of the LoRaWAN (Long Range Wide Area Network) technology, one of the most widely used LPWAN technology. We suggest that nodes use light-weight learning methods, namely, multi-armed bandit algorithms, to select the communication parameters (spreading factor and emission power). Extensive simulations show that such learning methods allow to manage the trade-off between energy consumption and packet loss much better than an Adaptive Data Rate (ADR) algorithm adapting spreading factors and transmission powers on the basis of Signal to Interference and Noise Ratio (SINR) values.

Vehicular networks. According to recent forecasts, constant population growth and urbanization will bring an additional load of 2.9 billion vehicles to road networks by 2050. This will certainly lead to increased air pollution concerns, highly congested roads putting more strain on an already deteriorated infrastructure, and may increase the risk of accidents on the roads as well. Therefore, to face these issues we need not only to promote the usage of smarter and greener means of transportation but also to design advanced solutions that leverage the capabilities of these means along with modern cities' road infrastructure to maximize its utility. To this end, we propose, in [47], an original Cognitive Radio inspired algorithm, named CRITIC, that aims to mimic the principle of Cognitive Radio technology used in wireless networks on road networks. The key idea behind CRITIC is to temporarily grant regular vehicles access to priority (e.g., bus or carpool) lanes whenever they are underutilized in order to reduce road traffic congestion. In [50], we explore novel ways of utilizing inter-vehicle and vehicle to infrastructure communication technology to achieve a safe and efficient lane change manoeuvre for Connected and Autonomous Vehicles (CAVs). The need for such new protocols is due to the risk that every lane change manoeuvre brings to drivers and passengers lives in addition to its negative impact on congestion level and resulting air pollution, if not performed at the right time and using the appropriate speed. To avoid this risk, we design two new protocols, one is built upon and extends an existing protocol, and it aims to ensure safe and efficient lane change manoeuvre, while the second is an original solution inspired from mutual exclusion concept used in operating systems. This latter complements the former by exclusively granting lane change permission in a way that avoids any risk of collision.

Adaptive Allocation for Virtual Network Functions in Wireless Access Networks Network Function Virtualization (NFV) is deemed as a mean to simplify deployment and management of network and telecommunication services. With wireless access networks, NFV has to take into account the radio resources at wireless nodes in order to provide an end-to-end optimal virtual network function (VNF) allocation. This topic has been well-studied in existing literature, however, the effects of variations of networks over time have not been addressed yet. In [67], we provide a model of the adaptive and dynamic VNF allocation problem considering VNF migration. Then, we formulate the optimisation problem as an Integer Linear Programming (ILP) problem and provide a heuristic algorithm for allocating multiple service function chains (SFCs).The proposed approach allows SFCs to be reallocated so as to obtain the optimal solution over time. The results confirm that the proposed algorithm is able to optimize the network utilization while limiting the reallocation of VNFs which could interrupt services.

7.6. Future networks and architectures

Service placement. The growing need for a simplified management of network infrastructures has recently led to the emergence of software-defined networking (SDN) and network function virtualization (NFV) paradigms. These concepts have, however, introduced new challenges and notably the service placement problem.

The problem of service placement, in its simplest version, consists in placing virtual machines in a network infrastructure. This placement sometimes also consists in placing flows, and therefore refers to a routing problem. In an even more elaborate version, this consists of combining the two approaches, which comes down to placing a service chain. In one of the most elaborated versions, it is necessary to add to the placement the dynamicity of the services to be deployed.

In [62] we demonstrate the feasibility of an extended and flexible Software Defined Network (SDN) control plane that allows to overcome the limitations of the Openflow protocol by achieving distributed and intelligent network services in SDN networks. This extended control plane is designed according to the following reference guidelines: 1) the concept of generic and programmable network nodes usually known as “white boxes”. They integrate a generic engine to execute the service and a library of elementary components as basic building blocks of any services; 2) a fine grained decomposition logic of network services into elementary components, which allows the services to be designed and customized on the fly using these building blocks available on each network node in libraries; 3) a mechanism for re-configuring or redefinition on the fly of the network services on generic nodes without service interruption; 4) some smart elementary agents called SDN controllers elements to provide and distribute the intelligence necessary to interact with the data plane at different levels of locality. This SDN control plane is illustrated in a proof of concept with the implementation of a distributed monitoring service use case. The monitoring service can act and evolve in a differentiated manner in the network depending on traffic requirements and monitoring usage.

In [63] 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.

As mentioned above, service placement is often closely linked to the routing problem. The latter is all the more complex when it comes to optimizing several metrics at once. An intuitive method is formulating the problem as an Integer Linear Programming and solving it by an approximation algorithm. This method tends to have a specific design and usually suffers from unacceptable computational delays to provide a sub-optimal solution.

Genetic algorithms (GAs) are deemed as a promising solution to cope with highly complex optimization problems. However, the convergence speed and the quality of solutions should be addressed in order to fit into practical implementations. In [28], we propose a genetic algorithm-based mechanism to address the multi-constrained multi-objective routing problem. Using a repairer to reduce the search space to feasible solutions, results confirm that the proposed mechanism is able to find the Pareto-optimal solutions within a short run-time.

Recent studies confirm the ability of Deep Reinforcement Learning (DRL) in solving complex routing problems; however, its performance in the network with QoS-sensitive flows has not been addressed. In [59], we exploit a DRL agent with convolutional neural networks in the context of SDN networks in order to enhance the performance of QoS-aware routing. The obtained results demonstrate that the proposed approach is able to improve the performance of routing configurations significantly even in complex networks.

One big advantage of using Virtual Network Functions (VNF), is the possibility of dynamically scaling, depending on traffic load (i.e. instantiate new resources to VNF when the traffic load increases, and reduce the number of resources when the traffic load decreases). In [13] and [36], we propose a novel mechanism to scale 5G core network resources by anticipating traffic load changes through forecasting via Machine Learning (ML) techniques. The traffic load forecast is achieved by using and training a Neural Network on a real dataset of traffic arrival in a mobile network. Two techniques were used and compared: (i) Recurrent Neural Network (RNN), more specifically Long Short Term Memory Cell (LSTM); and (ii) Deep Neural Network (DNN). Simulation results showed that the forecast-based scalability mechanism outperforms the threshold-based solutions, in terms of latency to react to traffic change, and delay to have new resources ready to be used by the VNF to react to traffic increase.

Content Centric Networking. During the last decade, Internet Service Providers (ISPs) infrastructure has undergone a major metamorphosis driven by new networking paradigms, namely: SDN and NFV. The upcoming advent of 5G will certainly represent an important achievement of this evolution. In this context, static (planning) or dynamic (on-demand) caching resources placement remains an open issue. In [40], we propose a new technique to achieve the best trade-off between the centralization of resources and their distribution, through an efficient placement of caching resources. To do so, we model the cache resources allocation problem as a multi-objective optimization problem, which is solved using Greedy Randomized Adaptive Search Procedures (GRASP). The obtained results confirm the quality of the outcomes compared to an exhaustive search method and show how a cache allocation solution depends on the network's parameters and on the performance metrics that we want to optimize.

Analysis of transmission schemes in networks of sensors. In Wireless Sensor Networks (WSNs), each node typically transmits several control and data packets in a contention fashion to the sink. In the literature, different adaptive schemes have been proposed for this purpose. Their common goal is to offer QoS guarantees in terms of system lifetime (related to energy consumption) and reporting delay (related to the cluster formation delay). In [61], we analyze and study three unscheduled transmission schemes for control packets in three cluster-based architectures: Fixed Scheme (FS), Adaptive by Estimation Scheme (AES) and Adaptive by Gamma Scheme (AGS). Based on the numerical results, we show that the threshold values are just as important in the system design as the actual value of the transmission probability in adaptive schemes (AES and AGS), to achieve QoS guarantees.

P2P networks for Video on Demand (VoD) services. In [25] we describe a novel scheme that efficiently distributes the resources that are provided by seeds in a P2P network for Video on Demand (VoD) services. In the proposed scheme, that we have called Prioritized-Windows Distribution (PWD), the amount of seed's resources assigned to a downloader depends on its current progress in the process of downloading the video. We demonstrate through a fluid model analysis and Markov chain numerical evaluations that PWD improves the P2P network performance in terms of the level of cooperation that is required from the seeds to keep the system under abundance conditions. Additionally, we analyze the performance of the system as a function of the initial playback delay, a parameter that highly influences the Quality of Service (QoS) as perceived by the users, and our results show that PWD also improves it.

DIVERSE Project-Team

6. New Results

6.1. Results on Variability modeling and management

6.1.1. 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. Publication at Empirical Software Engineering: [25] See also, in the rest of the report, the work on *Multimorphic Testing* that actually relies on variability techniques.

6.1.2. 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. Publication at SPLC (journal first) [29], workshop SPLTea'18 http://spltea.irisa.fr/ and repository: https://teaching.variability.io

6.1.3. Variability and machine learning

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 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. Publications: Temple et al. *Towards Adversarial Configurations for Software Product Lines*https://arxiv.org/abs/1805.12021, VaryLaTeX [30] a variability and learning-based tool to generate relevant paper variants written in latex.

TUXML (Tux is the mascotte of the Linux Kernel while ML stands for statistical machine learning). The goal of TuxML is to predict properties of Linux Kernel configurations (e.g., does the kernel compile? what's its size? does it boot?). The Linux Kernel provides near 15000 configuration options: there is an infinity of different kernels. As we cannot compile, measure, and observe all combinations of options (aka configurations), we're trying to learn Linux kernel properties out of a sample of configurations. The TuxML project is developing tools, mainly based on Docker and Python, to massively compile and gather data about thousand of configuration kernels https://github.com/TuxML/.

In general, we are currently exploring the use of machine learning for variability-intensive systems in the context of VaryVary ANR project https://varyvary.github.io.

6.2. Results on Software Language Engineering

6.2.1. 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.

6.2.2. Trace Comprehension Operators for Executable DSLs

Recent approaches contribute facilities to breathe life into metamodels, thus making behavioral models directly executable. Such facilities are particularly helpful to better utilize a model over the time dimension, e.g., for early validation and verification. However, when even a small change is made to the model, to the language definition (e.g., semantic variation points), or to the external stimuli of an execution scenario, it remains difficult for a designer to grasp the impact of such a change on the resulting execution trace. This prevents accessible trade-off analysis and design-space exploration on behavioral models. In [44], we propose a set of formally defined operators for analyzing execution traces. The operators include dynamic trace filtering, trace comparison with diff computation and visualization, and graph-based view extraction to analyze cycles. The operators are applied and validated on a demonstrative example that highlight their usefulness for the comprehension specific aspects of the underlying traces.

6.2.3. Model Transformation Reuse across Metamodels

Model transformations (MTs) are essential elements of model-driven engineering (MDE) solutions. MDE promotes the creation of domain-specific metamodels, but without proper reuse mechanisms, MTs need to be developed from scratch for each new metamodel. In [32], awarded by the **best paper award at ICMT 2018**, we classify reuse approaches for MTs across different metamodels and compare a sample of specific approaches – model types, concepts, a-posteriori typing, multilevel modeling, and design patterns for MTs – with the help of a feature model developed for this purpose, as well as a common example. We discuss strengths and weaknesses of each approach, provide a reading grid used to compare their features, and identify gaps in current reuse approaches.

6.2.4. Modular Language Composition for the Masses

The goal of modular language development is to enable the definition of new languages as assemblies of pre-existing ones. Recent approaches in this area are plentiful but usually suffer from two main problems: either they do not support modular language composition both at the specification and implementation levels, or they require advanced knowledge of specific paradigms which hampers wide adoption in the industry. In [36], awarded by the **best artefact award at SLE 2018**, we introduce a non-intrusive approach to modular development of language concerns with well-defined interfaces that can be composed modularly at the specification and implementation levels. We present an implementation of our approach atop the Eclipse Modeling Framework, namely Alex-an object-oriented metalanguage for semantics definition and language composition. We evaluate Alex in the development of a new DSL for IoT systems modeling resulting from the composition of three independently defined languages (UML activity diagrams, Lua, and the OMG Interface Description Language). We evaluate the effort required to implement and compose these languages using Alex with regards to similar approaches of the literature.

6.2.5. Shape-Diverse DSLs

Domain-Specific Languages (DSLs) manifest themselves in remarkably diverse shapes, ranging from internal DSLs embedded as a mere fluent API within a programming language, to external DSLs with dedicated syntax and tool support. Although different shapes have different pros and cons, combining them for a single language is problematic: language designers usually commit to a particular shape early in the design process, and it is hard to reconsider this choice later. In the new ideas paper [33] awarded as the **best new ideas paper at SLE 2018**, we envision a language engineering approach enabling (i) language users to manipulate language constructs in the most appropriate shape according to the task at hand, and (ii) language designers to combine the strengths of different technologies for a single DSL. We report on early experiments and lessons learned building Prism, our prototype approach to this problem. We illustrate its applicability in the engineering of a simple shape-diverse DSL implemented conjointly in Rascal, EMF, and Java. We hope that our initial contribution will raise the awareness of the community and encourage future research.

6.2.6. Fostering metamodels and grammars

Advanced and mature language workbenches have been proposed in the past decades to develop Domain-Specific Languages (DSL) and rich associated environments. They all come in various flavors, mostly depending on the underlying technological space (e.g., grammarware or modelware). However, when the time comes to start a new DSL project, it often comes with the choice of a unique technological space which later bounds the possible expected features. In [37], we introduce NabLab, a full-fledged industrial environment for scientific computing and High Performance Computing (HPC), involving several metamodels and grammars. Beyond the description of an industrial experience of the development and use of tool-supported DSLs, we report in this paper our lessons learned, and demonstrate the benefits from usefully combining metamodels and grammars in an integrated environment.

6.2.7. Automatic Production of End User Documentation for DSLs

The development of DSLs requires a significant software engineering effort: editors, code generators, etc., must be developed to make a DSL usable. Documenting a DSL is also a major and time-consuming task required to promote it and address its learning curve. Recent research work in software language engineering focus on easing the development of DSLs. This work focuses on easing the production of documentation of textual DSLs [27], [17]. The API documentation domain identified challenges we adapted to DSL documentation. Based on these challenges we propose a model-driven approach that relies on DSL artifacts to extract information required to build documentation. Our implementation, called Docywood, targets two platforms: Markdown documentation for static web sites and Xtext code fragments for live documentation while modeling. We used Docywood on two DSLs, namely ThingML and Target Platform Definition. Feedback from end users and language designers exhibits qualitative benefits of the proposal with regard to the DSL documentation challenges. End user experiments conducted on ThingML and Target Platform Definition show benefits on the correctness of the created models when using Docywood on ThingML.

6.3. 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 of resources for dynamically adaptive system, one in the field of temporal context model for dynamically adaptive system and a last one to improve the exploration of hidden real-time structures of programming behavior at runtime.

6.3.1. Resource-aware models@runtime layer for dynamically adaptive system

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 a whole development team. This year, we mainly explore two main axes.

In the first one, we try to improve the proposed models that could be used at runtime to improve resource usage in two domains: cloud computing and energy [34]. In the cloud computing domain, we try to improve resources usage in providing models to cloud provider to allow the reselling of unused resources to peers. Indeed, although Cloud computing techniques have reduced the total cost of ownership thanks to virtualization, the average usage of resources (e.g., CPU, RAM, Network, I/O) remains low. To address such issue, one may sell unused resources. Such a solution requires the Cloud provider to determine the resources available and estimate their future use to provide availability guarantees. In this work, we propose a technique that uses machine learning algorithms (Random Forest, Gradient Boosting Decision Tree, and Long Short Term Memory) to forecast 24-hour of available resources at the host level. Our technique relies on the use of quantile regression to provide a flexible trade-off between the potential amount of resources to reclaim and the risk of SLA violations. In addition, several metrics (e.g., CPU, RAM, disk, network) were predicted to provide exhaustive availability guarantees. Our methodology was evaluated by relying on four in production data center traces and our results show that quantile regression is relevant to reclaim unused resources. Our approach may increase the amount of savings up to 20% compared to traditional approaches.
In the energy domain, we work at proposing models that could be used at runtime to improve self-consumption of renewable energies [46]. Self-consumption of renewable energies is defined as electricity that is produced from renewable energy sources, not injected to the distribution or transmission grid or instantaneously withdrawn from the grid and consumed by the owner of the power production unit or by associates directly contracted to the producer. Designing solutions in favor of self-consumption for small industries or city districts is challenging. It consists in designing an energy production system made of solar panels, wind turbines, batteries that fit the annual weather prediction and the industrial or human activity. In this context, this we highlight the essentials of a domain specific modeling language designed to let domain experts run their own simulations.

6.3.2. A Temporal Model for Interactive Diagnosis of Adaptive Systems

The evolving complexity of adaptive systems impairs our ability to deliver anomaly-free solutions. Fixing these systems require a deep understanding on the reasons behind decisions which led to faulty or suboptimal system states. Developers thus need diagnosis support that trace system states to the previous circumstances targeted requirements, input context that had resulted in these decisions. However, the lack of efficient temporal representation limits the tracing ability of current approaches. To tackle this problem, we describe a novel temporal data model to represent, store and query decisions as well as their relationship with the knowledge (context, requirements, and actions) [38]. We validate our approach through a use case based-on the smart grid at Luxembourg.

Based on this work, we also enable a models@runtime approach in which we integrate the time required for a reconfiguration action to achieve the expected impact [39]. Indeed in most of the MAPE-K loop system, unfinished actions as well as their expected effects over time are not taken into consideration in MAPE-K loop processes, leading upcoming analysis phases potentially take sub-optimal actions. In this work, we propose an extended context model for MAPE-K loop that integrates the history of planned actions as well as their expected effects over time into the context representations. This information can then be used during the upcoming analysis and planning phases to compare measured and expected context metrics. We demonstrate on a cloud elasticity manager case study that such temporal action-aware context leads to improved reasoners while still be highly scalable.

6.3.3. Detection and analysis of behavioral T-patterns in debugging activities

A growing body of research in empirical software engineering applies recurrent patterns analysis in order to make sense of the developers' behavior during their interactions with IDEs. However, the exploration of hidden real-time structures of programming behavior remains a challenging task. In this work [40], we investigate the presence of temporal behavioral patterns (T-patterns) in debugging activities using the THEME software. Our preliminary exploratory results show that debugging activities are strongly correlated with code editing, file handling, window interactions and other general types of programming activities. The validation of our T-patterns detection approach demonstrates that debugging activities are performed on the basis of repetitive and well-organized behavioral events. Furthermore, we identify a large set of T-patterns that associate debugging activities with build success, which corroborates the positive impact of debugging practices on software development.

6.4. 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) observations of software diversity mainly through browser fingerprinting, and (ii) software testing to study and assess the validity of software.

6.4.1. Privacy and Security

6.4.1.1. FP-STALKER: Tracking Browser Fingerprint Evolutions

Browser fingerprinting has emerged as a technique to track users without their consent. Unlike cookies, fingerprinting is a stateless technique that does not store any information on devices, but instead exploits unique combinations of attributes handed over freely by browsers. The uniqueness of fingerprints allows them to be used for identification. However, browser fingerprints change over time and the effectiveness of tracking users over longer durations has not been properly addressed. In this work [42], we show that browser fingerprints tend to change frequently—from every few hours to days—due to, for example, software updates or configuration changes. Yet, despite these frequent changes, we show that browser fingerprints can still be linked, thus enabling long-term tracking. FP-STALKER is an approach to link browser fingerprint evolutions. It compares fingerprints to determine if they originate from the same browser. We created two variants of FP-STALKER, a rule-based variant that is faster, and a hybrid variant that exploits machine learning to boost accuracy. To evaluate FP-STALKER, we conduct an empirical study using 98,598 fingerprints we collected from 1,905 distinct browser instances. We compare our algorithm with the state of the art and show that, on average, we can track browsers for 54.48 days, and 26% of browsers can be tracked for more than 100 days.

6.4.1.2. Hiding in the Crowd: an Analysis of the Effectiveness of Browser Fingerprinting at Large Scale

Browser fingerprinting is a stateless technique, which consists in collecting a wide range of data about a device through browser APIs. Past studies have demonstrated that modern devices present so much diversity that fingerprints can be exploited to identify and track users online. With this work [35], we want to evaluate if browser fingerprinting is still effective at uniquely identifying a large group of users when analyzing millions of fingerprints over a few months.We analyze 2,067,942 browser fingerprints collected from one of the top 15 French websites. The observations made on this novel dataset shed a newlight on the ever-growing browser fingerprinting domain. The key insight is that the percentage of unique fingerprints in this dataset is much lower than what was reported in the past: only 33.6% of fingerprints are unique by opposition to over 80% in previous studies. We show that non-unique fingerprints tend to be fragile. If some features of the fingerprint change, it is very probable that the fingerprint will become unique. We also confirm that the current evolution of web technologies is benefiting users' privacy significantly as the removal of plugins brings down substantively the rate of unique desktop machines.

6.4.1.3. User Controlled Trust and Security Level of Web Real-Time Communications

In this work [16], we propose three main contributions. In our first contribution we study the WebRTC identity architecture and more particularly its integration with existing authentication delegation protocols. This integration has not been studied yet. To fill this gap, we implement components of the WebRTC identity architecture and comment on the issues encountered in the process. We then study this specification from a privacy perspective an identify new privacy considerations related to the central position of identity provider. In our second contribution, we aim at giving more control to users. To this end, we extend the WebRTC specification to allow identity parameters negotiation. We then propose a web API allowing users to choose their identity provider in order to authenticate on a third-party website. We validate our propositions by presenting prototype implementations. Finally, in our third contribution, we propose a trust and security model of a WebRTC session to help non-expert users to better understand the security of their WebRTC session. Our proposed model integrates in a single metric the security parameters used in the session establishment, the encryption parameters for the media streams, and trust in actors of the communication setup as defined by the user. We conduct a preliminary study on the comprehension of our model to validate our approach.

6.4.2. Software Testing

6.4.2.1. A Comprehensive Study of Pseudo-tested Methods

Pseudo-tested methods are defined as follows: they are covered by the test suite, yet no test case fails when the method body is removed, i.e., when all the effects of this method are suppressed. This intriguing concept was coined in 2016, by Niedermayr and colleagues [88], who showed that such methods are systematically present, even in well-tested projects with high statement coverage. This work presents a novel analysis of pseudo-tested methods [28]. First, we run a replication of Niedermayr's study with 28K+ methods, enhancing its external validity thanks to the use of new tools and new study subjects. Second, we perform a systematic characterization of these methods, both quantitatively and qualitatively with an extensive manual analysis of 101 pseudo-tested methods. The first part of the study confirms Niedermayr's results: pseudo-tested methods exist in all our subjects. Our in-depth characterization of pseudo-tested methods leads to two key insights: pseudo-tested methods are significantly less tested than the other methods; yet, for most of them, the developers would not pay the testing price to fix this situation. This calls for future work on targeted test generation to specify those pseudo-tested methods without spending developer time.

This work uses Descartes is a tool that implements extreme mutation operators and aims at finding pseudo-tested methods in Java projects [43]. It leverages the efficient transformation and runtime features of PITest.

6.4.2.2. Automatic Test Improvement with DSpot: a Study with Ten Mature Open-Source Projects

In the literature, there is a rather clear segregation between manually written tests by developers and automatically generated ones. In this work [23], we explore a third solution: to automatically improve existing test cases written by developers. We present the concept, design and implementation of a system called DSpot, that takes developer-written test cases as input (JUnit tests in Java) and synthesizes improved versions of them as output. Those test improvements are given back to developers as patches or pull requests, that can be directly integrated in the main branch of the test code base. We have evaluated DSpot in a deep, systematic manner over 40 real-world unit test classes from 10 notable and open-source software projects. We have amplified all test methods from those 40 unit test classes. In 26/40 cases, DSpot is able to automatically improve the test under study, by triggering new behaviors and adding new valuable assertions. Next, for ten projects under consideration, we have proposed a test improvement automatically synthesized by DSpot to the lead developers. In total, 13/19 proposed test improvements were accepted by the developers and merged into the main code base. This shows that DSpot is capable of automatically improving unit-tests in real-world, large scale Java software.

6.4.2.3. Multimorphic Testing

The functional correctness of a software application is, of course, a prime concern, but other issues such as its execution time, precision , or energy consumption might also be important in some contexts. Systematically testing these quantitative properties is still extremely difficult, in particular, because there exists no method to tell the developer whether such a test set is "good enough" or even whether a test set is better than another one. This work [41] proposes a new method, called Multimorphic testing, to assess the relative effectiveness of a test suite for revealing performance variations of a software system. By analogy with mutation testing, our core idea is to vary software parameters, and to check whether it makes any difference on the outcome of the tests: i.e. are some tests able to " kill " bad morphs (configurations)? Our method can be used to evaluate the quality of a test suite with respect to a quantitative property of interest, such as execution time or computation accuracy.

6.4.2.4. User Interface Design Smell: Automatic Detection and Refactoring of Blob Listeners

User Interfaces (UIs) intensively rely on event-driven programming: widgets send UI events, which capture users' interactions, to dedicated objects called controllers. Controllers use several UI listeners that handle these events to produce UI commands. In this work [20], we reveal the presence of design smells in the code that describes and controls UIs. We then demonstrate that specific code analyses are necessary to analyze and refactor UI code, because of its coupling with the rest of the code. We conducted an empirical study on four large Java Swing and SWT open-source software systems: Eclipse, JabRef, ArgouML, and FreeCol. We study to what extent the number of UI commands that a UI listener can produce has an impact on the change- and fault-proneness of the UI listener code. We develop a static code analysis for detecting UI commands in the code. We identify a new type of design smell, called Blob listener that characterizes UI listeners that can produce more than two UI commands. We propose a systematic static code analysis procedure that searches for Blob listener that we implement in InspectorGuidget. We conducted experiments on the four software systems for which we manually identified 53 instances of Blob listener. The results exhibit a precision of 81.25 % and a recall of 98.11 %. We then developed a semi-automatically and behavior-preserving refactoring process to remove Blob listeners. 49.06 % of the Blob listeners were automatically refactored. Patches for JabRef, and FreeCol have been accepted and merged. Discussions with developers of the four software systems assess the relevance of the Blob listener.

EASE Team

6. New Results

6.1. Smart City and ITS

Participants: Indra Ngurah, 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 autonomy covers 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 puts high constraints on the communications. Automated vehicles are just an of smart city automated processes that will have to share information within the surrounding to make their decisions.

In the continuation of our previous activities on the SEAS project, we contributed to the specification of the hybrid (ITS-G5 + Cellular) communication architecture of the French field operation test project SCOOP@F. The proposed solution relies on the MobileIP family of standards and the CALM architecture we contributed to standardize at IETF and ISO. On this topic our contribution mainly focussed on bringing concepts from the state of the art to real equipments. This includes the proposal of provisioning mechanisms to automatically configure and update security materials (ie. certificates and pseudonyms) to ensure an acceptable balance between confidentially, non repudiation and privacy. Extending these works on the On Board Unit (OBU) side, Rodrigo Silva's proposed an architecture and a decision making algorithm to optimize the binding of data flows to available networks while cars are moving [2]. In another field of applications, Indra Ngurah proposed a new routing algorithm for Delay Tolerant Application the context of smart cities[7].

6.2. Opportunistic and local communication/information sharing

Participants: Yoann Maurel, Jules Desjardin, Paul Couderc [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. Communication between devices and information sharing is a key to achieve this goal. In recent years, many products have been developed based on the cloud. This raises privacy and network access issues.We believe that communication and information sharing should be able to take place directly when possible, more efficient, confidential, or when the network is not available. To achieve this, applications must be provided with technologies that enable the opportunistic and rapid exchange of information based on simple and widespread technologies.

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.

To ease communication, we developed an opportunistic communication system that does not need any connection between participants, nor any preexisting infrastructure (e.g. WiFi network). The only condition for participants to exchange information is that they are close enough to each other. The communication protocol has been implemented over Bluetooth Low Energy advertisement packets. This protocol has been ported to ESP8266, ESP32 and Android platform. To ease information sharing, 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. The communication protocol has been extended to allow REST-like operations. The implementation in C of the protocol and a storage base was done in such a way as to take little memory and run on small chips (ESP8266, ESP32). The storage base can be accessed either opportunistically using the BLE protocol or via a COAP protocol for longer or bigger exchanges.

We have started validation work with a few applications, in particular regarding energy aspects and scalability with respect to the communication load. We also tested the system for building opportunistic games (e.g, capture the flags) and information sharing mechanism (e.g, sharing information when two devices cross paths). We are currently working on structuring knowledge information in the continuity of what has been done in the team in the past and provide encryption mechanism.

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 EASE, 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 . It has developed expertise in vertical-axis wind turbines,photovoltaic trackers, heat pump and energy storage devices. An interesting aspect of renewable energies is that they can be produced locally, close to the consumers, thus considerably reducing infrastructures and distribution costs.The autonomy of sites with micro-generation capabilities is then greatly increased by self-consumption of locally produced energy.

Designing solutions in favor of self-consumption for small industries or city districts is challenging. It consists in designing an energy production system made of solar panels, wind turbines, batteries that fit the annual weather prediction and the industrial or human activity. This 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)? What methods to use to optimize consumption and, whenever possible, act on installations and activities to reduce energy costs?

We aim to design an integrated tool-suite to assist the engineers in dimensioning an Energy Management System (EMS) for an isolated site to reduce the construction of new network infrastructure and reduce its dependence on the grid. We advocate that the MDE is a very good candidate to integrate the various technological and business knowledge on the renewable energy production and consumption forecasting techniques, the planning of processes, energy costs, grid, and batteries. 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. Using a DSL and components that clearly separate the different concerns would avoid code redundancies and would facilitate the work of domain experts.

In 2018, we developed a prototype of the Energy Management System (EMS) and a complete DSL that enables experts to quickly integrate their knowledge and algorithms, and to provide a library of reusable components and algorithms. The DSL reflects the different aspects of site modeling: batteries, producers, grid, machines used, and activities performed. It provides the necessary information and constraints so that the EMS can propose an arrangement of activities that optimizes the consumption of renewable energy. The system can be improved by extending existing components or adding new ones. Some of these components are also able to play back historical data, which is a common use for sizing purposes. The prototype is made of 4500 lines of Java code, 1300 loc of Lua and 78k loc of generated files.

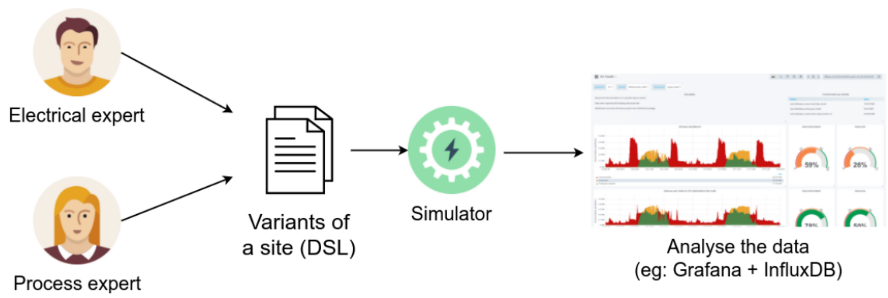


Figure 3. Experts express their concerns using the same DSL and can simulate various scenarios

This prototype has been tested in production to model agricultural sites. The great interest of our tool is that it enables to simulate easily a very wide range of situations and thus allows to determine quickly the best options. If we compare with the company's past practices, engineers mainly used homemade excel sheets and R script:

sharing information among experts was very difficult and detecting errors in site modeling was challenging. Building this domain specific language and its associated simulator saves lots of time and produces more precise results compared to the traditional manual approaches used before (see figure 3).

We are now conducting an experiment at several sites to see how adapting activities can improve production equipment profitability. This experience over a long period should provide us with relevant feedback on what can and cannot be requested from a site operator. This should allow us to use our tool not only to simulate upstream but also to make observations and recommendations on a weekly basis. We also want to improve the constraint systems to integrate the modeling of more resources (hot and cold water, number of employees, machine availability). Finally, we would like to explore how this model can be used as a basis for artificial intelligence algorithms to manage real-time operations.

This work has been published in the conference Models 2018 [11].

6.4. Location assessment from local observations

Participants: Yoann Maurel, Paul Couderc [contact].

Confidence in location is increasingly important in many applications, in particular for crowd-sensing systems integrating user contributed data/reports, and in augmented reality games. In this context, some users can have an interest in lying about their location, and this assumption has been ignored in several widely used geolocation systems because usually, location is provided by the user's device to enhance the user's experience. Two well known examples of applications vulnerable to location cheating are Pokemon Go and Waze.

Unfortunately, location reporting methods implemented in existing services are weakly protected: it is often possible to lie in simple cases or to emit signals that deceive the more cautious systems. For example, we have experimented simple and successful replay attacks against Google Location using this approach.

An interesting idea consists in requiring user devices to prove their location, by forcing a secure interaction with a local resource. This idea has been proposed by several works in the literature; unfortunately, this approach requires ad hoc deployment of specific devices in locations that are to be "provable".

We proposed an alternative solution using passive monitoring of Wi-Fi traffic from existing routers. The principle is to collect beacon timestamp observations (from routers) and other attributes to build a knowledge that requires frequent updates to remains valid, and to use statistical test to validate further observations sent by users. Typically, older data collected by a potential attacker will allow him to guess the current state of the older location for a limited timeframe, while the location validation server will get updates allowing him to determine a probability of cheating request.The main strength is its ability to work on existing Wi-Fi infrastructures, without specific hardware. Although it does not offer absolute proof, it makes attacks much more challenging and is simple to implement. The Figure 4 illustrates the basic architecture of the system.

This work was accepted for publication and will be presented at CCNC'2019 [5]. Several aspects would be interesting to further investigate, in particular using other attributes of Wi-Fi traffic beside beacon timestamps, and combining the timestamp solution with other type of challenges to propose a diversity of challenges for location validation servers.

6.5. Introducing Data Quality to the Internet of Things

Participants: Jean-Marie Bonnin, Jean-François Verdonck, Frédéric Weis [contact].

The Internet of Things (IoT) connects various distributed heterogeneous devices. Such Things sense and actuate their physical environment. The IoT pervades more and more into industrial environments forming the so-called Industrial IoT (IIoT). Especially in industrial environments such as smart factories, the quality of data that IoT devices provide is highly relevant. However, current frameworks for managing the IoT and exchanging data do not provide data quality (DQ) metrics. Pervasive applications deployed in the factory need to know how data are "good" for use. However, the DQ requirements differ from a process to another. Actually, specifying/expressing DQ requirements is a subjective tasks, depending to the specific needs of each targeted application. As an example this could mean how accurate a location of an object that is provided by an IoT

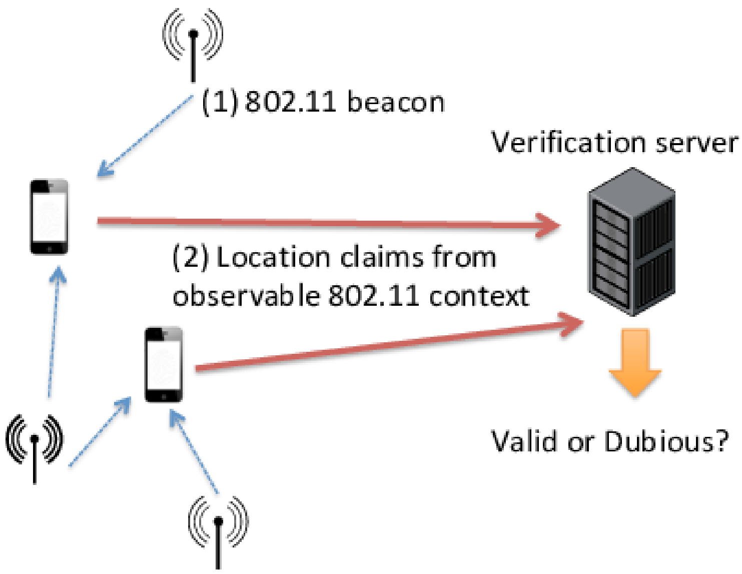


Figure 4. Location assessment from local observations: architecture

system differs from the actual physical position of the object. A Data Quality of 100% could mean that the value represents the actual position. A Data Quality of 0% could mean that the object is not at the reported position. In this example, the value 0% or 100% can be given by a specific software module that is able to filter raw data sent to the IoT system and to deliver the appropriate metric for Dev apps. Building ad hoc solutions for DQ management is perfectly acceptable. But the challenge of writing and deploying applications for the Internet of Things remains often understated. We believe that new approaches are needed, for thinking DQ management in the context of extremely dynamic systems that is the characteristic of the IoT.

In 2018, we started to define DQ software services that are able to query data and retrieve a collection of DQ metrics that the developer need. The goal is to enable developers to access, configure and tweak any DQ mechanisms in an easy way. Facilitating embedding of DQ capabilities will demand a new type of "endpoint" services, deployed to industrial pervasive environments. We obtained first results of our work towards establishing metrics and tools to enable IoT developers to know and use the quality of data they obtain from the IoT. Our approach combines continuous data analytics with modeling expected behavior of sensors in order to weight the inputs of different sensors to reduce the overall error. Key challenges of our work are semantic modeling of the data quality and modeling the expected behavior of sensors. We illustrated our approach at the example of localizing production robots in a factory. We demonstrated the potential of our first solutions with a demonstration at the AdHoc Now conference (see figure 5). We managed to significantly reduce the error introduced by faulty sensors. This should lead to publishing on both DQ and programming aspects of our approach.

This work has been done in collaboration with Technical University of Munich.

6.6. Risk Monitoring and Intrusion Detection

Participant: Jean-Marie Bonnin [contact].

Cyber-attacks on critical infrastructure such as electricity, gas, and water distribution, or power plants, are more and more considered to be a relevant and realistic threat to the European society. Whereas mature solutions

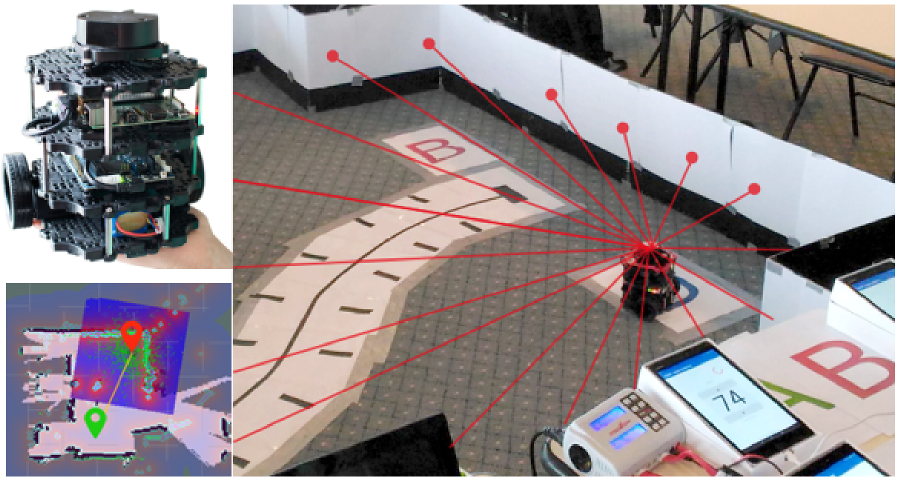


Figure 5. Demonstration at adhoc now 2018

like anti-malware applications, intrusion detection systems (IDS) and even intrusion prevention or self-healing systems have been designed for classic computer systems, these techniques have only been partially adapted to the world of Industrial Control Systems (ICS). This is most notably due to the fact that these industrial systems have been deployed several decades ago, when security was not such a big issue, and have not been replaced since. As a consequence, organisations and nations fall back upon risk management to understand the risks that they are facing. Today's trend is to combine risk management with real-time monitoring to enable prompt reactions in case of attacks. We provided techniques that assist security managers in migrating from a static risk analysis to a real-time and dynamic risk monitoring platform. Risk monitoring encompasses three steps [1]: the collection of risk-related information, the reporting of security events, and finally the inclusion of this real-time information into a risk analysis. The first step consists in designing agents that detect incidents in the system. They can either interpret the output of existing security appliances (such as firewalls), or monitor (part of) the system on their own. An intrusion detection system has been developed to this end, which focuses on an advanced persistent threat (APT) that particularly targets critical infrastructures. The second step copes with the translation of the obtained technical information in more abstract notions of risk, which can then be used in the context of a risk analysis. In the final step, the information collected from the various sources is correlated so as to obtain the risk faced by the entire system. A novel dependency model ties all parts together and thus constitutes the core of the risk monitoring framework we developed. The model is loosely based on attack trees, and can be intuitively visualized with boxes and arrows. Despite its visual simplicity, it allows risk assessors to encode the interdependencies of complex risk scenarios, and to quantify the risk originating from the former.

This work has been done in collaboration with University of Luxembourg.

6.7. Secure design of WoT services for Smart Cities

Participant: Jean-Marie Bonnin [contact].

The richness and the versatility of WebRTC, a new peer-to-peer, real-time and browser- based communication technology, allowed the imagination of new and innovative services. We analyzed the capabilities required to allow a participant in a WebRTC session to access the smart things belonging to his own environment as well as those of any other participant in the same session. The access to such environment (a Smart Space (SS)) can be either passive, for example by monitoring the contextual information provided by the sensors, or active by requesting the execution of commands by the actuators, or a mixture of both. This approach deserves attention because it allows to solve in an original way various issues such as allowing experts to remotely exercise and provide their expertises. From a technical point of view the issue is not trivial because it requires a smooth and mastered articulation between two different technologies: WebRTC and the Internet of Things (IoT) / Web of Things (WoT) [6].

We defined from scratch, of an architecture allowing a junction between WebRTC and the WoT. This architecture is illustrated through a set of innovative use cases. The latter relies essentially on a gateway connecting the two technologies. Since WebRTC is natively secure, its analysis allowed us to propose a set of mechanisms to secure the link between the gateway and the WebRTC client together with the access control to the SS. The implementation of an experimental prototype validates the feasibility of this approach. We also proposed a new smart home architecture encompassing several services, among them the healthcare and the energy management. The overall work targets the introduction of a real smart home, based in Aalborg University labs. Finally, we introduced an SDN controller in order to manage the various SSs that can be involved in a WebRTC session. The main idea consists in allowing an end-user to own more than one SS while keeping their management simple and effective. The principle of our approach consists in centralizing the decisions concerning the management of the various SSs. Due to the fact that routing concerns are intimately intertwined with those of security, the SDN clearly appears as a promising tool to solve these issues.

This work has been done in collaboration with IRISA-OCIF team.

KERDATA Project-Team

5. New Results

5.1. Convergence of HPC and Big Data

5.1.1. Large-scale logging for HPC and Big Data convergence

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

A critical objective set in this convergence context is to foster application portability across platforms. Cloud developers traditionally rely on purpose-specific services to provide the storage model they need for an application. In contrast, HPC developers have a much more limited choice, typically restricted to a centralized parallel file system for persistent storage. Unfortunately, these systems often offer low performance when subject to highly concurrent, conflicting I/O patterns.

This makes difficult the implementation of inherently concurrent data structures such as distributed shared logs. Shared log storage is indeed one of the storage models that are both unavailable and difficult to implement on HPC platforms using the available storage primitives. Yet, this data structure is key to applications such as computational steering, data collection from physical sensor grids, or discrete event generators. A shared log enables multiple processes to append data at the end of a single byte stream. Unfortunately, in such a case, the write contention at the tail of the log is among the worst-case scenarios for parallel file systems, yielding problematically low append performance.

In [25] we introduced SLoG, a shared log middleware providing a shared log abstraction over a parallel file system, designed to circumvent the aforementioned limitations. It features pluggable backends that enable it to leverage other storage models such as object stores or to transparently forward the requests to a shared log storage system when available (e.g., on cloud platforms). SLoG abstracts this complexity away from the developer, fostering application portability between platforms. We evaluated SLoG's performance at scale on a leadership-class supercomputer, using up to 100,000 cores. We measured append velocities peaking at 174 million appends per second, far beyond the capabilities of any shared log storage implementation on HPC platforms. For these reasons, we envision that SLoG could fuel convergence between HPC and big data.

5.1.2. Increasing small files access performance with dynamic metadata replication

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Small files are known to pose major performance challenges for file systems. Yet, such workloads are increasingly common in a number of Big Data Analytics workflows or large-scale HPC simulations. These challenges are mainly caused by the common architecture of most state-of-the-art file systems needing one or multiple metadata requests before being able to read from a file. Small input file size causes the overhead of this metadata management to gain relative importance as the size of each file decreases.

In our experiments, with small enough files, opening a file may take up to an order of magnitude more time than reading the data it contains. One key cause of this behavior is the separation of data and metadata inherent to the architecture of current file systems. Indeed, to read a file, a client must first retrieve the metadata for all folders in its access path, that may be located on one or more metadata servers, to check that the user has the correct access rights or to pinpoint the location of the data in the system. The high cost of network communication significantly exceeds the cost of reading the data itself.

In [22] we design a file system from the bottom up for small files without sacrificing performance for other workloads. This enables us to leverage some design principles that address the metadata distribution issues: consistent hashing and dynamic data replication. Consistent hashing enables a client to locate the data it seeks without requiring access to a metadata server, while dynamic replication adapts to the workload and replicates the metadata on the nodes from which the associated data is accessed. The former is often found in key-value stores, while the latter is mostly used in geo-distributed systems. These approaches allow to increase small file access performance up to one order of magnitude compared to other state-of-the-art file systems, while only causing a minimal impact on file write throughput.

5.1.3. 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 [28], we establish a lower bound for the duration of the commission operation. We then consider HDFS as a use case, and we show that our lower bound can be used to evaluate the performance of the commission algorithms. We show that the commission in HDFS can be greatly accelerated. With the highlights provided by our lower bound, we suggest improvements to speed the commission in HDFS.

In [29], we explore the possibility of relaxing the level of fault tolerance during the decommission in order to reduce the amount of data transfers needed before nodes are released, and thus return nodes to the resource manager faster. We quantify theoretically how much time and resources are saved by such a fast decommission strategy compared with a standard decommission. We establish lower bounds for the duration of the different phases of a fast decommission. We show that the method not only does not improve performance, but is also unsafe by nature.

In [24], we introduce Pufferbench, a benchmark for evaluating how fast one can scale up and down a distributed storage system on a given infrastructure and, thereby, how viably can one implement storage malleability on it. Besides, it can serve to quickly prototype and evaluate mechanisms for malleability in existing distributed storage systems. We validate Pufferbench against theoretical lower bounds for commission and decommission: it can achieve performance within 16% of them. We use Pufferbench to evaluate in practice these operations in HDFS: commission in HDFS could be accelerated by as much as 14 times! Our results show that: (1) the lower bounds for commission and decommission times we previously established are sound and can be approached in practice; (2) HDFS could handle these operations much more efficiently; most importantly, (3) malleability in distributed storage systems is viable and should be further leveraged for Big Data applications.

During a 3 months visit at Argonne National Lab, the design of an efficient rebalancing algorithm for rescaling operations have been started with Robert Ross. We use the rescaling operation to rebalance the load across the cluster. Performances cannot be sustained without minimizing the amount of data transferred per node, but also the amount of data stored per node. We evaluate a heuristic and show that good approximations of the optimal solutions can be achieved in reasonable time.

5.2. Scalable stream storage

5.2.1. KerA ingestion and storage

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu.

Big Data is now the new natural resource. Current state-of-the-art Big Data analytics architectures are built on top of a three layer stack: data streams are first acquired by the ingestion layer (e.g., Kafka) and then they flow through the processing layer (e.g., Flink) which relies on the storage layer (e.g., HDFS) for storing aggregated data or for archiving streams for later processing. Unfortunately, in spite of potential benefits brought by specialized layers (e.g., simplified implementation), moving large quantities of data through specialized layers is not efficient: instead, data should be acquired, processed and stored while minimizing the number of copies.

We argue that a plausible path to follow to alleviate from previous limitations is the careful design and implementation of the KerA unified architecture for stream ingestion and storage which can lead to the optimization of the processing of Big Data applications. This approach minimizes data movement within the analytics architecture, finally leading to better utilized resources. We identify a set of requirements for a unified stream ingestion/storage engine. We explain the impact of the different Big Data architectural choices on end-to-end performance. We propose a set of design principles for a scalable, unified architecture for data ingestion and storage: (1) dynamic partitioning based on semantic grouping and sub- partitioning, which enables more flexible and elastic management of stream partitions; (2) lightweight offset indexing (i.e., reduced stream offset management overhead) optimized for sequential record access; (3) adaptive and fine-grained replication to trade-off in-memory storage with performance (low-latency and high throughput with durability). We implement and evaluate the KerA prototype with the goal of efficiently handling diverse access patterns: low-latency access to streams and/or high throughput access to unbounded streams and/or objects [21].

5.2.2. Tailwind: fast and atomic RDMA-based replication

Participants: Yacine Taleb, Gabriel Antoniu.

Replication is essential for fault-tolerance. However, in in-memory systems, it is a source of high overhead. Remote direct memory access (RDMA) is attractive to create redundant copies of data, since it is low-latency and has no CPU overhead at the target. However, existing approaches still result in redundant data copying and active receivers. To ensure atomic data transfers, receivers check and apply only fully received messages.

Tailwind is a zero-copy recovery-log replication protocol for scale-out in-memory databases. Tailwind is the first replication protocol that eliminates *all* CPU-driven data copying and fully bypasses target server CPUs, thus leaving backups idle. Tailwind ensures all writes are atomic by leveraging a protocol that detects incomplete RDMA transfers. Tailwind substantially improves replication throughput and response latency compared with conventional RPC-based replication. In symmetric systems where servers both serve requests and act as replicas, Tailwind also improves normal-case throughput by freeing server CPU resources for request processing. We implemented and evaluated Tailwind on RAMCloud, a low-latency in-memory storage system. Experiments show Tailwind improves RAMCloud's normal-case request processing throughput by $1.7\times$. It also cuts down writes median and 99^{th} percentile latencies by 2x and 3x respectively [23].

5.3. Hybrid edge/cloud processing

5.3.1. Edge benchmarking

Participants: Pedro Silva, Alexandru Costan, Gabriel Antoniu.

The recent spectacular rise of the Internet of Things and the associated augmentation of the data deluge motivated the emergence of Edge computing as a means to distribute processing from centralized Clouds towards decentralized processing units close to the data sources. This led to new challenges regarding the ways to distribute processing across Cloud-based, Edge-based or hybrid Cloud/Edge-based infrastructures. In particular, a major question is: how much can one improve (or degrade) the performance of an application by performing computation closer to the data sources rather than keeping it in the Cloud?

In the paper “Investigating Edge vs. Cloud Computing Trade-offs for Stream Processing” submitted to CCGrid 2019, it is proposed a methodology to understand such performance trade-offs. Using two representative real-life stream processing applications and state-of-the-art processing engines, we perform an experimental evaluation based on the analysis of the execution of those applications in fully-Cloud computing and hybrid Cloud-Edge computing infrastructures. We derive a set of take-aways for the community, highlighting the limitations of each environment, the scenarios that could benefit from hybrid Edge-Cloud deployments, what relevant parameters impact performance and how.

5.3.2. Planner: cost-efficient execution plans for the uniform placement of stream analytics on Edge and Cloud

Participants: Laurent Prosperi, Alexandru Costan, Pedro Silva, Gabriel Antoniu.

Stream processing applications handle unbounded and continuous flows of data items which are generated from multiple geographically distributed sources. Two approaches are commonly used for processing: Cloud-based analytics and Edge analytics. The first one routes the whole data set to the Cloud, incurring significant costs and late results from the high latency networks that are traversed. The latter can give timely results but forces users to manually define which part of the computation should be executed on Edge and to interconnect it with the remaining part executed in the Cloud, leading to sub-optimal placements.

More recently, a new hybrid approach tries to combine both Cloud and Edge analytics in order to offer better performance, flexibility and monetary costs for stream processing. However, leveraging this dual approach in practice raises some significant challenges mainly due to the way in which stream processing engines organize the analytics workflow. Both Edge and Cloud engines create a dataflow graph of operators that are deployed on the distributed resources; they devise an execution plan by traversing this graph. In order to execute a request over such hybrid deployment, one needs a specific plan for the Edge engines, another one for the cloud engines and to ensure the right interconnection between them thanks to an ingestion system. Manually and empirically deploying this analytics pipeline (Edge-Ingestion-Cloud) can lead to sub-optimal computation placement with respect to the network cost (i.e., high latency, low throughput) between the Edge and the Cloud.

In this [26], we argue that a uniform approach is needed to bridge the gap between Cloud SPEs and Edge analytics frameworks in order to leverage a single, transparent execution plan for stream processing in both environments. We introduce Planner, a streaming middleware capable of finding cost-efficient cuts of execution plans between Edge and Cloud. Our goal is to find a distributed placement of operators on Edge and Cloud nodes to minimize the stream processing makespan. Real-world micro-benchmarks show that Planner reduces the network usage by 40 % and the makespan (end-to-end processing time) by 15 % compared to state-of-the-art.

5.3.3. Integrating KerA and Flink

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu.

Big Data real-time stream processing typically relies on message broker solutions that decouple data sources from applications. This translates into a three-stage pipeline: (1) event sources (e.g., smart devices, sensors, etc.) continuously generate streams of records; (2) in the *ingestion* phase, these records are acquired, partitioned and pre-processed to facilitate consumption; (3) in the *processing* phase, Big Data engines consume the stream records using a *pull-based* model. Since users are interested in obtaining results as soon as possible, there is a need to minimize the end-to-end latency of the three stage pipeline. This is a non-trivial challenge when records arrive at a fast rate (from producers and to consumers) and create the need to support a high throughput at the same time.

The weak link of the three-stage pipeline is the ingestion phase: it needs to acquire records with a high throughput from the producers, serve the consumers with a high throughput, scale to a large number of producers and consumers, and minimize the write latency of the producers and, respectively, the read latency of the consumers to facilitate low end-to-end latency. Since producers and consumers communicate with message brokers through RPCs, there is inevitably *interference* between these operations which can lead to increased processing times. Moreover, since consumers (i.e., source operators) depend on the networking infrastructure, its characteristics can limit the read throughput and/or increase the end-to-end read latency. One simple idea is to co-locate processing workers (source and other operators) with brokers managing stream partitions. We implement this approach by integrating KerA with Flink through a shared-memory approach. Experiments results demonstrate the effectiveness of our approach.

5.4. Scalable I/O, storage and in-situ visualization

5.4.1. HDF-based storage

Participants: Hadi Salimi, Gabriel Antoniu.

Extreme-scale scientific simulations that are deployed on thousands of cores usually store the resulted datasets in standard formats such as HDF5 or NetCDF. In the data storage process, two different approaches are traditionally employed: 1) file-per-process and 2) collective I/O. In the former approach, each computing core creates its own file at the end of each simulation iteration. However, this approach cannot scale up to thousands of cores because creating and updating thousands of files at the end of each iteration, leads to a poor performance. On the other hand, the latter is based on the coordination of processes to write on a single file that is also expensive in terms of performance.

The proposed approach in this research is to use Damaris for data aggregation and data storage. In the case, the computing resources are partitioned such that a subset of cores in each node or a subset of nodes of the underlying platform is dedicated to data management. The data generated by the simulation processes are transferred to these dedicated cores/nodes 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 asynchronously. Afterwards, the aggregated data can be stored in HDF5 format using out-of-the-box Damaris plug-in.

The benefits of using Damaris for storing simulation results into HDF5 is threefold: firstly, Damaris aggregates data from different processes in one process, as a result, the number of I/O writers is decreased; secondly, the write phase becomes entirely asynchronous, so the simulation processes do not have to wait for the write phase to be completed; and finally, the Damaris API is much more straightforward for simulation developers. Hence it can be easily integrated in simulation codes and easily maintained as well. The performance evaluation of the implemented prototype shows that using Damaris for storing simulation data can lead up to 297 % improvement compared to the standard file-per-process approach [32].

5.4.2. *Leveraging Damaris for in-situ visualization in support of GeoScience and CFD simulations*

Participants: Hadi Salimi, Gabriel Antoniu.

In the context of an industrial collaboration, KerData managed to sign a contract with Total around Damaris. Total is one of the industrial pioneers of HPC in France and owns the fastest supercomputer in France, named Pangea. On this machine, lots of geoscience simulations (oil exploration, oil extraction, seismic, etc.) are executed everyday and the results of these simulations are used by company's geoscientists.

This feasibility study on using Damaris on Total's geoscience simulations has been subject to an expertise contract between Total and KerData. The main goal of the contract is to show that Damaris is capable of supporting Total simulations to provide asynchronous I/O and in situ visualization. To this aim, by instrumenting two wave propagation simulation codes (prepared by Total), it was shown that Damaris can be applied to Total's wave propagation simulations in support of in situ visualization and asynchronous I/O.

The amount of changes made into the target simulations to support Damaris shows that for simple and complex simulations, the amount of changes in the simulation source code remain nearly the same. In addition, those part of the simulation code that are dedicated to dumping of the results can be totally removed, because Damaris supports this feature in a simpler and even more efficient way.

MYRIADS Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Fog Computing

Participants: Guillaume Pierre, Cédric Tedeschi, Arif Ahmed, Ali Fahs, Hamidreza Arkian, Mulugeta Tamiru, Mozhdeh Farhadi, Paulo Rodrigues de Souza Junior, Davaadorj Battulga, Genc Tato, Lorenzo Civolani, Trung Le.

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 traditional clouds: to provide computing resources in the 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, and demonstrated that major performance gains can be obtained with relatively simple modifications in the way Docker imports container images [12]. This work is currently being extended in a variety of ways: exploiting distributed storage services to share image among fog nodes, reorganizing the Docker images to allow them to be booted before the image has been fully downloaded, exploiting checkpoint/restart mechanisms to efficiently deploy application that have a long startup time. We expect a few publications on these topics in the coming year.

There does not yet exist any reference platform for fog computing platforms. We therefore investigate how Kubernetes could be adapted to support the specific needs of fog computing platforms. In particular we focused on the problem of redirecting end-user traffic to a nearby instance of the application. When different users impose various load on the system, any traffic routing system must necessarily implement a tradeoff between proximity and fair load-balancing between the application instances. We demonstrated how such customizeable traffic routing policies can be integrated in Kubernetes to help transform it in a suitable platform for fog computing. A paper on this topic is currently under review.

We investigated in collaboration with Etienne Riviere from UC Louvain the feasibility and possible benefits brought about by the *edgification* of a legacy micro-service-based application [31]. In other words, we devised a method to classify services composing the application as *edgifiable* or not, based on several criteria. We applied this method to the particular case of the ShareLatex application which enables the collaborative edition of LaTeX documents.

Thanks to the FogGuru MSCA H2020 project, five new PhD students have also started this year on various topics related to fog computing. We expect the first scientific results to appear in 2019.

7.1.2. Community Clouds

Participants: Jean-Louis Pazat, Bruno Stevant.

It is now feasible for consumers to buy inexpensive devices that can be installed at home and accessed remotely thanks to an Internet connection. Such a simple “self-hosting” paradigm can be an alternative to traditional cloud providers, especially for privacy-conscious users. We discuss how a community of users can pool their devices in order to host microservices-based applications, where each microservice is deployed on a different device. The performance of such an application depends heavily on the computing and network resources that are available and on the placement of each microservice. Finding the placement that minimizes the application response time is an NP-hard problem. We show that, thanks to well known optimization techniques (Particle Swarm Optimization), it is possible to quickly find a service placement resulting in a response time close to the optimal one. Thanks to an emulation platform, we evaluate the robustness of this solution to changes in the Quality of Service under conditions typical of a residential access network [30].

7.1.3. Stream Processing

Participants: Cédric Tedeschi, Mehdi Belkhiria.

We investigated a decentralized scaling mechanism for stream processing applications where the different operators composing the processing topology are able to take their own scaling decisions independently, based on local information. We built a simulation tool to validate the ability of our algorithm to react to load variation. We plan to submit a paper on this topic by the end of 2018.

7.1.4. QoS-aware and Energy-efficient Resource Management for Function as a Service

Participants: Yasmina Bouizem, Christine Morin, Nikos Parlavantzas.

Recent years have seen the widespread adoption of serverless computing, and in particular, Function-as-a-Service (FaaS) systems. These systems enable users to execute arbitrary functions without managing underlying servers. However, existing FaaS frameworks provide no quality of service guarantees to FaaS users in terms of performance and availability. Moreover, they provide no support for FaaS providers to reduce energy consumption. The goal of this work is to develop an automated resource management solution for FaaS plaforms that takes into account performance, availability, and energy efficiency in a coordinated manner. This work is performed in the context of the thesis of Yasmina Bouizem. In 2018, we analysed the challenges of designing FaaS platforms and performed a detailed evaluation of three open-source FaaS frameworks, all based on Kubernetes, with respect to performance, fault-tolerance, energy consumption, and extensibility [13].

7.1.5. Cost-effective Reconfiguration for Multi-cloud Applications

Participants: Christine Morin, Nikos Parlavantzas, Linh Manh Pham.

Modern applications are increasingly being deployed on resources delivered by Infrastructure-as-a-Service (IaaS) cloud providers. A major challenge for application owners is continually managing the application deployment in order to satisfy the performance requirements of application users, while reducing the charges paid to IaaS providers. This work developed an approach for adaptive application deployment that explicitly considers adaptation costs and benefits in making deployment decisions. The approach relies on predicting the duration of reconfiguration actions as well as workload changes. The work builds on the Adapter system, developed by Myriads in the context of the PaaSage European project (2012-2016). We have evaluated the approach using experiments in a real cloud testbed, demonstrating its ability to perform multi-cloud adaptation while optimizing the application owner profit under diverse circumstances [25].

7.1.6. Adaptive Resource Management for High-performance, Real-time Embedded Systems

Participants: Baptiste Goupille-Lescar, Christine Morin, Nikos Parlavantzas.

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 2018, we proposed an online scheduling approach for executing real-time applications on heavily-constrained embedded architectures. The approach enables dynamically allocating resources to fulfill requests coming from several sensors, making the most of the computing platform, while providing guaranties on quality of service levels. The approach was tested in an industrial use case concerning the operation of a multi-function surface active electronically scanned array (AESA) radar. We showed that the approach allows us to obtain lower execution latencies than current mapping solutions while maintaining high predictability and allowing gradual performance degradation in overload scenarios [22].

7.2. Greening Clouds

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.

Finally, a research result dating from 2015 was finally published after a long review and publication process [4]: to help the energy-aware co-design of IaaS and PaaS platforms, we conducted an extensive experimental evaluation of the effect of a range of Cloud infrastructure operations (start, stop, migrate VMs) on their computing throughput and energy consumption, and derived a model to help drive cloud reconfiguration operations according to performance/energy requirements.

7.2.2. End-to-end Energy Models for Internet of Things

Participant: 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 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). This work has been published in [10].

7.2.3. Exploiting Renewable Energy in Distributed Clouds

Participants: Benjamin Camus, Anne-Cécile Orgerie.

The growing appetite of Internet services for Cloud resources leads to a consequent increase in data center (DC) facilities worldwide. This increase directly impacts the electricity bill of Cloud providers. Indeed, electricity is currently the largest part of the operation cost of a DC. Resource over-provisioning, energy non-proportional behavior of today's servers, and inefficient cooling systems have been identified as major contributors to the high energy consumption in DCs.

In a distributed Cloud environment, on-site renewable energy production and geographical energy-aware load balancing of virtual machines allocation can be associated to lower the brown (i.e. not renewable) energy consumption of DCs. Yet, combining these two approaches remains challenging in current distributed Clouds. Indeed, the variable and/or intermittent behavior of most renewable sources – like solar power for instance – is not correlated with the Cloud energy consumption, that depends on physical infrastructure characteristics and fluctuating unpredictable workloads.

We proposed NEMESIS: a Network-aware Energy-efficient Management framework for distributEd cloudS Infrastructures with on-Site photovoltaic production. The originality of NEMESIS lies in its combination of a greedy VM allocation algorithm, a network-aware live-migration algorithm, a dichotomous consolidation algorithm and a stochastic model of the renewable energy supply in order to optimize both green and brown energy consumption of a distributed cloud infrastructure with on-site renewable production. Our solution employs a centralized resource manager to schedule VM migrations in a network-aware and energy-efficient way, and consolidation techniques distributed in each data center to optimize the Cloud's overall energy consumption. This work has been published in [15] and [38].

7.2.4. Smart Grids

Participants: Anne Blavette, Benjamin Camus, Anne-Cécile Orgerie, Martin Quinson.

Smart grids allow to efficiently perform demand-side management in electrical grids in order to increase the integration of fluctuating and/or intermittent renewable energy sources in the energy mix. In this work, we consider a distributed computing cloud partially powered by photovoltaic panels as a self-consumer that can also benefit from geographical flexibility: the computing load can be moved from one data center to another one benefiting from better solar irradiance conditions. The various data centers composing the cloud can then cooperate to better synchronise their consumption with their photovoltaic production.

We aim at optimizing the self-power consumption of a distributed Cloud infrastructure with on-site photovoltaic electricity generation. We propose to rely on the flexibility brought by Smart Grids to exchange renewable energy between data centers and thus, to further increase the overall Cloud's self-consumption of the locally-produced renewable energy. Our solution is named SCORPIUS: Self-Consumption Optimization of Renewable energy Production In distribUted cloudS. It optimizes the Cloud's self-consumption by trading-off between VM migration and renewable energy exchange. This optimization is based on an original Smart Grid model to exchange renewable energy between distant sites. This work has been published in the distributed computing community [14] and in the electrical engineering community [37].

7.2.5. 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 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. This work was explored in the context of David Guyon's PhD thesis (defended on December 7, 2018). For 2018, it resulted in one publication in the International Journal of Grid and Utility Computing [8] and two publications in conferences: IC2E [23] and SBAC-PAD [24].

7.3. Securing Clouds

7.3.1. Security Monitoring in Clouds

Participants: Christine Morin, Louis Rilling, Amir Teshome Wonjiga, Clément Elbaz.

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.

After having proposed a verification method for security monitoring SLOs [33], we have worked on defining security monitoring SLOs that are at the same time relevant for the tenant, achievable for the provider, and verifiable. Indeed the experiments done when studying verification showed the costs of verying the configuration of an NIDS, in time and in network overhead on the tenant's virtual infrastructure. This allows us to propose trade-offs in the verification part of an SLO. In order to allow a provider to propose achievable SLOs, we also propose methods to predict metrics of evaluation for an NIDS configured according to the specific needs of a tenant. These predictions are based on measurements done on a set of basic setups of the NIDS, the basic setups covering together the variety of NIDS rules that may interest tenants. Finally we propose extensions to an existing cloud SLA language to define security monitoring SLOs. These results will be submitted for publication in beginning of 2019.

To make security monitoring SLOs adaptable to context changes like the evolution of threats and updates to the tenants' software, we first studied the economic feasibility for a provider to guarantee new threats mitigation in SLAs. Our study of 3 years on the lifecycle of public vulnerabilities from their publication to the publication of mitigations (either as intrusion detection rules or as software patches) shows that there is room for providers to propose profitable SLAs. The results of this study incite us to investigate in two directions: how to incite tenants to apply security patches on the software they manage, and how to mitigate new threats during time window in which no intrusion detection rule exist and no security patch is applied yet (if available).

Our results were published in [33], [34], [20], [21].

A demo of SAIDS, our prototype of self-adaptable network intrusion detection systems was also presented at FIC 2018, Lille, France in January 2018.

7.4. Experimenting with Clouds

7.4.1. Simulating Distributed IT Systems

Participants: Toufik Boubehziz, Benjamin Camus, Anne-Cécile Orgerie, Millian Poquet, Martin Quinson.

Our team plays a major role in the advance of the SimGrid simulator of IT systems. This framework has a major impact on the community. Cited by over 900 papers, it was used as a scientific instrument by more than 300 publications over the years.

This year, we pursued our effort to ensure that SimGrid becomes a *de facto* standard for the simulation of distributed IT platforms. We further polished the new interface to ensure that it correctly captures the concepts needed by the experimenters. To that extend, we also added several complex applications to our Continuous Integration (CI) testing framework, to ensure that we correctly cover the needs of our existing users. We also worked toward our potential users by reworking the documentation, and by proposing new pedagogical resources. Making SimGrid usable in the classroom should greatly increase its impact. A publication on this effort was recognized as Best Paper in the Workshop on Education for High-Performance Computing [17].

The work on SimGrid is fully integrated to the other research efforts of the Myriads team. This year, we added the ability to co-simulate IT systems with SimGrid and physical systems modeled with equational systems [16]. This work, developed to study the co-evolution of thermal systems or of the electic grid with the IT system, is now distributed as an official plugin of the SimGrid framework.

7.4.2. Formal Methods for IT Systems

Participants: The Anh Pham, Martin Quinson.

The SimGrid framework also provide a state of the art Model-Checker for MPI applications. This can be used to formally verify whether the application entails synchronization issues such as deadlocks or livelocks [7].

This year, we pursued our effort (in collaboration with Thierry Jéron, EPI SUMO) to improve the reduction techniques proposed to mitigate the state space explosion issue. We are leveraging event folding structures to improve the performance and accuracy of dynamic partial ordering reduction techniques. We plan to submit a publication on this work by the beginning of 2019.

7.4.3. Executing Epidemic Simulation Applications in the Cloud

Participants: Christine Morin, Nikos Parlavantzas, Manh Linh Pham.

In the context of the DiFFuSE ADT and in collaboration with INRA researchers, we transformed a legacy application for simulating the spread of Mycobacterium avium subsp. paratuberculosis (MAP) to a cloud-enabled application based on the DiFFuSE framework (Distributed framework for cloud-based epidemic simulations). This is the second application to which the DiFFuSE framework is applied. The first application was a simulator of the spread of the bovine viral diarrhea virus, developed within the MIHMES project (2012-2017). Using both the MAP and BVDV applications, we performed extensive experiments showing the advantages of the DiFFuSE framework. Specifically, we showed that DiFFuSE enhances application performance and allows exploring different cost-performance trade-offs while supporting automatic failure handling and elastic resource acquisition from multiple clouds. These results are described in a journal article under submission. In 2018, we also released the fist major version of the DiFFuSE software (v1.0) under the CeCILL-B licence.

7.4.4. Implicit locality awareness of Remote Procedure Calls evaluation

Participants: Javier Rojas Balderrama, Matthieu Simonin.

Cloud computing depends on communication mechanisms implying location transparency. Transparency is tied to the cost of ensuring scalability and an acceptable request responses associated to the locality. Current implementations, as in the case of OpenStack, mostly follow a centralized paradigm but they lack the required service agility that can be obtained in decentralized approaches. In an edge scenario, the communicating entities of an application can be dispersed. In this context, we focus our study on the inter-process communication of OpenStack when its agents are geo-distributed regarding two key metrics: scalability and locality. Scalability refers to the ability of the communication middleware to handle a massive number of clients while consuming a reasonnable amount of resources. Locality refers to the ability of the communication middleware to serve requests as localy as possible while mitigating long-haul data transfers.

Results show that scalability and locality are very limited when considering the traditionnal broker-based approaches [28]. Novel solution such as router-based communication middleware offers better scalability and a good level of implicit locality. This work is an initial step towards building locality-aware geo-distributed systems.

7.4.5. *Tools for the experimentation*

Participant: Matthieu Simonin.

In collaboration with the STACK team and in the context of the Discovery IPL, novel experimentation tools have been developed. In this context experimenting with large software stacks (OpenStack, Kubernetes) was required. These stacks are often tedious to handle. However, practitioners need a right abstraction level to express the moving nature of experimental targets. This includes being able to easily change the experimental conditions (e.g underlying hardware and network) but also the software configuration of the targeted system (e.g service placement, fined-grained configuration tuning) and the scale of the experiment (e.g migrate the experiment from one small testbed to another bigger testbed).

In this spirit we discuss in [19] a possible solution to the above desiderata. We illustrate its use in a real world use case study which has been completed in [28]. We show that an experimenter can express their experimental workflow and execute it in a safe manner (side effects are controlled) which increases the repeatability of the experiments.

STACK Team

7. New Results

7.1. Resource Management

Participants: Mohamed Abderrahim, Ronan-Alexandre Cherrueau, Bastien Confais, Jad Darrous, Shadi Ibrahim, Adrien Lebre, Matthieu Simonin, Emile Cadorel, Hélène Coullon, Jean-Marc Menaud.

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

7.1.1. Geo-distributed Infrastructures

In [15], we provide reflections regarding how fog/edg infrastructures can be operated. While it is clear that edge infrastructures are required for emerging use-cases related to IoT, VR or NFV, there is currently no resource management system able to deliver all features for the edge that made cloud computing successful (*e.g.*, an OpenStack for the edge). Since building a system from scratch is seen by many as impractical, our community should investigate different appraoches. This study, which has been achieved with Ericsson colleagues, provides a list of the features required to operate and use edge computing resources, and investigate how an existing IaaS manager (i.e., OpenStack) satisfies these requirements. Finally, we identify from this study two approaches to design an edge infrastructure manager that fulfils our requirements, and discuss their pros and cons.

In [18], we propose a new novel VMI management system for distributed cloud infrastructures. Most large cloud providers, like Amazon and Microsoft, replicate their Virtual Machine Images (VMIs) on multiple geographically distributed data centers to offer fast service provisioning. Provisioning a service may require to transfer a VMI over the wide-area network (WAN) and therefore is dictated by the distribution of VMIs and the network bandwidth in-between sites. Nevertheless, existing methods to facilitate VMI management (ie, retrieving VMIs) overlook network heterogeneity in geo-distributed clouds. To deal with such a limitation, we design, implement and evaluate Nitro, a novel VMI management system that helps to minimize the transfer time of VMIs over a heterogeneous WAN. To achieve this goal, Nitro incorporates two complementary features. First, it makes use of deduplication to reduce the amount of data which will be transferred due to the high similarities within an image and in-between images. Second, Nitro is equipped with a network-aware data transfer strategy to effectively exploit links with high bandwidth when acquiring data and thus expedites the provisioning time. Experimental results show that our network-aware data transfer strategy offers the optimal solution when acquiring VMIs while introducing minimal overhead. Moreover, Nitro outperforms state-of-the-art VMI storage systems (eg, OpenStack Swift) by up to 77%.

In [22] we perform a performance evaluation of two communication bus mechanisms available in the openstack eco-system. Cloud computing depends on communication mechanisms implying location transparency. Transparency is tied to the cost of ensuring scalability and an acceptable request responses associated to the locality. Current implementations, as in the case of OpenStack, mostly follow a centralized paradigm but they lack the required service agility that can be obtained in decentralized approaches. In an edge scenario, the communicating entities of an application can be dispersed. In this context, we perform a study on the inter-process communication of OpenStack when its agents are geo-distributed. More precisely, we are interested in the different Remote Procedure Calls (OARPCs) implementations of OpenStack and their behaviours with regards to three classical communication patterns: anycast, unicast and multicast. We discuss how the communication middleware can align with the geo-distribution of the RPC agents regarding two key factors: scalability and locality. We reached up to ten thousands communicating agents, and results show that a router-based deployment offers a better trade-off between locality and load-balancing. Broker-based suffers from its centralized model which impact the achieved locality and scalability.

In [5], we give a complete overview of VMPlaceS, a dedicated framework we have been implementing since 2015 in order to evaluate and compare VM placement algorithms. Most current infrastructures for cloud computing leverage static and greedy policies for the placement of virtual machines. Such policies impede the optimal allocation of resources from the infrastructure provider viewpoint. Over the last decade, more dynamic and often more efficient policies based, *e.g.*, on consolidation and load balancing techniques, have been developed. Due to the underlying complexity of cloud infrastructures, these policies are evaluated either using limited scale testbeds/in-vivo experiments or ad-hoc simulators. These validation methodologies are unsatisfactory for two important reasons: they (i) do not model precisely enough real production platforms (size, workload variations, failure, etc.) and (ii) do not enable the fair comparison of different approaches. More generally, new placement algorithms are thus continuously being proposed without actually identifying their benefits with respect to the state of the art. In this article, we present and discuss most of the features provided by VMPlaceS, a dedicated simulation framework that enables researchers (i) to study and compare VM placement algorithms from the infrastructure perspective, (ii) to detect possible limitations at large scale and (iii) to easily investigate different design choices. Built on top of the SimGrid simulation platform, VMPlaceS provides programming support to ease the implementation of placement algorithms and runtime support dedicated to load injection and execution trace analysis. To illustrate the relevance of VMPlaceS, we first discuss a few experiments that enabled us to study in details three well known VM placement strategies. Diving into details, we also identify several modifications that can significantly increase their performance in terms of reactivity. Second, we complete this overall presentation of VMPlaceS by focusing on the energy efficiency of the well-know FFD strategy. We believe that VMPlaceS will allow researchers to validate the benefits of new placement algorithms, thus accelerating placement research and favouring the transfer of results to IaaS production platforms.

In [27], we present different heuristics that address the placement challenge in Fog/Edge infrastructures. As Fog Computing brings processing and storage resources to the edge of the network, there is an increasing need of automated placement (i.e., host selection) to deploy distributed applications. Such a placement must conform to applications' resource requirements in a heterogeneous Fog infrastructure, and deal with the complexity brought by Internet of Things (IoT) applications tied to sensors and actuators. In this study, we present and evaluate four heuristics to address the problem of placing distributed IoT applications in the fog. By combining proposed heuristics, our approach is able to deal with large scale problems, and to efficiently make placement decisions fitting the objective: minimizing placed applications' average response time. The proposed approach has been validated through comparative simulation of different heuristic combinations with varying sizes of infrastructures and applications.

In [35], we introduce the premises of monitoring function chaining conepts with the ultimate goal of delievering an holistic monitoring system for Fog/Edge infrastuctures. By relying on small sized and massively distributed infrastructures, the Edge computing paradigm aims at supporting the low latency and high bandwidth requirements of the next generation services that will leverage IoT devices (e.g., video cameras, sensors). To favor the advent of this paradigm, management services, similar to the ones that made the success of Cloud computing platforms, should be proposed. However, they should be designed in order to cope with the limited capabilities of the resources that are located at the edge. In that sense, they should mitigate as much as possible their footprint. Among the different management services that need to be revisited, we investigate in this study the monitoring one. Monitoring functions tend to become compute-, storage- and network-intensive, in particular because they will be used by a large part of applications that rely on real-time data. To reduce as much as possible the footprint of the whole monitoring service, we propose to mutualize identical processing functions among different tenants while ensuring their quality-of-service (QoS) expectations.We formalize our approach as a constraint satisfaction problem and show through micro-benchmarks its relevance to mitigate compute and network footprints.

In [17], we discuss the limitations of meta-data management in Fog/Edge infrastructures. A few storage systems have been proposed to store data in those infrastructures. Most of them are relying on a Distributed Hash Table (DHT) to store the location of objects which is not efficient because the node storing the location of the data may be placed far away from the object replicas. In this paper, we propose to replace the DHT by a tree-based approach mapping the physical topology. Servers look for the location of an object by requesting successively their ancestors in the tree. Location records are also relocated close to the object replicas not only to limit the network traffic when requesting an object, but also to avoid an overload of the root node. We also propose to modify the Dijkstra's algorithm to compute the tree used. Finally, we evaluate our approach using the object store InterPlanetary FileSystem (IPFS) on Grid'5000 using both a micro experiment with a simple network topology and a macro experiment using the topology of the French National Research and Education Network (RENATER). We show that the time to locate an object in our approach is less than 15 ms on average which is around 20% better than using a DHT.

7.1.2. Cloud and HPC convergence

Geo-distribution of Cloud Infrastructures is not the only current trend of utility computing. Another important challenge is to favor the convergence of Cloud and HPC infrastructures, in other words on-demand HPC. Among challenges of this convergence is, for example, how to exploit HPC systems to execute data-intensive workflows effectively, as well as how to schedule tasks and jobs in Cloud, HPC, or hybrid HPC/Cloud infrastructures to meet data volatility and the ever-growing heterogeneity in the computation demands of workflows.

With the growing needs of users and size of data, commodity-based infrastructure will strain under the heavy weight of Big Data. On the other hand, HPC systems offer a rich set of opportunities for Big Data processing. As first steps toward Big Data processing on HPC systems, several research efforts have been devoted to understanding the performance of Big Data applications on these systems. Yet the HPC specific performance considerations have not been fully investigated. In [28], we conduct an experimental campaign to provide a clearer understanding of the performance of Spark, the de facto in-memory data processing framework, on HPC systems. We ran Spark using representative Big Data workloads on Grid'5000 testbed to evaluate how the latency, contention and file system's configuration can influence the application performance. We discuss the implications of our findings and draw attention to new ways (e.g., burst buffers) to improve the performance of Spark on HPC systems.

Motivated by the our work [28] , we extend Eley [107], a burst buffer solution that aims to accelerate the performance of Big Data applications, to be interference-aware. Specifically, while data prefetching reduce the response time of Big data applications as data inputs will be stored on a low-latency device close to computing nodes, it may come at a high cost for the HPC applications: the continuous interaction with the parallel file system (i.e., I/O read requests) may introduce a huge interference at the parallel file system level and thus end up with a degraded and unpredictable performance for HPC applications. In [7], we introduce interference and performance models for both HPC and Big Data applications in order to identify the performance gain and the interference cost of the prefetching technique of Eley; and demonstrate how Eley chooses the best action to optimize the prefetching while guaranteeing the pre-defined QoS requirement of HPC applications. For example, with 5% QoS requirement of the HPC application, Eley reduces the execution time of Big Data applications by up to 30% compared to the Naive burst buffer solution (NaiveBB) while guaranteeing the QoS requirement. On the other hand, the NaiveBB violates the QoS requirement by up to 58%.

Besides Clouds, Data Stream Processing (DSP) applications are widely deployed in HPC systems, especially the ones which require timely responses. DSP applications are often modelled as a directed acyclic graph: operators with data streams among them. Inter-operator communications can have a significant impact on the latency of DSP applications, accounting for 86% of the total latency. Despite their impact, there has been relatively little work on optimizing inter-operator communications, focusing on reducing inter-node traffic but not considering inter-process communication (IPC) inside a node, which often generates high latency due to the multiple memory-copy operations. In [26], we introduce a new DSP system designed specifically to address the high latency caused by inter-operator communications, called TurboStream. To achieve this goal, we introduce (1) an improved IPC framework with OSRBuffer, a DSP-oriented buffer, to reduce memory-copy operations and waiting time of each single message when transmitting messages between the operators inside one node, and (2) a coarse-grained scheduler that consolidates operator instances and assigns them to nodes to diminish the inter-node IPC traffic. Using a prototype implementation, we show that our improved IPC framework reduces the end-to-end latency of intra-node IPC by 45.64% to 99.30%. Moreover, TurboStream reduces the latency of DSP by 83.23% compared to JStorm.

Current data stream or operation stream paradigms cannot handle data burst efficiently, which probably results in noticeable performance degradation. In [25], we introduce a dual-paradigm stream processing, called DO (Data and Operation) that can adapt to stream data volatility. It enables data to be processed in micro-batches (ie, operation stream) when data burst occurs to achieve high throughput, while data is processed record by record (ie, data stream) in the remaining time to sustain low latency. DO embraces a method to detect data bursts, identify the main operations affected by the data burst and switch paradigms accordingly. Our insight behind DO's design is that the trade-off between latency and throughput of stream processing frameworks can be dynamically achieved according to data communication among operations in a fine-grained manner (ie, operation level) instead of framework level. We implement a prototype stream processing framework that adopts DO. Our experimental results show that our framework with DO can achieve 5x speedup over operation stream under low data stream sizes, and outperforms data stream on throughput by 2.1 x to 3.2 x under data burst.

In the context of the Hydda project, where hybrid HPC/Cloud infrastructures are studied, heterogeneous dataflows, composed of coarse-grain tasks interconnected through data dependencies, are scheduled. Indeed, in heterogeneous dataflows, genomics dataflows for instance, some tasks may need HPC infrastructures (*e.g.*, simulation) while other are suited for Cloud infrastructures (*e.g.*, Big Data). Different quality of services are also expected from one task to the other. In [31] the scheduling of heterogeneous scientific dataflows is studied while minimizing the Cloud provider operational costs, by introducing a deadline-aware algorithm. Scheduling in a Cloud environment is a difficult optimization problem. Usually, works around the scheduling of scientific dataflows focus on public Clouds where the management of the infrastructure is an unknown black box. Thus, many works offer scheduling algorithms built to choose the best set of virtual machines through time such that the cost of the enduser is minimized. This paper presents a new algorithm based on HEFT that aims at minimizing the number of machines used by the Cloud provider, by taking deadlines into account.

7.2. Programming Support

Participants: Zakaria Al-Shara, Frederico Alvares, Maverick Chardet, Hélène Coullon, Thomas Ledoux, Jacques Noyè, Dimitri Pertin.

Our contributions regarding the programming support are divided in two topics. First, we have contributed to automated deployment and reconfiguration with three publications. Second, we have contributed to autonomic computing and self-management in the Cloud with two publications. While these topics are strongly related (*i.e.*, a reconfiguration system is an autonomic controller), we have decided to distinguish two different levels of contributions, one being based on deployment and reconfiguration execution, or software commissioning (low level system commands), while the other uses model-driven software engineering techniques to build common self-management models for the Cloud (high level abstractions).

7.2.1. Deployment and reconfiguration in the Cloud

Distributed software architecture is composed of multiple interacting modules, or components. Deploying such software consists in installing them on a given infrastructure and leading them to a functional state. However, since each module has its own life cycle and might have various dependencies with other modules, deploying such software is a very tedious task, particularly on massively distributed and heterogeneous infrastructures. To address this problem, many solutions have been designed to automate the deployment process. In [14], we introduce Madeus, a component-based deployment model for complex distributed software. Madeus accurately describes the life cycle of each component by a Petri net structure, and is able to finely express the dependencies between components. The overall dependency graph it produces is then used to reduce deployment time by parallelizing deployment actions. While this increases the precision and performance of the model, it also increases its complexity. For this reason, the operational semantics needs to be clearly defined to prove results such as the termination of a deployment. In this paper, we formally describe the operational semantics of Madeus, and show how it can be used in a use-case: the deployment of OpenSatck, a real and large distributed software.

Distributed software and infrastructures also become more and more dynamic. Therefore, there is a need for models assisting their management, including their reconfiguration. We focus on three properties for reconfigurations. First, we think that the efficiency of a reconfiguration is of first importance as a running service should not be interrupted for a long period of time (downtime minimization). Second, we think that it is important to offer generic reconfiguration models to help developers building complex reconfigurations. Such models offer safety properties and a clear expressivity to guide the developer. Third, multiple actors are involved in reconfigurations. On one side, developers of components are responsible for describing components life cycles, while on the other side, different developers or IT adminitstrators could be responsible for the reconfiguration design of a complete distributed software composed of multiple connected components. To be able to simplify the reconfiguration design, it is important to offer the good abstraction level to each actor by guaranteeing a separation of concerns. Existing reconfiguration models are either specific to a subset of reconfigurations or are unable to provide both good performance and high separation of concerns between the actors interacting with them. In [32], we present an extension that could be applied both to Aeolus (an existing reconfiguration model) and Madeus (our deployment model). This extension introduces reconfiguration to Madeus, and enhances the separation of concerns compared to Aeolus. To this purpose, we introduce the behavior concept such that more ellaborated life-cycles can be handled by Madeus. The obtained life-cycle defined by the component developer is complex and not adapted to the reconfiguration designer. Thus, we also introduce a minimal view of each life-cycle, namely behavioral interfaces, such that the reconfiguration is still possible but hides intricate details of each component life-cycle.

In [13] we present our complete plans to extend Madeus to support reconfiguration and to provide a good separation of concerns.

7.2.2. Autonomic computing and self-management

A Cloud needs autonomic controllers to be handled efficiently. Such controllers mostly follow a loop with four steps: monitor the system or the infrastructure, analyze the situation according to the monitoring and a set of models, plan and execute actions in consequences. In the Cloud management, multiple autonomic controllers have to be designed at each level of service (*e.g.*, IaaS, PaaS, SaaS etc.). Moreover, each autonomic controller is connected to the others. In the context of massively geo-distributed infratstructures such as Fog computing, autonomic controllers will also be decentralized, thus increasing the need for generic models of autonomic controllers and their coordination.

In the CoMe4ACloud project [3], [12], we propose a generic model-based architecture for autonomic management of Cloud systems. We derive a generic unique Autonomic Manager (AM) capable of managing any Cloud service, regardless of its XaaS layer. This AM is based on a constraint solver which aims at finding the optimal configuration for the modeled XaaS, i.e. the best balance between costs and revenues while meeting the constraints established by the SLA between the producer and the consumer of the Cloud service. In [12], we introduce the designed model-based architecture, and notably its core generic XaaS modeling language. We present as well the interoperability with a Cloud standard (TOSCA). In [3], we evaluate our approach in two different ways. Firstly, we analyze qualitatively the impact of the AM behavior on the system configuration when a given series of events occurs. We show that the AM takes decisions in less than 10 s for several hundred nodes simulating virtual/physical machines. Secondly, we demonstrate the feasibility of the integration with real Cloud systems, such as OpenStack, while still remaining generic.

7.3. Energy-aware computing

Participants: Jean-Marc Menaud, Shadi Ibrahim, Thomas Ledoux, Emile Cadorel, Yewan Wang, Jonathan Pastor.

Energy consumption is one of the major challenges of modern datacenters and supercomputers. Our works in Energy-aware computing can be categorized into two subdomains: Software level (SaaS, PaaS) and Infrastructure level (IaaS).

At Software level, we worked on the general Cloud applications architecture and HPC applications.

In particular, in his habilitation thesis [2], Thomas Ledoux shows that dynamic reconfiguration in Cloud computing can provide an answer to an important societal challenge, namely digital and energetic transitions. Unlike current work providing solutions in the lower layers of the Cloud to improve the energy efficiency of data centers, Thomas Ledoux advocates a software eco-elasticity approach on the high layers of the Cloud. Inspired by both the concept of frugal innovation (Jugaad) and the mechanism of energy brownout, he proposes a number of original artifacts – such as Cloud SLA, eco-elasticity in the SaaS layers, virtualization of energy or green energy-aware SaaS applications, etc. – to reduce the carbon footprint of Cloud architectures.

However, by applying Green Programming techniques, developers have to iteratively implement and test new versions of their software, thus evaluating the impact of each code version on their energy, power and performance objectives. This approach is manual and can be long, challenging and complicated, especially for High Performance Computing applications. In [21], we formally introduces the definition of the Code Version Variability (CVV) leverage and present a first approach to automate Green Programming (i.e., CVV usage) by studying the specific use-case of an HPC stencil-based numerical code, used in production. This approach is based on the automatic generation of code versions thanks to a Domain Specific Language (DSL), and on the automatic choice of code version through a set of actors. Moreover, a real case study is introduced and evaluated though a set of benchmarks to show that several trade-offs are introduced by CVV. Finally, different kinds of production scenarios are evaluated through simulation to illustrate possible benefits of applying various actors on top of the CVV automation. While this work takes HPC applications as a use-case the presented automated green programming technique could be applied to any kind of production application onto any kind of infrastructures.

In general, many Big Data processing applications nowadays run on large-scale multi-tenant clusters. Due to hardware heterogeneity and resource contentions, straggler problem has become the norm rather than the exception in such clusters. To handle the straggler problem, speculative execution has emerged as one of the most widely used straggler mitigation techniques. Although a number of speculative execution mechanisms have been proposed, as we have observed from real-world traces, the questions of “when” and “where” to launch speculative copies have not been fully discussed and hence cause inefficiencies on the performance and energy of Big Data applications. In [29], we propose a performance model and an energy consumption model to reveal the performance and energy variations with different speculative execution solutions. We further propose a window-based dynamic resource reservation and a heterogeneity-aware copy allocation technique to answer the “when” and “where” questions for speculative executions. Evaluations using real-world traces show that our proposed technique can improve the performance of Big Data applications by up to 30% and reduce the overall energy consumption by up to 34%.

At infrastructure level, we worked on power and thermal management from server to datacenter. In fact, with the advent of Cloud Computing, the size of datacenters is ever increasing and the management of servers and their power consumption and heat production have become challenges. The management of the heat produced by servers has been experimentally less explored than the management of their power consumption. It can be partly explained by the lack of a public testbed that provides reliable access to both thermal and power metrics of server rooms. In [34], [20], [19] we had describe SeDuCe, a testbed that targets research on power and thermal management of servers, by providing public access to precise data about the power consumption and the thermal dissipation of 48 servers integrated in Grid’5000 as the new ecotype cluster. We presented the chosen software and hardware architecture for the SeDuCe testbed. Future work will focus on two areas: adding renewable energy capabilities to the SeDuCe testbed, and improving the precision of temperature sensors.

If SeDuCe testbed is focused on the management of the power consumption and heat produced by servers at room level, we realized in [30], [24], [23] , studies on power consumption (and heat impact) of physical servers. First, we characterized some potential factors on the power variation of the servers, such as: original fabrication, position in the rack, voltage variation and temperature of components on motherboard. The results show that certain factors, such as original fabrication, ambient temperature and CPU temperature, have noticeable effects on the power consumption of servers. The experimental results emphasize the importance of adding these external factors into the metric, so as to build an energy predictive model adaptable in real situations.

7.4. Security and Privacy

Participants: Mario Südholt, Mohammad Mahdi Bazm, Fatima-Zahra Boujdad, Jean-Marc Menaud.

This year the team has provided two major contributions on security and privacy challenges in distributed systems. First, we have developed our models and techniques for the detection and mitigation of side-channel attacks. Second, we have provided a first model and implementation techniques for secure and privacy-preserving distributed biomedical analyses, notably genomic ones.

7.4.1. Side-channel attacks and trusted Fog/Edge infrastructures

In [4], we investigate Cloud computing infrastructures, which are based on the sharing of hardware resources among different clients. The infrastructures leverage virtualization to share physical resources among several self-contained execution environments like virtual machines and Linux containers. Isolation is a core security challenge for such a paradigm. It may be threatened through side-channels, created due to the sharing of physical resources like caches of the processor or by mechanisms implemented in the virtualization layer. Side-channel attacks (SCAs) exploit and use such leaky channels to obtain sensitive data like kernel information. We clarify the nature of this threat for cloud infrastructures. Current SCAs are done locally and exploit isolation challenges of virtualized environments to retrieve sensitive information. We also introduce the concept of distributed side-channel attack (DSCA). We explore how such attacks can threaten isolation of any virtualized environments. Finally, we study a set of different applicable countermeasures for attack mitigation in cloud infrastructures.

In [9], we investigate Fog and Edge computing for the provision of large pools of resources at the edge of the network that may be used for distributed computing. Fog infrastructure heterogeneity also results in complex configuration of distributed applications on computing nodes. Linux containers are a mainstream technique allowing to run packaged applications and micro services. However, running applications on remote hosts owned by third parties is challenging because of untrusted operating systems and hardware maintained by third parties. To meet such challenges, we may leverage trusted execution mechanisms. In this work, we propose a model for distributed computing on Fog infrastructures using Linux containers secured by Intel's Software Guard Extensions (SGX) technology. We implement our model on a Docker and OpenSGX platform. The result is a secure and flexible approach for distributed computing on Fog infrastructures.

In [10], we contribute to the research on cache-based side-channel attacks and show the security impact of these attacks on cloud computing. The detection of cache-based side-channel attacks has received more attention in IaaS cloud infrastructures because of improvements in the attack techniques. However, detection of such attacks requires high resolution information, and it is also a challenging task because of the fine-granularity of the attacks. In this paper, we present an approach to detect cross-VM cache-based side-channel attacks through using hardware fine-grained information provided by Intel Cache Monitoring Technology (CMT) and Hardware Performance Counters (HPCs) following the Gaussian anomaly detection method. The approach shows a high detection rate with a 2% performance overhead on the computing platform.

7.4.2. Secure and privacy-aware biomedical analyses

In [11], we study the need for the sharing of genetic data, for instance, in genome-wide association studies, which is incessantly growing. In parallel, serious privacy concerns rise from a multi-party access to genetic information. Several techniques , such as encryption, have been proposed as solutions for the privacy-preserving sharing of genomes. However, existing programming means do not support guarantees for privacy properties and the performance optimization of genetic applications involving shared data. We propose two contributions in this context. First, we present new cloud-based architectures for cloud-based genetic applications that are motivated by the needs of geneticians. Second, we propose a model and implementation for the composition of watermarking with encryption, fragmentation, and client-side computations for the secure and privacy-preserving sharing of genetic data in the cloud.

7.5. Experiment-Driven Resarch

Participants: Adrien Lebre, Bastien Confais, Ronan-Alexandre Cherrueau, Matthieu Simonin, Thuy-Linh Nguyen.

Because STACK members have to perform a significant number of evaluations of complex software stack at large scale, the team contributes to the recent area of software-defined experiments and reproducible research.

In [36], [16] we propose a new approach to ensure reproducibility and repeatability of scientific experiments. Similar to the LAMP stack that considerably eased the web developers life, we advocate the need of an analogous software stack to help the experimenters making reproducible research. In 2018, we propose the EnosStack, an open source software stack especially designed for reproducible scientific experiments. EnosStack enables to easily describe experimental workflows meant to be re-used, while abstracting the underlying infrastructure running them. Being able to switch experiments from a local to a real testbed deployment greatly lower code development and validation time. In this paper, we describe the abstractions that have driven its design, before presenting a real experiment we deployed on Grid'5000 to illustrate its usefulness. We also provide all the experiment code, data and results to the community.

Similar to the previous work, we discuss in [37] a large experimental campaign that allows us to understand in details the boot duration of both virtualization techniques under various storage devices and resources contentions. While many studies have been focusing on reducing the time to manipulate Virtual Machine/Container images in order to optimize provisioning operations in a Cloud infrastructure, only a few studies have considered the time required to boot these systems. Some previous researches showed that the whole boot process can last from a few seconds to few minutes depending on co-located workloads and the number of concurrent deployed machines. The paper explains how we analyzed thoroughly the boot time of VMs, Dockers on top of bare-metal servers, and Dockers inside VMs. We discuss a methodology that enables us to perform fully-automatized and reproducible experimental campaigns on a scientific testbed. Thanks to this methodology, we conducted more than 14.400 experiments on Grid'5000 testbed for a bit more than 500 hours. The results we collected provide an important information related to the boot time behavior of these two virtualization technologies.

In [33], we presented the first experiment that has been done, as far as we know, on top of the Grid'5000 and FIT testbeds. More precisely, we discuss how we evaluated a new storage service for edge/IoT scenarios. Our proof-of-concept relies on the Interplanetary Object Store (IPFS), a Scale-Out NAS deployed on each site and a tree-based approach for the meta-data management [17]. This proposal enables (i) IoT devices to write locally on their closest site and (ii) to relocate automatically the objects on the sites they are requested, leading to low access times. The contribution of this work is a discussion of our attempt of using the two platforms simultaneously as well as the problems we encountered to interconnect them. Our ultimage goal is to give guidelines on how can researchers perform evaluations in a realistic environment : IoT devices comes from the FIT/IoT-lab and Fog nodes from Grid'5000.

WIDE Project-Team

6. New Results

6.1. Scalable Systems

6.1.1. Nobody cares if you liked Star Wars: KNN graph construction on the cheap.

Participants: Olivier Ruas, François Taïani.

K-Nearest-Neighbors (KNN) graphs play a key role in a large range of applications. A KNN graph typically connects entities characterized by a set of features so that each entity becomes linked to its k most similar counterparts according to some similarity function. As datasets grow, KNN graphs are unfortunately becoming increasingly costly to construct, and the general approach, which consists in reducing the number of comparisons between entities, seems to have reached its full potential. In work [27] we propose to overcome this limit with a simple yet powerful strategy that samples the set of features of each entity and only keeps the least popular features. We show that this strategy outperforms other more straightforward policies on a range of four representative datasets: for instance, keeping the 25 least popular items reduces computational time by up to 63%, while producing a KNN graph close to the ideal one.

This work was done in collaboration with Anne-Marie Kermarrec (Mediego/EPFL).

6.1.2. Pleiades: Distributed structural invariants at scale

Participants: Simon Bouget, David Bromberg, Adrien Luxey, François Taïani.

Modern large scale distributed systems increasingly espouse sophisticated distributed architectures characterized by complex distributed structural invariants. Unfortunately, maintaining these structural invariants at scale is time consuming and error prone, as developers must take into account asynchronous failures, loosely coordinated sub-systems and network delays. To address this problem, we propose Pleiades [31], a new framework to construct and enforce large-scale distributed structural invariants under aggressive conditions. Pleiades combines the resilience of self-organizing overlays, with the expressiveness of an assembly-based design strategy. The result is a highly survivable framework that is able to dynamically maintain arbitrary complex distributed structures under aggressive crash failures. Our evaluation shows in particular that Pleiades is able to restore the overall structure of a 25,600 node system in less than 11 asynchronous rounds after half of the nodes have crashed.

6.1.3. CASCADE: Reliable distributed session handoff for continuous interaction across devices

Participants: David Bromberg, Adrien Luxey, François Taïani.

Allowing users to navigate seamlessly between their personal devices while protecting their privacy remains today an ongoing challenge. Existing solutions rely on peer-to-peer designs, and blindly flood the network with session messages. It is particularly hard to come up with proposals that are both cost-efficient and dependable while relying on poorly connected mobile appliances. In [24] we propose Cascade, a distributed protocol to share applicative sessions among one's devices. Our proactive session handoff algorithm takes inspiration from the BitTorrent P2P file sharing protocol, but adapts it to the specific characteristics of our problem. It eschews in particular trackers, and limits the seeders of each session to the devices most likely to be used next, as computed by a decentralized aggregation protocol. A key aspect of our approach is to trade off network costs for reliability, while providing a faster session handoff than centralized solutions in the vast majority of the cases.

6.1.4. Sprinkler: A probabilistic dissemination protocol to provide fluid user interaction in multi-device ecosystems

Participants: David Bromberg, Adrien Luxey, François Taïani.

Offering fluid multi-device interactions to users while protecting their privacy largely remains an ongoing challenge. Existing approaches typically use a peer-to-peer design and flood session information over the network, resulting in costly and often unpractical solutions. In [29], we propose Sprinkler, a decentralized probabilistic dissemination protocol that uses a gossip-based learning algorithm to intelligently propagate session information to devices a user is most likely to use next. Our solution allows designers to efficiently trade off network costs for fluidity, and is for instance able to reduce network costs by up to 80% against a flooding strategy while maintaining a fluid user experience.

This work was done in collaboration with Fabio Costa, Ricardo Da Rocha and Vinicius Lima from the Universidade Federal de Goias (UFG).

6.2. Personalization and Privacy

6.2.1. GoldFinger

Participants: Olivier Ruas, François Taïani.

In work [37] we propose fingerprinting, a new technique that consists in constructing compact, fast-to-compute and privacy-preserving representation of datasets. We illustrate the effectiveness of our approach on the emblematic big data problem of K-Nearest-Neighbor (KNN) graph construction and show that fingerprinting can drastically accelerate a large range of existing KNN algorithms, while efficiently obfuscating the original data, with little to no overhead. Our extensive evaluation of the resulting approach (dubbed GoldFinger) on several realistic datasets shows that our approach delivers speed-ups up to 78.9% compared to the use of raw data while only incurring a negligible to moderate loss in terms of KNN quality. To convey the practical value of such a scheme, we apply it to item recommendation, and show that the loss in recommendation quality is negligible.

This work was done in collaboration with Rachid Guerraoui (EPFL) and Anne-Marie Kermarrec (Mediego/EPFL).

6.2.2. Collaborative filtering under a Sybil attack: Similarity metrics do matter!

Participant: Davide Frey.

Recommendation systems help users identify interesting content, but they also open new privacy threats. For this reason, in [22] we deeply analyzed the effect of a Sybil attack that tries to infer information on users from a user-based collaborative-filtering recommendation systems. We evaluated the impact of different similarity metrics used to identity users with similar tastes in the trade-off between recommendation quality and privacy. Based on our results, we proposed and evaluated a novel similarity metric that combines the best of both worlds: a high recommendation quality with a low prediction accuracy for the attacker. Our experiments, on a state-of-the-art recommendation framework and on real datasets showed that existing similarity metrics exhibit a wide range of behaviors in the presence of Sybil attacks, while our new similarity metric consistently achieves the best trade-off while outperforming state-of-the-art solutions.

This work was carried out in collaboration with Antoine Boutet from INSA Lyon, former-intern Florestan De Moor, Rachid Guerraoui and Antoine Rault from EPFL, and Anne-Marie Kermarrec from Mediego.

6.3. Network and Graph Algorithms

6.3.1. Rumor spreading and conductance

Participant: George Giakkoupis.

In [16], we study the completion time of the PUSH-PULL variant of rumor spreading, also known as randomized broadcast. We show that if a network has n nodes and conductance ϕ then, with high probability, PUSH-PULL will deliver the message to all nodes in the graph within $O(\log n/\phi)$ many communication rounds. This bound is best possible. We also give an alternative proof that the completion time of PUSH-PULL is bounded by a polynomial in $\log n/\phi$, based on graph sparsification. Although the resulting asymptotic bound is not optimal, this proof shows an interesting and, at the outset, unexpected connection between rumor spreading and graph sparsification. Finally, we show that if the degrees of the two endpoints of each edge in the network differ by at most a constant factor, then both PUSH and PULL alone attain the optimal completion time of $O(\log n/\phi)$, with high probability.

This work was done in collaboration with Flavio Chierichetti (Sapienza University of Rome), Silvio Lattanzi (Google Research), and Alessandro Panconesi (Sapienza University of Rome).

6.3.2. 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 [21], we study 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 $O(\log n/\phi)$ 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 show 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. Since our bounds coincide with the worst-case time bounds for Push rumor spreading on regular graphs until all nodes are informed, this raises the question whether CoBra walks and Push rumor spreading perform similarly in general. We answer this negatively by separating the cover time of CoBra walks and the rumor spreading time of Push by a super-polylogarithmic factor on a family of tree-like regular graphs.

This work was done in collaboration with Petra Berenbrink and Peter Kling from the University of Hamburg.

6.3.3. The quadratic shortest path problem: Complexity, approximability, and solution methods

Participant: Davide Frey.

In work [20] we considered the problem of finding a shortest path in a directed graph with a quadratic objective function (the QSPP). We show that the QSPP cannot be approximated unless $\mathsf{P} = \mathsf{NP}$. For the case of a convex objective function, we presented an n-approximation algorithm, where n is the number of nodes in the graph, and we proved APX-hardness. Furthermore, we proved that even if only adjacent arcs play a part in the quadratic objective function, the problem still cannot be approximated unless $\mathsf{P} = \mathsf{NP}$. In order to solve the general problem we first proposed a mixed integer programming formulation, and then devised an efficient exact Branch-and-Bound algorithm for the general QSPP. This algorithm computes lower bounds by considering a reformulation scheme that is solvable through a number of minimum-cost-flow problems. We carried out computational experiments solving to optimality different classes of instances with up to 1000 nodes.

This work was carried out in collaboration with Borzou Rostami from Polytechnique Montréal, Adreé Chasssein and Michael Hopf from TU Kaiserslautern, Christoph Buchheim from TU Dortmund, Federico Malucelli from Politecnico di Milano, and Marc Goerigk from Lancaster University.

6.3.4. Weighting past on the geo-aware state deployment problem

Participant: François Taïani.

The geographical barrier between mobile devices and mobile application servers (typically hosted in the Cloud) imposes an unavoidable latency and jitter that negatively impacts the performance of modern mobile systems. Fog Computing architectures can mitigate this impact if there is a middleware service able to correctly partition and deploy the state of an application at optimal locations. Geo-aware state deployment is challenging as it must consider the mobility of the devices and the dependencies arising when multiple devices concurrently manipulate the same application state. In [28], we propose a range of new object-graph-based strategies for geo-aware state deployment. In particular, our investigation focuses on understanding the role of preserving previously observed associations between state items on application performance.

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

6.3.5. Mind the gap: Autonomous detection of partitioned MANET systems using opportunistic aggregation

Participants: Simon Bouget, David Bromberg, François Taïani.

Mobile Ad-hoc Networks (MANETs) use limited-range wireless communications and are thus exposed to partitions when nodes fail or move out of reach of each other. Detecting partitions in MANETs is unfortunately a nontrivial task due to their inherently decentralized design and limited resources such as power or bandwidth. In [32], we propose a novel and fully decentralized approach to detect partitions (and other *large* membership changes) in MANETs that is both accurate and resource efficient. We monitor the current composition of a MANET using the lightweight aggregation of compact membership-encoding filters. Changes in these filters allow us to infer the likelihood of a partition with a quantifiable level of confidence. We first present an analysis of our approach, and show that it can detect close to 100% of partitions under realistic settings, while at the same time being robust to false positives due to churn or dropped packets. We perform a series of simulations that compare against alternative approaches and confirm our theoretical results, including above 90% accurate detection even under a 40% message loss rate.

This work was performed in collaboration with Etienne Rivière from UC Louvain (Belgium) and Hugues Mercier from University of Neuchâtel (Switzerland).

6.4. Theory of Distributed Systems

6.4.1. An improved bound for random binary search trees with concurrent insertions

Participant: George Giakkoupis.

Recently, Aspnes and Ruppert (DISC 2016) defined the following simple random experiment to determine the impact of concurrency on the performance of binary search trees: n randomly permuted keys arrive one at a time. When a new key arrives, it is first placed into a buffer of size c. Whenever the buffer is full, or when all keys have arrived, an adversary chooses one key from the buffer and inserts it into the binary search tree. The ability of the adversary to choose the next key to insert among c buffered keys, models a distributed system, where up to c processes try to insert keys concurrently. Aspnes and Ruppert showed that the expected average depth of nodes in the resulting tree is $O(\log n + c)$ for a comparison-based adversary, which can only take the relative order of arrived keys into account. In work [25], we generalize and strengthen this result. In particular, we allow an adversary that knows the actual values of all keys that have arrived, and show that the resulting expected average node depth is $D_{avg}(n) + O(c)$, where $D_{avg}(n) = 2ln(n) - Theta(1)$ is the expected average node depth of a random tree obtained in the standard unbuffered version of this experiment. Extending the bound by Aspnes and Ruppert to this stronger adversary model answers one of their open questions.

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

6.4.2. Acyclic strategy for silent self-stabilization in spanning forests

Participant: Anaïs Durand.

Self-stabilization is a general paradigm to enable the design of distributed systems tolerating any finite number of transient faults. Many self-stabilizing algorithms are designed using the same patterns. In [30] we formalize some of those design patterns to obtain general statements regarding both correctness and time complexity. Precisely, we study a class of algorithms devoted to networks endowed with a sense of direction describing a spanning forest whose characterization is a simple (i.e., quasi-syntactic) condition. We show that any algorithm of this class is (1) silent and self-stabilizing under the distributed unfair daemon (the weakest scheduling assumption in the considered model), and (2) has a stabilization time polynomial in moves and asymptotically optimal in rounds. Our condition mainly uses the concept of acyclic strategy, which is based on the notions of top-down and bottom-up actions. We have combined this formalization together with a notion of acyclic causality between actions and a last criteria called correct-alone (n.b., only this criteria is not syntactic) to obtain the notion of acyclic strategy. We show that any algorithm following an acyclic strategy reaches a terminal configuration in a polynomial number of moves, assuming a distributed unfair daemon. Hence, if its terminal configurations satisfy the specification, the algorithm is both silent and self-stabilizing. Unfortunately, we show that this condition is not sufficient to obtain an asymptotically optimal stabilization time in rounds. So, we enforce the acyclic strategy with the property of local mutual exclusivity to have an asymptotically optimal round complexity. We also propose a simple method to make any algorithm, that follows an acyclic strategy, locally mutually exclusive. This method has no overhead in moves. Finally, to show the versatility of our approach, we review works where our results apply.

This work was done in collaboration with Karine Altisen and Stéphane Devismes (VERIMAG, Université Grenoble Alpes).

6.4.3. Set agreement and renaming in the presence of contention-related crash failures

Participants: Anaïs Durand, Michel Raynal.

Given a predefined contention threshold λ, consider all executions in which process crashes are restricted to occur only when process contention is smaller than or equal to λ. If crashes occur after contention bypassed λ, there are no correctness guarantees (e.g., termination is not guaranteed). It is known that, when $\lambda = n-1$, consensus can be solved in an n-process asynchronous read/write system despite the crash of one process, thereby circumventing the well-known FLP impossibility result. Furthermore, it was shown that when $\lambda = n-k$ and $k \geq 2$, k-set agreement can be solved despite the crash of $2k-2$ processes.

In work [33] we consider two types of process crash failures: λ-constrained crash failures (as previously defined), and classical crash failures (that we call *any time* failures). We present two algorithms suited to these types of failures. The first algorithm solves k-set agreement, where $k = m + f$, in the presence of $t = 2m + f-1$ crash failures, $2m$ of them being $(n-k)$-constrained failures, and $(f-1)$ being any time failures. The second algorithm solves $(n + f)$-renaming in the presence of $t = m + f$ crash failures, m of them being $(n-t-1)$-constrained failures, and f being any time failures. It follows that the differentiation between λ-constrained crash failures and any time crash failures enlarges the space of executions in which the impossibility of k-set agreement and renaming in the presence of asynchrony and process crashes can be circumvented. In addition to its behavioral properties, both algorithms have a noteworthy first class property, namely, their simplicity.

This work was done in collaboration with Gadi Taubenfeld (IDC Herzliya).

6.4.4. Anonymous obstruction-free (n, k)-set agreement with $n-k + 1$ atomic read/write registers

Participant: Michel Raynal.

The k-set agreement problem is a generalization of the consensus problem. Namely, assuming that each process proposes a value, every non-faulty process must decide one of the proposed values, under the constraint that at most k different values are decided. This is a hard problem in the sense that it cannot be solved in a pure read/write asynchronous system, in which k or more processes may crash. One way to sidestep this impossibility result consists in weakening the termination property, requiring only that a process decides if it executes alone during a long enough period of time. This is the well-known obstruction-freedom progress condition. Consider a system of n anonymous asynchronous processes that communicate through atomic read/write registers, and such that any number of them may crash. In work [14] we address and solve the challenging open problem of designing an obstruction-free k-set agreement algorithm with only $(n{-}k+1)$ atomic registers. From a shared memory cost point of view, our algorithm is the best algorithm known to date, thereby establishing a new upper bound on the number of registers needed to solve this problem. For the consensus case ($k = 1$), the proposed algorithm is up to an additive factor of 1 close to the best known lower bound. Further, the paper extends this algorithm to obtain an x-obstruction-free solution to the k-set agreement problem that employs $(n{-}k+x)$ atomic registers, as well as a space-optimal solution for the repeated version of k-set agreement. Using this last extension, we prove that n registers are enough for every colorless task that is obstruction-free solvable with identifiers and any number of registers.

This work was done in collaboration with Zohir Bouzid and Pierre Sutra (CNRS).

HYBRID Project-Team

7. New Results

7.1. Virtual Reality Tools and Usages

7.1.1. Virtual Embodiment

Studying the Sense of Embodiment in VR Shared Experiences
Participants: Rebecca Fribourg, Ferran Argelaguet, Anatole Lécuyer

In [35], we explored the influence of sharing a virtual environment with another user on the sense of embodiment in virtual reality. For this aim, we conducted an experiment where users were immersed in a virtual environment while being embodied in an anthropomorphic virtual representation of themselves. To evaluate the influence of the presence of another user, two situations were studied: either users were immersed alone, or in the company of another user (see Figure 3). During the experiment, participants performed a virtual version of the well-known whac-a-mole game, therefore interacting with the virtual environment, while sitting at a virtual table. Our results show that users were significantly more "efficient" (i.e., faster reaction times), and accordingly more engaged, in performing the task when sharing the virtual environment, in particular for the more competitive tasks. Also, users experienced comparable levels of embodiment both when immersed alone or with another user. These results are supported by subjective questionnaires but also through behavioural responses, e.g. users reacting to the introduction of a threat towards their virtual body. Taken together, our results show that competition and shared experiences involving an avatar do not influence the sense of embodiment, but can increase user engagement. Such insights can be used by designers of virtual environments and virtual reality applications to develop more engaging applications.

This work was done with collaboration with Mimetic Inria team.

Figure 3. Studying the sense of embodiment in VR shared experiences: Setup of the experiment. Each user was able to interact in the virtual environment with his own avatar, while the physical setup provided both a reference frame and passive haptic feedback. From left to right: experimental conditions Alone, Mirror and Shared; Physical setup of the experiment.

Towards Novel Approaches to Characterise, Manipulate and Measure the Sense of Agency in Virtual Environments
Participants: Camille Jeunet, Ferran Argelaguet, Anatole Lécuyer

While the Sense of Agency (SoA) has so far been predominantly characterised in VR as a component of the Sense of Embodiment, other communities (e.g., in psychology or neurosciences) have investigated the SoA from a different perspective proposing complementary theories. Yet, despite the acknowledged potential benefits of catching up with these theories a gap remains. In [18], we first aimed to contribute to fill this gap by introducing a theory according to which the SoA can be divided into two components, the feeling and the judgment of agency, and relies on three principles, namely the principles of priority, exclusivity and consistency. We argued that this theory could provide insights on the factors influencing the SoA in VR systems. Second, we proposed novel approaches to manipulate the SoA in controlled VR experiments (based on these three principles) as well as to measure the SoA, and more specifically its two components based on neurophysiological markers, using ElectroEncephaloGraphy (EEG). We claim that these approaches would enable us to deepen our understanding of the SoA in VR contexts. Finally, we validated these approaches in an experiment (see Figure 4). Our results (N=24) suggest that our approach was successful in manipulating the SoA as the modulation of each of the three principles induced significant decreases of the SoA (measured using questionnaires). In addition, we recorded participants' EEG signals during the VR experiment, and neurophysiological markers of the SoA, potentially reflecting the feeling and judgment of agency specifically, were revealed. Our results also suggest that users' profile, more precisely their Locus of Control (LoC), influences their level of immersion and SoA.

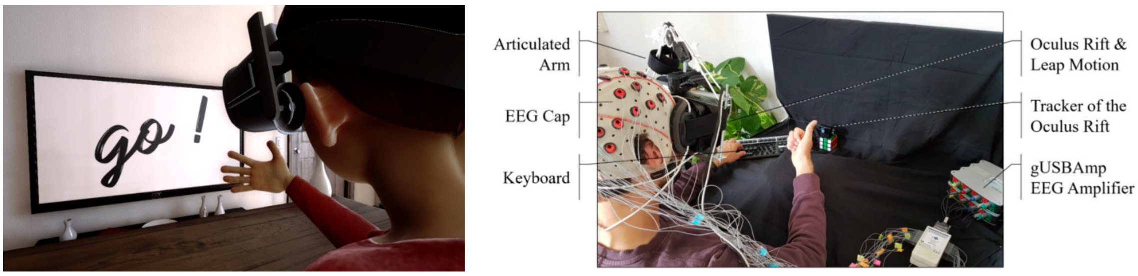


Figure 4. Studying the sense of agency in VR. Left: Third-person perspective: the participant receives a go signal and starts to perform the movement. Right: experimental set-up. The participant is equipped with an EEG cap, plugged to g.USBAmp amplifiers. In addition, he is immersed in the virtual environment using an Oculus Rift attached to his head and supported by an articulated arm (to avoid any pressure on the EEG cap and reduce the risk of muscular fatigue). Finally, his head is tracked by the Oculus tracker and his right hand is tracked by a Leap Motion fixed in front of the Oculus Rift.

Virtual Shadows for Real Humans in a CAVE: Influence on Virtual Embodiment and 3D Interaction
Participants: Guillaume Cortes, Ferran Argelaguet, Anatole Lécuyer

In immersive projection systems (IPS), the presence of the user's real body limits the possibility to elicit a virtual body ownership illusion. But, is it still possible to embody someone else in an IPS even though the users are aware of their real body? In order to study this question, we proposed to consider using a virtual shadow in the IPS, which can be similar or different from the real user's morphology [29]. We conducted an experiment (N=27) to study the users' sense of embodiment whenever a virtual shadow was or was not present (see Figure 5). Participants had to perform a 3D positioning task in which accuracy was the main requirement. The results showed that users widely accepted their virtual shadow (agency and ownership) and felt more comfortable when interacting with it (compare to no virtual shadow). Yet, due to the awareness of their real body, the users have less acceptance of the virtual shadow whenever the shadow gender differs from their own. Furthermore, the results showed that virtual shadows increase the users' spatial perception of the virtual environment by decreasing the inter-penetrations between the user and the virtual objects. Taken together, our results promote the use of dynamic and realistic virtual shadows in IPS and pave the way for further studies on “virtual shadow ownership” illusion.

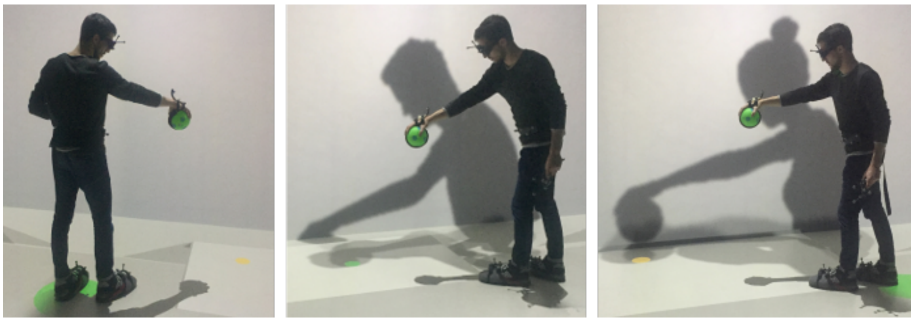


Figure 5. Various virtual shadows conditions. The participants performed a positioning task with 3 different virtual shadow conditions: None (N) (left), Male (M) (middle), Female (F) (right). The real shadow of the user is visible on the floor but does not match the natural behavior of a shadow in the virtual environment and is not taken into consideration.

This work was done with collaboration with Rainbow Inria team.

Influence of Being Embodied in an Obese Virtual Body on Shopping Behavior and Products Perception in VR

Participants: Jean-Marie Normand, Guillaume Moreau

In [26], we studied the changes an obese virtual body has on products perception (e.g., taste, etc.) and purchase behavior (e.g., number purchased) in an immersive virtual retail store. Participants (of a normal BMI on average) were embodied in a normal (N) or an obese (OB) virtual body and were asked to buy and evaluate food products in the immersive virtual store (see Figure 6 . Based on stereotypes that are classically associated with obese people, we expected that the group embodied in obese avatars would show a more unhealthy diet, (i.e., buy more food products and also buy more products with high energy intake, or saturated fat) and would rate unhealthy food as being tastier and healthier than participants embodied in “normal weight” avatars. Our participants also rated the perception of their virtual body: the OB group perceived their virtual body as significantly heavier and older. Stereotype activation failed for our participants embodied in obese avatars, who did not exhibit a shopping behavior following the (negative) stereotypes related to obese people. Participants might have rejected their virtual bodies when performing the shopping task, while the embodiment and presence ratings did not show significant differences, and purchased products based on their real (non-obese) bodies. This could mean that stereotype activation is more complex that previously thought.

7.1.2. VR and Building Information Modeling

OpenBIM-based Ontology for Interactive Virtual Environments

Participants: Anne-Solène Dris, François Lehericey, Valérie Gouranton, Bruno Arnaldi

We proposed an ontology improving the use of Building Information Modelling (BIM) models as an Interactive Virtual Environment (IVE) generator [33]. Our results enable to create a bidirectional link between the informed 3D database and the virtual reality application, and to automatically generate object-specific functions and capabilities according to their taxonomy. We presented an illustration of our results based on a Risk-Hunting training application. In such contexts, the notions of objects handling and scheduling of the construction are essential for the immersion of the future trainee as well as for the success of the training.

Risk-Hunting Training in Interactive Virtual Environments

Participants: Anne-Solène Dris, François Lehericey, Valérie Gouranton, Bruno Analdi

Figure 6. Being embodied in an obese virtual body. From left to right: A participant in a motion capture suit; The obese male avatar; A close-up on some products of our virtual store; The male avatar with a "normal" Body Mass Index.

Safety is an everlasting concern in construction environments. In such applications, when an accident happens it is rarely harmless. To raise awareness and train workers to safety procedures, training centers propose risk-hunting courses in which real-life equipment is set up in an incorrect way. Trainees can safely observe these environments and are supposed to point at risk situations. In [34], we proposed a risk-hunting course in Virtual Reality. With VR, we can put the trainee in a full construction environment with potentially dangerous hazards without engaging his safety. Contrary to others risk-hunting courses, we have designed a virtual environment with interactions to emphasize the importance of learning to correct the errors. First, instead of only having to spot the errors, the trainee had to fix them. Then, a second way to exploit VR interaction capabilities consisted in introducing consequences of not fixing an error. For example, not fixing an error in a scaffolding would make it collapse later. This implies to rely on script-writing the virtual environment to add causality on specific actions. Our goal was here to educate the trainee about the dramatic consequences that could arise when errors are not corrected.

7.1.3. Augmented Reality Methods and Applications

MoSART: Mobile Spatial Augmented Reality for 3D Interaction With Tangible Objects
Participants: Guillaume Cortes, Anatole Lécuyer

In [11] we introduced MoSART: a novel approach for Mobile Spatial Augmented Reality on Tangible objects. MoSART is dedicated to mobile interaction with tangible objects in single or collaborative situations. It is based on a novel "all-in-one" Head-Mounted Display (AMD) including a projector (for the SAR display) and cameras (for the scene registration). Equipped with the HMD the user is able to move freely around tangible objects and manipulate them at will. The system tracks the position and orientation of the tangible 3D objects and projects virtual content over them. The tracking is a feature-based stereo optical tracking providing high accuracy and low latency. A projection mapping technique is used for the projection on the tangible objects which can have a complex 3D geometry. Several interaction tools have also been designed to interact with the tangible and augmented content, such as a control panel and a pointer metaphor, which can benefit as well from the MoSART projection mapping and tracking features. The possibilities offered by our novel approach are illustrated in several use cases, in single or collaborative situations, such as for virtual prototyping, training or medical visualization.

This work was done with collaboration with Rainbow Inria team.
Evaluation of 2D and 3D Ultrasound Tracking Algorithms
Participants: Maud Marchal

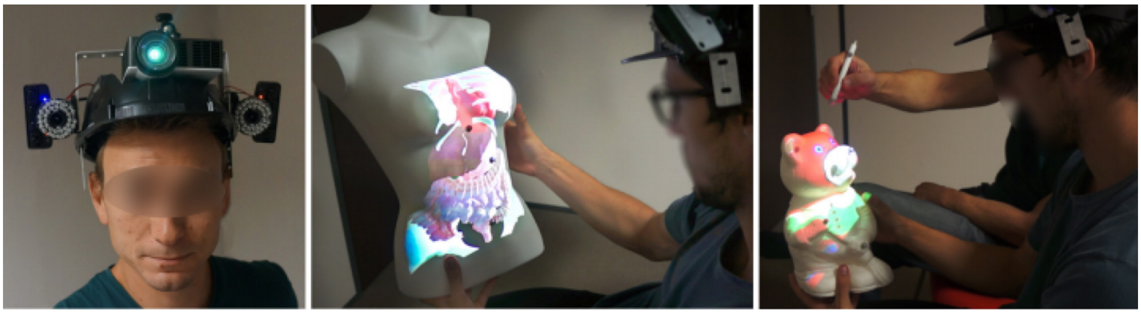


Figure 7. The MoSART headset for wearable augmented reality on tangible objects. Our novel system relies on an "all-in-one" Head-Mounted-Display (Left) which embeds a pico-projector for projection mapping and two cameras for feature-based stereo optical tracking of 3D tangible objects. The user can freely walk around and manipulate tangible objects superimposed with the projected images, such as for medical visualization purposes (Center). Tangible tools can also be used to interact with the virtual content such as for annotating or painting the objects in single or collaborative scenarios (Right).

Compensation for respiratory motion is important during abdominal cancer treatments. In [12], the results of the 2015 MICCAI Challenge on Liver Ultrasound Tracking are reported. These results extend the 2D results to relate them to clinical relevance in form of reducing treatment margins and hence sparing healthy tissues, while maintaining full duty cycle. The different methodologies of the MICCAI challenge are described for estimating and temporally predicting respiratory liver motion from continuous ultrasound imaging, used during ultrasound-guided radiation therapy. Furthermore, the trade-off between tracking accuracy and runtime in combination with temporal prediction strategies and their impact on treatment margins is also investigated. The paper follows the work of the PhD of Lucas Royer defended in 2016 and his methodology that was ranked first in the MICCAI challenge.

Evaluation of AR Inconsistencies on AR Placement Tasks: A VR Simulation Study

Participants: Romain Terrier, Jean-Marie Normand, Ferran Argelaguet

One of the major challenges of Augmented Reality (AR) is the registration of virtual and real contents. When errors occur during the registration process, inconsistencies between real and virtual contents arise and can alter user interaction. In this work, we assessed the impact of registration errors on the user performance and behaviour during an AR pick-and-place task in a Virtual Reality (VR) simulation [41]. The VR simulation ensured the repeatability and control over experimental conditions. The paper describes the VR simulation framework used and three experiments studying how registration errors (e.g., rotational errors, positional errors, shaking) and visualization modalities (e.g., transparency, occlusion) modify the user behaviour while performing a pick-and-place task. Our results show that users kept a constant behavior during the task, i.e., the interaction was driven either by the VR or the AR content, except if the registration errors did not enable to efficiently perform the task. Furthermore, users showed preference towards an half-transparent AR in which correct depth sorting is provided between AR and VR contents. Taken together, our results open perspectives for the design and evaluation of AR applications through VR simulation frameworks.

7.1.4. The 3DUI Contest 2018

Every year, the international IEEE Virtual Reality Conference organizes an annual 3D User Interfaces contest. This year, Hybrid submitted two different proposals.

Toward Intuitive 3D User Interfaces for Climbing, Flying and Stacking

Participants: Antonin Bernardin, Guillaume Cortes, Rebecca Fribourg, Tiffany Luong, Florian Nouviale, Hakim Si-Mohammed

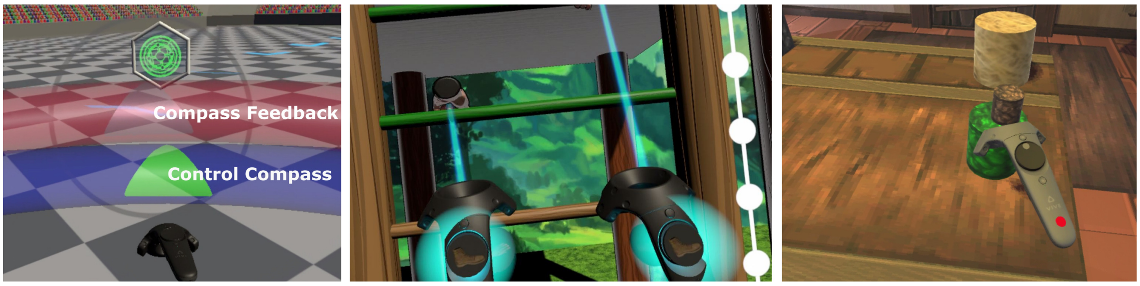


Figure 8. First solution proposed to the 2018 3DUI Contest. First-Person Drone Flying: The 3D User Interface used to control the drone (left). Ladder Climbing: First person point of view of the ladder climbing (center). Object Stacking: Physical object manipulation with frame recording as indicated by the red round on the controller and time control (right).

In this first solution, we proposed 3D user interfaces that are adapted to specific Virtual Reality tasks: climbing a ladder using a puppet metaphor, piloting a drone thanks to a 3D virtual compass and stacking 3D objects with physics-based manipulation and time control [28]. These metaphors have been designed to provide the user with an intuitive, playful and efficient way to perform each task (see Figure 8).

Climb, Fly, Stack: Design of Tangible and Gesture-based Interfaces for Natural and Efficient Interaction

Participants: Alexandre Audinot, Emeric Goga, Vincent Goupil, Carl-Johan Jorgensen, Adrien Reuzeau, Ferran Argelaguet

In this second solution we proposed three different 3D interaction metaphors conceived to fulfill the three tasks proposed in the IEEE VR 3DUI Contest. We proposed the VladdeR, a tangible interface for Virtual laddeR climbing, the FPDrone, a First Person Drone control flying interface, and the Dice Cup, a tangible interface for virtual object stacking [27]. All three metaphors take advantage of body proprioception and previous knowledge of real life interactions without the need of complex interaction mechanics (see Figure 9): climbing a tangible ladder through arm and leg motions, control a drone like a child flies an imaginary plane by extending your arms or stacking objects as you will grab and stack dice with a dice cup.

7.2. Physically-Based Simulation and Haptic Feedback

7.2.1. Haptic Methods and Rendering

KinesTouch: 3D Force-Feedback Rendering for Tactile Surfaces

Participants: Antoine Costes, Ferran Argelaguet, Anatole Lécuyer

Haptic enhancement of touchscreens has been mostly addressed through the use of various types of vibrations, altering the physics of the finger sliding on the screen, in order to provide friction forces and even small relief sensations. However, such approaches do not allow for displaying other haptic properties such as stiffness or large-scale shapes. In [30], we introduced the "KinesTouch", a novel approach for touchscreen enhancement providing four types of haptic feedback with a single force-feedback device: compliance, friction, fine roughness, and shape. Regarding friction in particular, we proposed a novel effect based on large lateral motion that increases or diminishes the sliding velocity between the finger and the screen. Our results show that this effect is able to produce distinct sliding sensations. Our general approach is also illustrated through a set of interactive use cases of 2D/3D content manipulation in various contexts.

This work was done in collaboration with Technicolor.

Haptic Material: a Holistic Approach for Haptic Texture Mapping

Participants: Antoine Costes, Ferran Argelaguet, Anatole Lécuyer

Figure 9. Second solution proposed to the 2018 3DUI Contest. Left, flying interface. Center, climbing interface. Left, stacking interface.

The development of 3D scanning technologies made common the digitizing of objects in realistic virtual copies, but still at the cost of most of their haptic properties. Besides, while haptic devices and setups spread widely, little attention is paid to the reuse and compatibility of haptic data, which are most of the time context- or hardware-specific. In [31], we proposed a new format for haptic texture mapping which is not dependent on the haptic rendering setup hardware. Our "haptic material" format encodes ten elementary haptic features in dedicated maps, similarly to "materials" used in computer graphics. These ten different features enable the expression of compliance, surface geometry and friction attributes through vibratory, cutaneous and kinesthetic cues, as well as thermal rendering. The diversity of haptic data allows various hardware to share this single format, each of them selecting which features to render depending on its capabilities.

This work was done in collaboration with Technicolor.

Combining Tangible Objects and Wearable Haptics

Participants: Xavier de Tinguy, Maud Marchal, Anatole Lécuyer

In [32], we studied the combination of tangible objects and wearable haptics for improving the display of stiffness sensations in virtual environments. Tangible objects enable to feel the general shape of objects, but they are often passive or unable to simulate several varying mechanical properties. Wearable haptic devices are portable and unobtrusive interfaces able to generate varying tactile sensations, but they often fail at providing convincing stiff contacts and distributed shape sensations. We propose to combine these two approaches in virtual and augmented reality (VR/AR), becoming able of arbitrarily augmenting the perceived stiffness of real/tangible objects by providing timely tactile stimuli at the fingers. We developed a proof-of-concept enabling to simulate varying elasticity/stiffness sensations when interacting with tangible objects by using wearable tactile modules at the fingertips. We carried out a user study showing that wearable haptic stimulation can well alter the perceived stiffness of real objects, even when the tactile stimuli is not delivered at the contact point. We illustrated our approach both in VR and AR, within several use cases and different tangible settings, such as when touching surfaces, pressing buttons and pistons, or holding an object (see Figure 12). Taken together, our results pave the way for novel haptic sensations in VR/AR by better exploiting the multiple ways of providing simple, unobtrusive, and low-cost haptic displays.

This work was done in collaboration with Rainbow Inria team.

7.2.2. Haptic Applications

A Survey on the Use of Haptic and Tactile Information in the Car to Improve Driving Safety

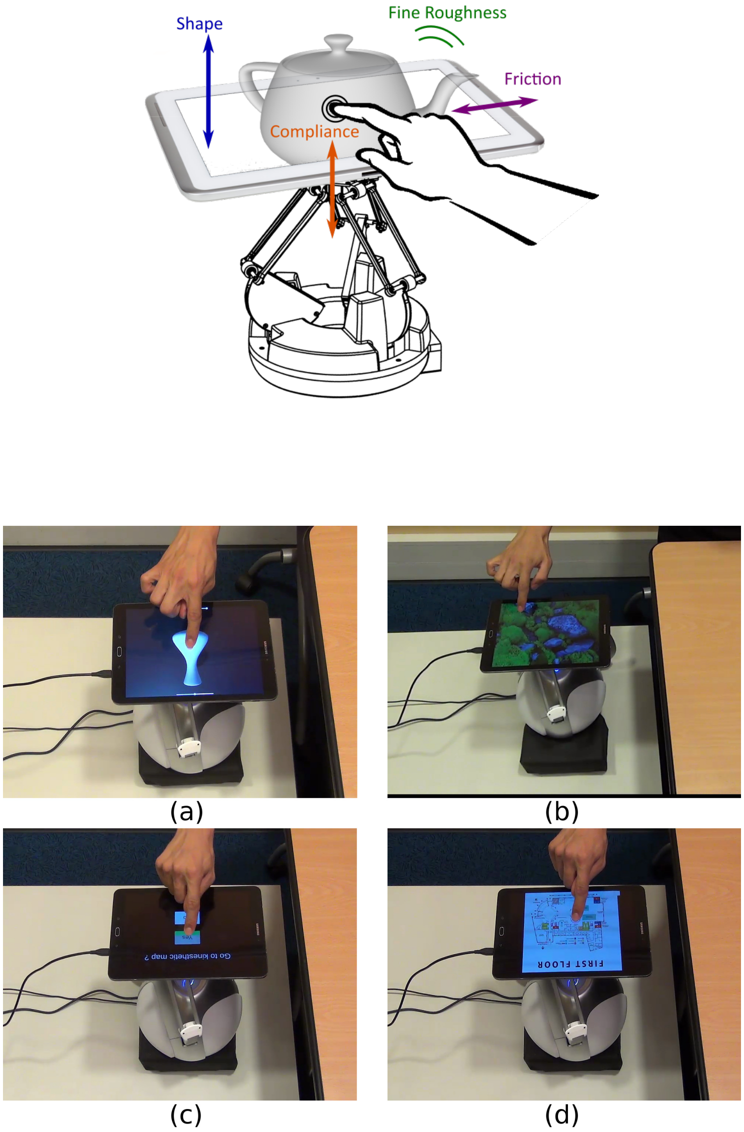


Figure 10. The KinesTouch approach. Top: concept of KinesTouch to provide four different types of haptic feedback to a touchscreen. Bottom: Use cases illustrating our approach: (a) Interaction with a 3D object, (b) Texture of a 2D image, (c) GUI and haptic buttons, (d) Interactive map.

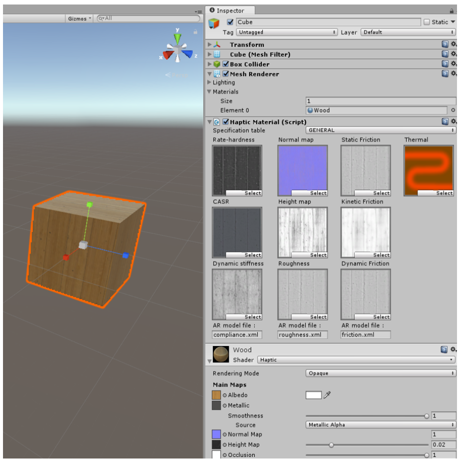


Figure 11. Implementation of our haptic material format in Unity3D.

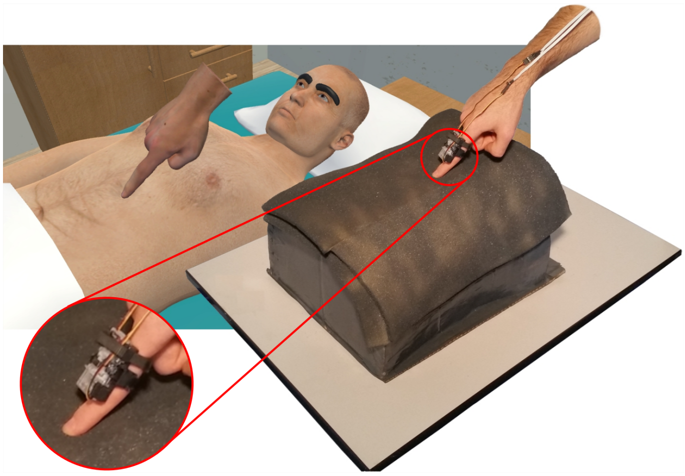


Figure 12. Combining tangible objects and wearable haptics in a VR medical palpation simulator. Passive tangible objects (a tangible chest here) provide haptic information about the global shape/percept of the virtual objects, while wearable haptic devices provide haptic information about dynamically changing mechanical properties (local elasticity here).

Participants: Yoren Gaffary, Anatole Lécuyer

In [15], we presented an overview of haptic technologies deployed in cars and their uses to enhance drivers' safety during manual driving. These technologies enable to deliver haptic (tactile or kinesthetic) feedback at various areas of the car, such as the steering wheel or the pedal. The paper explores two main uses of the haptic modality to fulfill the safety objective: providing driving assistance and warning. Driving assistance concerns the transmission of information usually conveyed with other modalities for controlling the cars' functions, maneuvering support, and guidance. Warning concerns the prevention of accidents using emergency warnings, increasing the awareness of surroundings, and preventing collisions, lane departures, and speeding. This paper discusses how haptic feedback has been introduced so far for these purposes and provides perspectives regarding the present and future of haptic cars meant to increase drivers' safety.

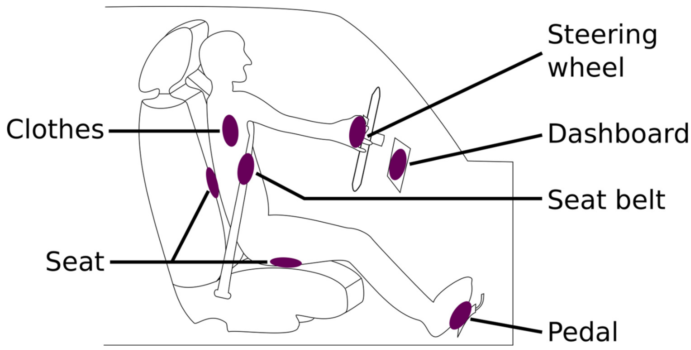


Figure 13. Using haptics in the car to improve driving safety: Different areas covered with haptic stimulations.

Toward Haptic Communication and Tactile Alphabets
Participants: Yoren Gaffary, Maud Marchal, Fernando Argelaguet Sanz, Anatole Lécuyer

In [14], we studied the possibility to convey information using tactile stimulation on fingertips. We designed and evaluated three tactile alphabets which are rendered by stretching the skin of the index's fingertip: (1) a Morse-like alphabet, (2) a symbolic alphabet using two successive dashes, and (3) a display of Roman letters based on the Unistrokes alphabet. All three alphabets (26 letters each) were evaluated through a user study in terms of recognition rate, intuitiveness and learning. Participants were able to perceive and recognize the letters with very good results (80%-97% recognition rates). Tactile alphabets with representations closer to Roman alphabet seem easier to learn. Taken together, our results pave the way to novel kinds of information communication using tactile modality.

This work was done in collaboration with CEA LIST.

7.3. Brain-Computer Interfaces

7.3.1. BCI Methods and Techniques

SimBCI: Novel Software Framework for Studying BCI Methods

Participants: Jussi Lindgren and Anatole Lécuyer

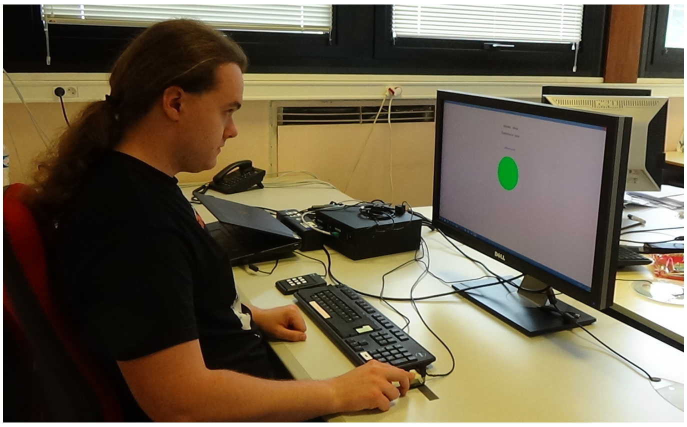


Figure 14. Toward tactile alphabets: A participant perceives a letter haptically stimulated using skin stretching at the level of his index.

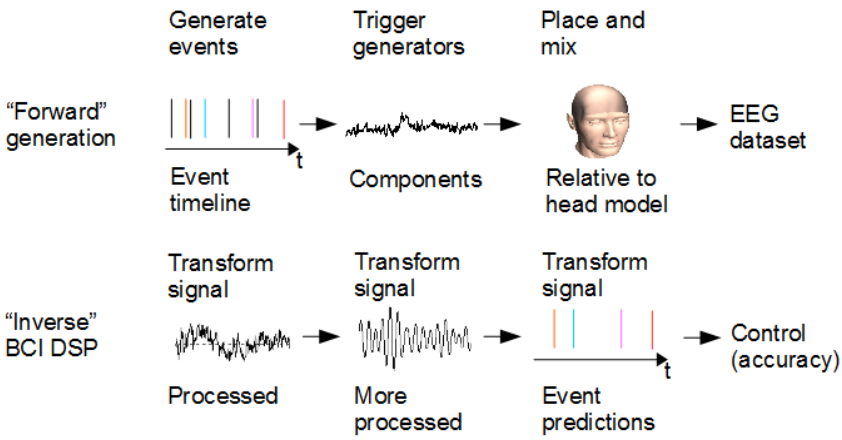


Figure 15. Simulated BCI data generation and testing in simBCI.

How to investigate the applicability of physiology-based source reconstruction for Brain-Computer Interfaces (BCIs)? The classic way is human experiments, but these unfortunately lack ground truth. The electrical activity inside the human brain is not fully described by external EEG measurements. In the CominLabs project SABRE, we have developed a BCI simulator framework called simBCI [22] to help in such studies. The framework allows modifying and changing generative models and their parameters inside a model brain, and studying what effects such changes have on the signal and subsequently the BCI signal processing. The modifiable parameters can include artifact properties, generative source locations, background activity characteristics and so on. We have released the framework as open source to the community (https://gitlab.inria.fr/sb/simbci/).

This work was done in collaboration with IMT Atlantique.

Novel Control Strategy for BCI Exploiting Visual Imagery and Attention

Participants: Jussi Lindgren and Anatole Lécuyer

Current paradigms for Brain-Computer Interfaces (BCIs) leave a lot to be desired in their accuracy and usability. We studied visual imagery as a potential new paradigm. In visual imagery, the user imagines objects or scenes visually, and the BCI is based on trying to classify the imagination type based on the EEG measurements. In [20], we studied to what extent can we distinguish the different mental processes of observing visual stimuli and imagining them based on the EEG. We found in a study of 26 users that we could somewhat differentiate (i) visual imagery vs. visual observation task (71% of classification accuracy), (ii) visual observation task towards different visual stimuli (classifying one observation cue versus another observation cue with an accuracy of 61%) and (iii) resting vs. observation/imagery (77% accuracy between imagery task versus resting state, and the accuracy of 75% between observation task versus resting state). All reported accuracies are averages over the users. Our results suggest that the presence of visual imagery and related alpha power changes may be useful to broaden the range of BCI control strategies.

Figure 16. Imagining or perceiving flowers and hammers. Can we tell from EEG which task the user is performing?

7.3.2. BCI Applications

BCI-based Interfaces for Augmented Reality: Feasibility, Design and Evaluation

Participants: Hakim Si-Mohammed, Camille Jeunet, Ferran Argelaguet and Anatole Lécuyer

In [25], we have studied the combination of BCI and Augmented Reality (AR). We first tested the feasibility of using BCI in AR settings based on Optical See-Through Head-Mounted Displays (OST-HMDs). Experimental results showed that a BCI and an OST-HMD equipment (EEG headset and Hololens in our case) are well compatible and that small movements of the head can be tolerated when using the BCI. Second, we introduced a design space for command display strategies based on BCI in AR, when exploiting a famous brain pattern called Steady-State Visually Evoked Potential (SSVEP). Our design space relies on five dimensions concerning the visual layout of the BCI menu ; namely: orientation, frame-of-reference, anchorage, size and explicitness. We implemented various BCI-based display strategies and tested them within the context of mobile robot control in AR. Our findings were finally integrated within an operational prototype based on a real mobile robot that is controlled in AR using a BCI and a HoloLens headset. Taken together our results (from four user studies) and our methodology could pave the way to future interaction schemes in Augmented Reality exploiting 3D User Interfaces based on brain activity and BCIs.

This work was done in collaboration with Loki Inria team.

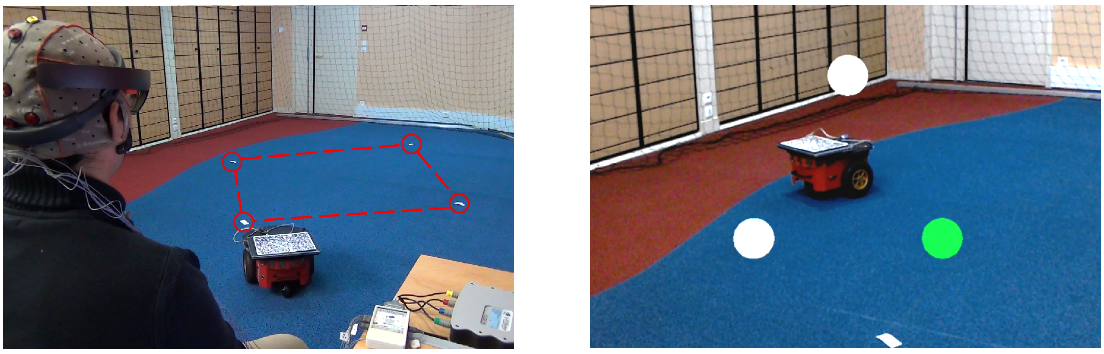


Figure 17. BCI-based interfaces in augmented reality: Illustration of our final prototype in use. (Left) general overview of the setup with the user equipped with EEG, sitting and facing the real mobile robot. (Right) First-person view, as seen from the HoloLens. The dashed line represents the path that the robot moved through during testing sessions.

Neurofeedback for Stroke Rehabilitation: A Case Report

Participants: Giulia Lioi, Mathis Fleury and Anatole Lécuyer

Neurofeedback (NF) consists in training self-regulation of brain activity by providing real-time information about the participant brain function. Few works have shown the potential of NF for stroke rehabilitation however its effectiveness has not been investigated yet. NF approaches are usually based on real-time monitoring of brain activity using a single imaging technique. Recent studies have revealed the potential of combining EEG and fMRI to achieve a more efficient and specific self-regulation. In this case report [49], we tested the feasibility of applying bimodal EEG-MRI NF on two stroke patients affected by left hemiplegia participated. The protocol included a calibration step (motor imagery of hemiplegic hand) and two NF sessions (5 minutes each). The experiment was run using a NF platform performing real-time EEG-fMRI processing and NF presentation. Both patients were found able to self-regulate their brain activity during the NF sessions. The EEG activity was harder to modulate than the BOLD activity. The patients were highly motivated to engage and satisfied with the NF animation, as assessed with a qualitative questionnaire. These results showed the feasibility and the potential of applying EEG-fMRI NF for stroke rehabilitation.

This work was done in collaboration with Visages Inria team.

Using EEG in Sport Performance Analysis

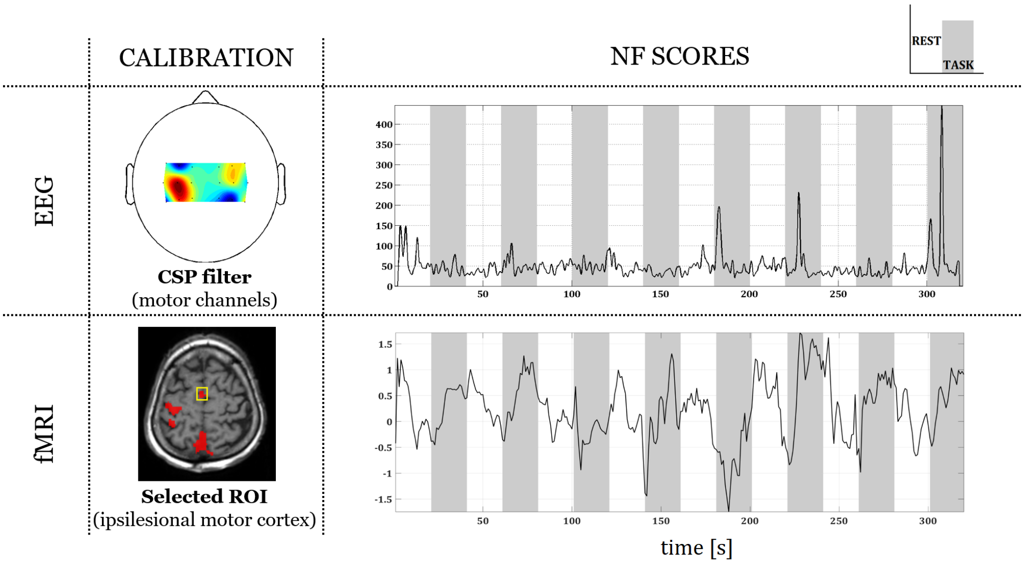


Figure 18. Neurofeedback for stroke patients: Examples of EEG and fMRI neurofeedback (NF) scores for a single session (patient 1). The left column represent the regions of interest selected to compute the NF signal during calibration. Resting blocks (20s) are indicated in white, NF training blocks (20s) in gray.

Participants: Ferran Argelaguet and Anatole Lécuyer

Competition changes the environment for athletes. The difficulty of training for such stressful events can lead to the well-known effect of "choking" under pressure, which prevents athletes from performing at their best level. To study the effect of competition on the human brain we recorded [24] pilot electroencephalography (EEG) data while novice shooters were immersed in a realistic virtual environment representing a shooting range. We found a differential between-subject effect of competition on mu (8–12 Hz) oscillatory activity during aiming; compared to training, the more the subject was able to desynchronize his mu rhythm during competition, the better was his shooting performance. Because this differential effect could not be explained by differences in simple measures of the kinematics and muscular activity, nor by the effect of competition or shooting performance per se, we interpret our results as evidence that mu desynchronization has a positive effect on performance during competition. It remains to show whether this effect can be generalized to expert shooters. Our findings could be relevant in sports training to help athletes avoid choking under pressure during competition. Confirmation through further experimental validation is however needed.

This work was done in collaboration with EPFL.

7.4. Cultural Heritage

Through several collaborations with Cultural Heritage partners such as archaeologists, historians, or curators, the Hybrid team has developed a methodology to propose new practices and tools in this domain. This methodology combines different technologies of digitization, such as CT scan, photogrammetry, or lidar, 3D production, such as 3D modelling or 3D printing, and 3D interactions in VR and AR.

7.4.1. 3D Printing and AR Applications

Lift the Veil of the Block Samples from the Warcq Chariot Burial

Participants: Ronan Gaugne and Valérie Gouranton

Figure 19. Experimental setup used in our sport performance study. (Left) Subjects were standing in our immersive projection system and were able to interact with the system using an ART Flystick. (Right) Subjects were wearing a high-density 64 channels EEG cap.

Cultural Heritage (CH) professionals such as archaeologists and conservators regularly experience the problem of working on concealed artifacts and face the potential destruction of source material without real understanding of the internal structure or state of decay or modification of the initial context by the micro-excavation process. Medical images-based digitization, such as MRI or CT scan, are increasingly used in CH as they provide information on the internal structure of archaeological material. Likewise, additive technologies are used more and more in the Cultural Heritage process, for example, in order to reproduce, complete, study or exhibit artifacts. 3D copies are based on digitization techniques such as laser scan or photogrammetry. In this case, the 3d copy remains limited to the external surface of objects. Different previous works illustrated the interest of combining 3D printing and Computed Tomography (CT) scans in order to extract concealed artifacts from larger archaeological material. The method was based on 3D segmentation techniques within volume data obtained by CT scans to isolate nested objects. This approach was useful to perform a digital extraction, but in some case it is also interesting to observe the internal spatial organization of an intricate object in order to understand its production process. Then, we proposed a method for the representation of a complex internal structure based on a combination of CT scan and emerging 3D printing techniques mixing colored and transparent parts of an aggregate of objects (see Figure 20), with very small pieces, from an exceptional aristocratic Gallic grave in the context of a preventive archaeological investigation [39].

This project was done in collaboration with UMR Trajectoires, Inrap and Image ET/BCRX.

Digital Introspection of a Mummy Cat

Participants: Ronan Gaugne and Valérie Gouranton

In the last decade, thanks to the dissemination of novel medical imaging technologies, research on the study of animal mummies of Ancient Egypt has became more and more important, leading to a better understanding of the history and culture of this civilization. Modern 3D technologies such as virtual reality, augmented reality and 3D printing enable to enrich the research process and open innovative possibilities for scenography in scientific mediation. In [36] we focused on one particular mummy cat and proposed to combine CT scan, 3D printing and augmented reality in a global process to accompany and support at the same time a scientific study of the object and a preparation of a mediation action in a Museum (see Figure 21 and Figure 22).

Figure 20. Transparent 3D printings from CT scan of archaeological materials.

This project was done in collaboration with Inrap, UMR Trajectoires, HISoMA and Musée des Beaux-Arts, Rennes.

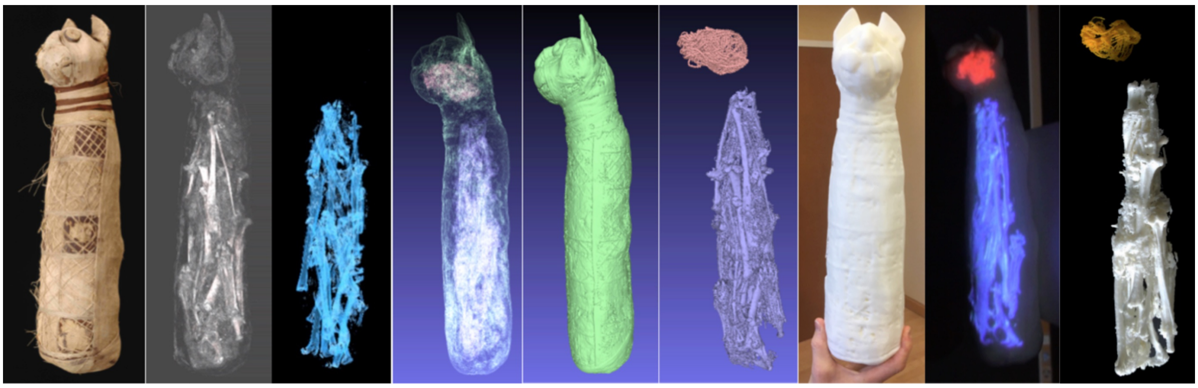


Figure 21. Digital introspection of a mummy cat. From left to right: the original mummy, CT scan of the mummy, volume rendering of the bones, mesh generation from CT scan, mesh of the external shape, meshes of internal parts, 3D printing of the external shape, projective AR of internal parts, 3D printing of internal parts.

7.4.2. VR Applications

EvoluSon: Walking through an Interactive History of Music

Participants: Ronan Gaugne, Florian Nouviale and Valérie Gouranton

The EvoluSon project [16] proposes an immersive experience where the spectator explores an interactive visual and musical representation of the main periods of the history of Western music (see Figure 23). The musical content is constituted of original musical compositions based on the theme of Bach's Art of Fugue to illustrate

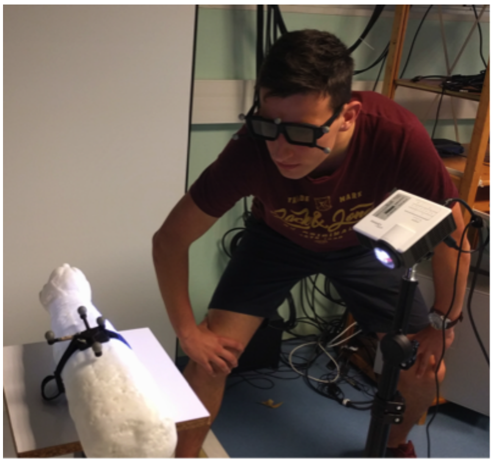


Figure 22. Projective AR system used for the visualization of the internal content of a mummy cat.

the 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. This project is the result of collaboration between a computer science research laboratory and a research laboratory on art and music. It was first presented to a public event on science and music organised by the computer science research laboratory.

This project was done in collaboration with the Research Laboratory on Art and Music of University Rennes 2.

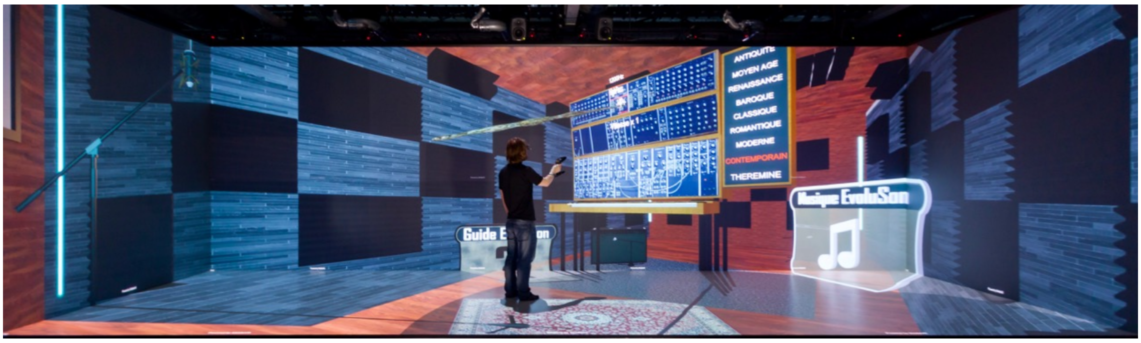


Figure 23. Interacting with the music through the ages inside Evoluson.

INSIDE Interactive and Non-destructive Solution for Introspection in Digital Environments

Participants: Flavien Lécuyer, Valérie Gouranton, Ronan Gaugne and Bruno Arnaldi

The development of scanning technologies allowed to limit the destructiveness induced by the excavation. However, it is not enough, as the rendering is not enough to study a scanned artifact. We proposed to use virtual reality as a legitimate tool for the inspection of artifacts modelled in 3D: INSIDE [38], with tools to lead a complete virtual excavation (see Figure 24). This tool opens a new way of practicing archaeology, more efficient and safer for the content being excavated.

This project was done in collaboration with the Research Laboratory on Archeology and History, UMR CReAAH, UMR Trajectoires, and Inrap.

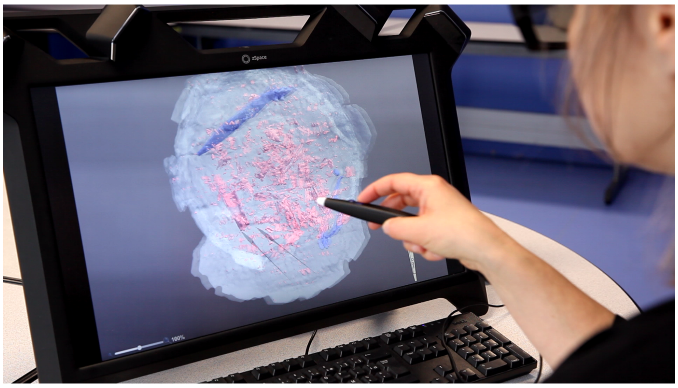


Figure 24. The INSIDE system used by an archaeologist within a workbench system.

VR Interactions with Multiple Interpretations of Archaeological Artefacts

Participants: Ronan Gaugne and Valérie Gouranton

The incorporation of 3D printed artefacts into Virtual Reality and Augmented Reality experiences is gaining strong interest from Cultural Heritage professionals. Indeed, in most cases, virtual environments cannot convey information such as the physical properties of artefacts. In [37], we presented a methodology for the development of VR experiences which incorporate 3D replicas of artefacts as user interfaces. The methodology is applied on the development of an experience to present various interpretations of an urn which was found at the edge of a cliff on the south east coastal area of the United Kingdom in 1910. In order to support the understanding of the multiple interpretations of this artefact, the system deploys a virtual environment and a physical replica to allow users to interact with the artefacts and the environment (see Figure 25). Feedback from heritage users suggests VR technologies along with digitally fabricated replicas can meaningfully engage audiences with multiple interpretations of cultural heritage artefacts.

This project was done in collaboration with University of Brighton (UK), Inrap, CNRS and UMR CReAAH.

Figure 25. Interaction in VR using the physical replica of a funeral urn.

LACODAM Project-Team

7. New Results

7.1. Introduction

In this section, we organize the bulk of our contributions this year along two of our research axes, namely Pattern Mining and Decision Support. Some other contributions lie within the domains of query optimization and machine learning.

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 mining of negative patterns [24], [14] and periodic patterns [18].
- *Data Mining for the masses.* In [11], 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.
- *Efficient pattern mining.* In [10], we propose a method to sample itemsets efficiently on streaming data. This contribution tackles two limitations of the state of the art in pattern mining: (1) the so-called pattern explosion —the user is confronted to too many patterns—, and (2) the assumption of static data.
- *Data Mining for Data Science*. One of the most basic types of patterns is to know if the data makes one single group, i.e., is *unimodal*, or can be clustered into several groups. In [13], we propose a new statistical test of unimodality, that is both independent of the input distribution and computationally efficient.

7.1.2. Decision Support

In regards to the axis of decision support, our contributions can be organized in two categories: forecasting & prediction, and anomaly detection.

- *Forecasting & prediction.* In [15], [12], we propose solutions to automate the task of capacity planning in the context of a large data network as the one available at Orange. The work in [19] offers a tool to predict the nutritional needs of sows in lactation.
- *Anomaly Detection.* The work in [20] tackles the problem of fraud detection under imbalanced data.

7.1.3. Others

- *Machine Learning.*[16] proposes a novel algorithm to weight the importance of classification errors when training a classifier. [8] proposes a classification algorithm optimized for highly imbalanced data.
- Query optimization. In [9] we propose a query-load-agnostic caching approach to speed-up provenance-aware queries in RDF data cubes.

7.2. Mining Periodic Patterns with a MDL Criterion

Participants: E. Galbrun, P. Cellier, N. Tatti, A. Termier, B. Crémilleux

The quantity of event logs available is increasing rapidly, be they produced by industrial processes, computing systems, or life tracking, for instance. It is thus important to design effective ways to uncover the information they contain. Because event logs often record repetitive phenomena, mining periodic patterns is especially relevant when considering such data. Indeed, capturing such regularities is instrumental in providing condensed representations of the event sequences. The work in [18] presents an approach for mining periodic patterns from event logs while relying on a Minimum Description Length (MDL) criterion to evaluate candidate patterns. Our goal is to extract a set of patterns that suitably characterises the periodic structure present in the data. We evaluate the interest of our approach on several real-world event log datasets.

7.3. NegPSpan: Efficient Extraction of Negative Sequential Patterns with Embedding Constraints

Participants: T. Guyet, R. Quinou
Mining frequent sequential patterns consists in extracting recurrent behaviors, modeled as patterns, in a big sequence dataset. Such patterns inform about which events are frequently observed in sequences, i.e., what does really happen. Sometimes, knowing that some specific event does not happen is more informative than extracting a lot of observed events. Negative sequential patterns (NSP) formulate recurrent behaviors by patterns containing both observed events and absent events. Few approaches have been proposed to mine such NSPs. In addition, the syntax and semantics of NSPs differ in the different methods which makes it difficult to compare them. [24] provides a unified framework for the formulation of the syntax and the semantics of NSPs. Then, it introduces a new algorithm, NegPSpan, that extracts NSPs using a PrefixSpan depth-first scheme and enabling maxgap constraints that other approaches do not take into account. The formal framework allows for highlighting the differences between the proposed approach w.r.t. to the methods from the literature, especially w.r.t. the state of the art approach eNSP. Intensive experiments on synthetic and real datasets show that NegPSpan can extract meaningful NSPs and that it can process bigger datasets than eNSP thanks to significantly lower memory requirements and better computation times.

7.4. NTGSP: Mining Negative Temporal Patterns

Participants: K. Tsesmeli, M. Boumghar, T. Guyet, R. Quiniou, L. Pierre
In [14] the authors study the problem of extracting frequent patterns containing positive events, negative events specifying the absence of events as well as temporal information on the delay between these events. [14] defines the semantics of such patterns and proposes the NTGSP method based on state-of-the-art approaches. The performance of the method is evaluated on commercial data provided by EDF (Électricité de France).

7.5. Accelerating Itemset Sampling using Satisfiability Constraints on FPGA

Participants: M. Gueguen, O. Sentieys, A. Termier
Finding recurrent patterns within a data stream is important for fields as diverse as cybersecurity or e-commerce. This requires to use pattern mining techniques. However, pattern mining suffers from two issues. The first one, known as “pattern explosion”, comes from the large combinatorial space explored and is the result of too many patterns output for analysis. Recent techniques, called *output space sampling* solve this problem by outputting only a sample of the results, with a target size provided by the user. The second issue is that most algorithms are designed to operate on static datasets or low throughput streams. In [10], the authors propose a contribution to tackle both issues, by designing an FPGA accelerator for pattern mining with output space sampling. They show that their accelerator can outperform a state-of-the-art implementation on a server class CPU using a modest FPGA product.

7.6. Are your data data gathered? The Folding Test of Unimodality

Participants: A. Siffer, C. Largouët, A. Termier

Understanding data distributions is one of the most fundamental research topics in data analysis. The literature provides a great deal of powerful statistical learning algorithms to gain knowledge on the underlying distribution given multivariate observations. We are likely to find out a dependence between features, the appearance of clusters or the presence of outliers. Before such deep investigations, [13] proposes the folding test of unimodality. As a simple statistical description, it allows to detect whether data are gathered or not (unimodal or multimodal). To the best of our knowledge, this is the first multivariate and purely statistical unimodality test. It makes no distribution assumption and relies only on a straightforward p-value. Experiments on real world data show the relevance of the test and how to use it for the task of clustering.

7.7. Day-ahead Time Series Forecasting: Application to Capacity Planning

Participants: C. Leverger, V. Lemaire, S. Malinowski, T. Guyet, L. Rozé

In the context of capacity planning, forecasting the evolution of server usage enables companies to better manage their computational resources. The work in [12] addresses this problem by collecting key indicator time series. The article proposes a method to forecast the evolution of server usage one day-ahead. The method assumes that data is structured by a daily seasonality, but also that there is typical evolution of indicators within a day. Then, it uses the combination of a clustering algorithm and Markov Models to produce day-ahead forecasts. Our experiments on real datasets show that the data satisfies our assumption and that, in the case study, our method outperforms classical approaches (AR, Holt-Winters).

7.8. PerForecast: A Tool to Forecast the Evolution of Time Series for Capacity Planning.

Participants: C. Leverger, R. Marguerie, V. Lemaire, T. Guyet, S. Malinowski

The work published in [15] presents PerForecast, a tool for automatic capacity planning. The tool relies on univariate temporal data and automatically configured predictive models. The aim is to anticipate *capacity problems* in the infrastructure of Orange in order to ensure the delivery of services to customers. For example, PerForecast can predict the overhead of a server at the earliest possible stage, so that new machines can be ordered before the deterioration of the service in question. The purchase procedures being long and costly, the earlier they are done, the better the quality of service.

7.9. Tree-based Cost-Sensitive Methods for Fraud Detection in Imbalanced Data

Participants: G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, M. Sebban

Bank fraud detection is a difficult classification problem where the number of frauds is much smaller than the number of genuine transactions. The authors of [20] present cost sensitive tree-based learning strategies applied in this context of highly imbalanced data. The paper first proposes a cost sensitive splitting criterion for decision trees that takes into account the cost of each transaction. Then the criterion is extended with a decision rule for classification with tree ensembles. The authors then propose a new cost-sensitive loss for gradient boosting. Both methods have been shown to be particularly relevant in the context of imbalanced data. Experiments on a proprietary dataset of bank fraud detection in retail transactions show that the presented cost sensitive algorithms increase the retailer's benefits by 1,43% compared to non cost-sensitive ones and that the gradient boosting approach outperforms all its competitors.

7.10. An Algorithm to Optimize the F-measure by Proper Weighting of Classification Errors

Participants: K. Bascol, R. Emonet, E. Fromont, A. Habrard, G. Metzler, M. Sebban

[16] proposes an F-Measure optimization algorithm with theoretical guarantees that can be used with any error-weighting learning method. The algorithm, iteratively generates a set of costs from the training set so that the final classifier has an F-measure close to optimal. The optimality of the F-measure is expressed using a finer upper bound as presented in [31]. Furthermore, we show that the costs obtained at each iteration of our method can drastically reduce the search space and thus converge quickly to the optimal parameters. The efficiency of the method is shown both in terms of F-measurement but also in terms of speed of convergence on several unbalanced datasets.

7.11. Learning Maximum excluding Ellipsoids from Imbalanced Data with Theoretical Guarantees

Participants: G. Metzler, X. Badiche, B. Belkasmi, E. Fromont, A. Habrard, M. Sebban

[8] addresses the problem of learning from imbalanced data. The authors consider the scenario where the number of negative examples is much larger than the number of positive ones. This work proposes a theoretically-founded method, which learns a set of local ellipsoids centered at the minority class examples while excluding the negative examples of the majority class. This task is addressed from a Mahalanobis-like metric learning point of view, which allows deriving generalization guarantees on the learned metric using the uniform stability framework. The experimental evaluation on classic benchmarks and on a proprietary dataset in bank fraud detection shows the effectiveness of the approach, particularly when the imbalance is huge.

7.12. Answering Provenance-Aware Queries on RDF Data Cubes under Memory Budgets

Participants: L. Galárraga, K. Ahlstrøm, K. Hose, T. B. Pedersen

The steadily-growing popularity of semantic data on the Web and the support for aggregation queries in SPARQL 1.1 have propelled the interest in Online Analytical Processing (OLAP) and data cubes in RDF. Query processing in such settings is challenging because SPARQL OLAP queries usually contain many triple patterns with grouping and aggregation. Moreover, one important factor of query answering on Web data is its provenance, i.e., metadata about its origin. Some applications in data analytics and access control require to augment the data with provenance metadata and run queries that impose constraints on this provenance. This task is called provenance-aware query answering. The work in [9] investigates the benefit of caching some parts of an RDF cube augmented with provenance information when answering provenance-aware SPARQL queries. [9] proposes provenance-aware caching (PAC), a caching approach based on a provenance-aware partitioning of RDF graphs, and a benefit model for RDF cubes and SPARQL queries with aggregation. The results on real and synthetic data show that PAC outperforms significantly the LRU strategy (least recently used) and the Jena TDB native caching in terms of hit-rate and response time.

LINKMEDIA Project-Team

6. New Results

6.1. Low-level content description and indexing

6.1.1. Scalability of the NV-tree: Three Experiments

Participants: Laurent Amsaleg, Björn Þór Jónsson [Univ. Copenhagen], Herwig Lejsek [Videntifier Tech.].

The NV-tree is a scalable approximate high-dimensional indexing method specifically designed for large-scale visual instance search. We report in [10] on three experiments designed to evaluate the performance of the NV-tree. Two of these experiments embed standard benchmarks within collections of up to 28.5 billion features, representing the largest single-server collection ever reported in the literature. The results show that indeed the NV-tree performs very well for visual instance search applications over large-scale collections.

6.1.2. Prototyping a Web-Scale Multimedia Retrieval Service Using Spark

Participants: Laurent Amsaleg, Gylfi Þór Gudmundsson [School of Computer Science, Reykjavik], Björn Þór Jónsson [Univ. Copenhagen], Michael Franklin [Computer Science Division, Berkeley].

The world has experienced phenomenal growth in data production and storage in recent years, much of which has taken the form of media files. At the same time, computing power has become abundant with multi-core machines, grids, and clouds. Yet it remains a challenge to harness the available power and move toward gracefully searching and retrieving from web-scale media collections. Several researchers have experimented with using automatically distributed computing frameworks, notably Hadoop and Spark, for processing multimedia material, but mostly using small collections on small computing clusters. In [3] we describe a prototype of a (near) web-scale throughput-oriented MM retrieval service using the Spark framework running on the AWS cloud service. We present retrieval 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. We also present a publicly available demonstration retrieval system, running on our own servers, where the implementation of the Spark pipelines can be observed in practice using standard image benchmarks, and downloaded for research purposes. Finally, we describe a method to evaluate retrieval quality of the ever-growing high-dimensional index of the prototype, without actually indexing a web-scale media collection.

6.1.3. Extreme-value-theoretic estimation of local intrinsic dimensionality

Participants: Laurent Amsaleg, Teddy Furon, Oussama Chelly [National Institute of Informatics], Stéphane Girard [MISTIS, Inria Grenoble], Michael Houle [National Institute of Informatics], Ken-Ichi Kawarabayashi [National Institute of Informatics], Michael Nett [Google].

This work is concerned with the estimation of a local measure of intrinsic dimensionality (ID) recently proposed by Houle. The local model can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the distribution of distances to a query point is modeled in terms of a continuous random variable. This form of intrinsic dimensionality can be particularly useful in search, classification, outlier detection, and other contexts in machine learning, databases, and data mining, as it has been shown to be equivalent to a measure of the discriminative power of similarity functions. Several estimators of local ID are proposed and analyzed based on extreme value theory, using maximum likelihood estimation, the method of moments, probability weighted moments, and regularly varying functions, see [2]. An experimental evaluation is also provided, using both real and artificial data.

6.1.4. Intrinsic dimensionality for Information Retrieval

Participant: Vincent Claveau.

Examining the properties of representation spaces for documents or words in Information Retrieval (IR) brings precious insights to help the retrieval process. Following the work presented in the previous paragraph, it has been shown that intrinsic dimensionality is chiefly tied with the notion of indiscriminateness among neighbors of a query point in the vector space. In this work [13], we revisit this notion in the specific case of IR. More precisely, we show how to estimate indiscriminateness from IR similarities in order to use it in representation spaces used for documents and words. We show that indiscriminateness may be used to characterize difficult queries; moreover we show that this notion, applied to word embeddings, can help to choose terms to use for query expansion.

6.1.5. Heat Map Based Feature Ranker

Participants: Christian Raymond, Carlos Huertas [Autonomous University of Baja California, Mexico], Reyes Uarez-Ramirez [Autonomous University of Baja California, Mexico].

In [6], we present Heat Map Based Feature Ranker, an algorithm to estimate feature importance purely based on its interaction with other variables. A compression mechanism reduces evaluation space up to 66% without compromising efficacy. Our experiments show that our proposal is very competitive against popular algorithms, producing stable results across different types of data. We also show how noise reduction through feature selection aids data visualization using emergent self-organizing maps.

6.1.6. Time series retrieval and indexing using DTW-preserving shapelets

Participants: Laurent Amsaleg, Ricardo Carlini Sperandio, Simon Malinowski, Romain Tavenard [Univ. Rennes 2].

Dynamic Time Warping (DTW) is a very popular similarity measure used for time series classification, retrieval or clustering. DTW is, however, a costly measure, and its application on numerous and/or very long time series is difficult in practice. We have proposed a new approach for time series retrieval: time series are embedded into another space where the search procedure is less computationally demanding, while still accurate. This approach is based on transforming time series into high-dimensional vectors using DTW-preserving shapelets. That transform is such that the relative distance between the vectors in the Euclidean transformed space well reflects the corresponding DTW measurements in the original space. We have also proposed in [12] strategies for selecting a subset of shapelets in the transformed space, resulting in a trade-off between the complexity of the transformation and the accuracy of the retrieval. Experimental results using the well known time series datasets demonstrate the importance of this trade-off. This transformation can then be used to build efficient time series indexing schemes.

6.1.7. Fast Spectral Ranking for Similarity Search

Participants: Yannis Avrithis, Teddy Furon, Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Ondra Chum [Univ. Prague].

Despite the success of deep learning on representing images for particular object retrieval, recent studies show that the learned representations still lie on manifolds in a high dimensional space. This makes the Euclidean nearest neighbor search biased for this task. Exploring the manifolds online remains expensive even if a nearest neighbor graph has been computed offline. This work introduces an explicit embedding reducing manifold search to Euclidean search followed by dot product similarity search. This is equivalent to linear graph filtering of a sparse signal in the frequency domain. To speed up online search, we compute an approximate Fourier basis of the graph offline. We improve the state of art on particular object retrieval datasets including the challenging Instre dataset containing small objects. At a scale of 10^5 images, the offline cost is only a few hours, while query time is comparable to standard similarity search [15].

6.1.8. Mining on Manifolds: Metric Learning without Labels

Participants: Yannis Avrithis, Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Ondra Chum [Univ. Prague].

In this work we present a novel unsupervised framework for hard training example mining [17]. The only input to the method is a collection of images relevant to the target application and a meaningful initial representation, provided e.g. by pre-trained CNN. Positive examples are distant points on a single manifold, while negative examples are nearby points on different manifolds. Both types of examples are revealed by disagreements between Euclidean and manifold similarities. The discovered examples can be used in training with any discriminative loss. The method is applied to unsupervised fine-tuning of pre-trained networks for fine-grained classification and particular object retrieval. Our models are on par or are outperforming prior models that are fully or partially supervised.

6.1.9. Hybrid Diffusion: Spectral-Temporal Graph Filtering for Manifold Ranking

Participants: Yannis Avrithis, Teddy Furon, Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Ondra Chum [Univ. Prague].

State of the art image retrieval performance is achieved with CNN features and manifold ranking using a k-NN similarity graph that is pre-computed off-line. The two most successful existing approaches are temporal filtering, where manifold ranking amounts to solving a sparse linear system online, and spectral filtering, where eigen-decomposition of the adjacency matrix is performed off-line and then manifold ranking amounts to dot-product search online. The former suffers from expensive queries and the latter from significant space overhead. Here we introduce a novel, theoretically well-founded hybrid filtering approach allowing full control of the space-time trade-off between these two extremes. Experimentally, we verify that our hybrid method delivers results on par with the state of the art, with lower memory demands compared to spectral filtering approaches and faster compared to temporal filtering [16].

6.1.10. Transactional Support for Visual Instance Search

Participants: Laurent Amsaleg, Björn Þór Jónsson [Univ. Copenhagen], Herwig Lejsek [Videntifier Tech.].

This work addresses the issue of dynamicity and durability for scalable indexing of very large and rapidly growing collections of local features for visual instance retrieval. By extending the NV-tree, a scalable disk-based high-dimensional index, we show how to implement the ACID properties of transactions which ensure both dynamicity and durability. We present a detailed performance evaluation of the transactional NV-tree, showing that the insertion throughput is excellent despite the effort to enforce the ACID properties [20].

6.1.11. Time-series prediction for capacity planning

Participants: Simon Malinowski, Colin Leverger [Orange Labs], Thomas Guyet [AgroCampus Ouest], Vincent Lemaire [Orange Labs].

In a collaboration with Orange Labs, we have worked on KPI time series prediction in order to improve capacity planning. A software has been develloped. This software is detailed in [32]. It aims at visualizing and comparing different time series prediction techniques on user-defined input data. We have also developed a novel prediction algorithm that focuses on time series for with a seasonality [21]. It uses the combination of a clustering algorithm and Markov Models to produce day-ahead forecasts. Our experiments on real datasets show that in the case study, our method outperforms classical approaches (AR, Holt-Winters).

6.1.12. Scale-adaptive CNN for Crowd counting

Participants: Miaojing Shi, Lu Zhang [Fudan Univ.], Qiaobo Chen [Shanghai Jiaotong Univ.].

The task of crowd counting is to automatically estimate the pedestrian number in crowd images. To cope with the scale and perspective changes that commonly exist in crowd images, this work proposes a scale-adaptive CNN (SaCNN) architecture with a backbone of fixed small receptive fields. We extract feature maps from multiple layers and adapt them to have the same output size; we combine them to produce the final density map. The number of people is computed by integrating the density map. We also introduce a relative count loss along with the density map loss to improve the network generalization on crowd scenes with few pedestrians, where most representative approaches perform poorly on. We conduct extensive experiments and demonstrate significant improvements of SaCNN over the state-of-the-art [31].

6.1.13. Revisiting Perspective information for Efficient Crowd counting

Participants: Miaojing Shi, Zhaohui Yang [Peking Univ.], Chao Xu [Peking Univ.], Qijun Chen [Tongji Univ.].

A major challenge of crowd counting lies in the perspective distortion, which results in drastic person scale change in an image. Density regression on the small person area is in general very hard. In this work, we propose a perspective-aware convolutional neural network (PACNN) for efficient crowd counting, which integrates the perspective information into density regression to provide additional knowledge of the person scale change in an image. Ground truth perspective maps are firstly generated for training; PACNN is then specifically designed to predict multi-scale perspective maps, and encode them as perspective-aware weighting layers in the network to adaptively combine the outputs of multi-scale density maps. The weights are learned at every pixel of the maps such that the final density combination is robust to the perspective distortion. We conduct extensive experiments to demonstrate the effectiveness and efficiency of PACNN over the state-of-the-art [42].

6.1.14. Phone-Level Embeddings for Unit Selection Speech Synthesis

Participants: Laurent Amsaleg, Antoine Perquin [EXPRESSION team, IRISA], Gwénolé Lecorvé [EXPRESSION team, IRISA], Damien Lolive [EXPRESSION team, IRISA].

Deep neural networks have become the state of the art in speech synthesis. They have been used to directly predict signal parameters or provide unsupervised speech segment descriptions through embeddings. In [25] we present four models with two of them enabling us to extract phone-level embeddings for unit selection speech synthesis. Three of the models rely on a feed-forward DNN, the last one on an LSTM. The resulting embeddings enable replacing usual expert-based target costs by an euclidean distance in the embedding space. This work is conducted on a French corpus of an 11 hours audiobook. Perceptual tests show the produced speech is preferred over a unit selection method where the target cost is defined by an expert. They also show that the embeddings are general enough to be used for different speech styles without quality loss. Furthermore, objective measures and a perceptual test on statistical parametric speech synthesis show that our models perform comparably to state-of-the-art models for parametric signal generation, in spite of necessary simplifications, namely late time integration and information compression.

6.1.15. Disfluency Insertion for Spontaneous TTS: Formalization and Proof of Concept

Participants: Pascale Sébillot, Raheel Qader [EXPRESSION team, IRISA], Gwénolé Lecorvé [EXPRESSION team, IRISA], Damien Lolive [EXPRESSION team, IRISA].

This is an exploratory work to automatically insert disfluencies in text-to-speech (TTS) systems. The objective is to make TTS more spontaneous and expressive. To achieve this, we propose to focus on the linguistic level of speech through the insertion of pauses, repetitions and revisions. We formalize the problem as a theoretical process, where transformations are iteratively composed. This is a novel contribution since most of the previous work either focus on the detection or cleaning of linguistic disfluencies in speech transcripts, or solely concentrate on acoustic phenomena in TTS, especially pauses. We present a first implementation of the proposed process using conditional random fields and language models. The objective and perceptual evaluation conducted on an English corpus of spontaneous speech show that our proposition is effective to generate disfluencies, and highlights perspectives for future improvements [26]

6.1.16. Bi-directional Recurrent End-to-End Neural Network Classifier for Spoken Arab Digit Recognition

Participants: Christian Raymond, Naima Zerari [University of Batna 2, Algeria], Hassen Bouzgou [University of Batna 2, Algeria].

In [30], we propose a general end-to-end approach to sequence learning that uses Long Short-Term Memory (LSTM) to deal with the non-uniform sequence length of the speech utterances. The neural architecture can recognize the Arabic spoken digit spelling of an isolated Arabic word using a classification methodology, with the aim to enable natural human-machine interaction. The proposed system consists to, first, extract the relevant features from the input speech signal using Mel Frequency Cepstral Coefficients (MFCC) and then these features are processed by a deep neural network able to deal with the non uniformity of the sequences length. A recurrent LSTM or GRU architecture is used to encode sequences of MFCC features as a fixed size.

6.1.17. Are Deep Neural Networks good for blind image watermarking?

Participants: Teddy Furon, Vedran Vukotić [Lamark, France], Vivien Chappelier [Lamark, France].

Image watermarking is usually decomposed into three steps: i) some features are extracted from an image, ii) they are modified to embed the watermark, iii) and they are projected back into the image space while avoiding the creation of visual artefacts. The feature extraction is usually based on a classical image representation given by the Discrete Wavelet Transform or the Discrete Cosine Transform for instance. These transformations need a very accurate synchronisation and usually rely on various registration mechanisms for that purpose. This paper investigates a new family of transformation based on Deep Learning networks. Motivations come from the Computer Vision literature which has demonstrated the robustness of these features against light geometric distortions. Also, adversarial sample literature provides means to implement the inverse transform needed in the third step. This work [29] shows that this approach is feasible as it yields a good quality of the watermarked images and an intrinsic robustness.

6.2. Description and structuring

6.2.1. Automatic classification of radiological reports for clinical care

Participants: Anne-Lyse Minard, Alfonso Gerevini [Università degli Studi di Brescia, Italy], Alberto Lavelli [Fondazione Bruno Kessler, Italy], Alessandro Maffi [Università degli Studi di Brescia, Italy], Roberto Maroldi [Università degli Studi di Brescia, Italy, Azienda Socio Sanitaria Territoriale Spedali Civili di Brescia, Italy], Ivan Serina [Università degli Studi di Brescia, Italy], Guido Squassina [Azienda Socio Sanitaria Territoriale Spedali Civili di Brescia, Italy].

Radiological reporting generates a large amount of free-text clinical narratives, a potentially valuable source of information for improving clinical care and supporting research. The use of automatic techniques to analyze such reports is necessary to make their content effectively available to radiologists in an aggregated form. In this paper we focus on the classification of chest computed tomography reports according to a classification schema proposed for this task by radiologists of the Italian hospital ASST Spedali Civili di Brescia. The proposed system is built exploiting a training data set containing reports annotated by radiologists. Each report is classified according to the schema developed by radiologists and textual evidences are marked in the report. The annotations are then used to train different machine learning based classifiers. We present in this paper a method based on a cascade of classifiers which make use of a set of syntactic and semantic features. The resulting system is a novel hierarchical classification system for the given task, that we have experimentally evaluated [5].

6.2.2. Revisiting the medial axis for planar shape decomposition

Participants: Yannis Avrithis, N. Papanelopoulos [NTU Athens], S. Kollias [Univ. Lincoln].

We introduce a simple computational model for planar shape decomposition that naturally captures most of the rules and salience measures suggested by psychophysical studies, including the minima and short-cut rules, convexity, and symmetry [7]. It is based on a medial axis representation in ways that have not been explored before and sheds more light into the connection between existing rules like minima and convexity. In particular, vertices of the exterior medial axis directly provide the position and extent of negative minima of curvature, while a traversal of the interior medial axis directly provides a small set of candidate endpoints for part-cuts. The final selection follows a prioritized processing of candidate part-cuts according to a local convexity rule that can incorporate arbitrary salience measures. Neither global optimization nor differentiation is involved. We provide qualitative and quantitative evaluation and comparisons on ground-truth data from psychophysical experiments. With our single computational model, we outperform even an ensemble method on several other competing models.

6.2.3. Is ATIS too shallow to go deeper for benchmarking Spoken Language Understanding models?

Participants: Christian Raymond, Frédéric Béchet [Aix Marseille University].

We started a collaboration about benchmarking scientific benchmarks. We started in [11] by the ATIS (Air Travel Information Service) corpus, that will be soon celebrating its 30th birthday. Designed originally to benchmark spoken language systems, it still represents the most well-known corpus for benchmarking Spoken Language Understanding (SLU) systems. In 2010, in a paper titled "What is left to be understood in ATIS?", Tur et al. discussed the relevance of this corpus after more than 10 years of research on statistical models for performing SLU tasks. Nowadays, in the Deep Neural Network (DNN) era, ATIS is still used as the main benchmark corpus for evaluating all kinds of DNN models, leading to further improvements, although rather limited, in SLU accuracy compared to previous state-of-the-art models. We propose in this paper to investigate these results obtained on ATIS from a qualitative point of view rather than just a quantitative point of view and answer the two following questions: what kind of qualitative improvement brought DNN models to SLU on the ATIS corpus? Is there anything left, from a qualitative point of view, in the remaining 5% of errors made by current state-of-the-art models?

6.2.4. KRAUTS: A German Temporally Annotated News Corpus

Participants: Anne-Lyse Minard, Strötgen Jannik [Max Planck Institute for Informatics, Germany], Lukas Lange [Max Planck Institute for Informatics, Germany], Manuela Speranza [Fondazione Bruno Kessler, Italy], Bernardo Magnini [Fondazione Bruno Kessler, Italy].

In recent years, temporal tagging, i.e., the extraction and normalization of temporal expressions, has become a vibrant research area. Several tools have been made available, and new strategies have been developed. Due to domain-specific challenges, evaluations of new methods should be performed on diverse text types. Despite significant efforts towards multilinguality in the context of temporal tagging, for all languages except English, annotated corpora exist only for a single domain. In the case of German, for example, only a narrative style corpus has been manually annotated so far, thus no evaluations of German temporal tagging performance on news articles can be made. In this paper, we present KRAUTS, a new German temporally annotated corpus containing two subsets of news documents: articles from the daily newspaper DOLOMITEN and from the weekly newspaper DIE ZEIT. Overall, the corpus contains 192 documents with 1,140 annotated temporal expressions, and has been made publicly available to further boost research in temporal tagging citejannik:hal-01844834.

6.2.5. Active learning to assist annotation of aerial Images in environmental surveys

Participants: Ewa Kijak, Mathieu Laroze [OBELIX team, IRISA], Romain Dambreville [OBELIX team, IRISA], Chloe Friguet [OBELIX team, IRISA], Sébastien Lefèvre [OBELIX team, IRISA].

Remote sensing technologies greatly ease environmental assessment over large study areas using aerial images, e.g. for monitoring and counting animals or ships. Such data are most often analyzed by a manual operator, leading to costly and non scalable solutions. If object detection algorithms are used to fasten and automate the counting processes, these algorithms need to have prior ground truth available, which is a time-consuming and tedious process for field experts or engineers. We introduced a method to assist the annotation process in aerial images by introducing an active learning algorithm, allowing interaction with the expert such as class confirmation or correction at the labeling stage, and querying the expert with groups of samples taken from the same image to ease user annotation. Usual active learning algorithms perform instance selection from the whole set of input data. In this work, the selection of the queried instances is constrained by requiring that they belong to a group, (a part of) an image in our case, to ease the annotator task as the queried instances are proposed in their comprehensive context. We defined a score to rank the images and identify the one that should be annotated at each iteration, based on both uncertainty and true positives. The main objective is to reduce the number of human interactions on the overall process, starting from a first annotated image, rather than reaching the maximum final accuracy. Therefore, the annotation cost is measured through the gain in interactions (corrections of the classifier decisions by the annotator) with respect to a labeling task from scratch. At each iteration, the classifier is retrained according to a specific subset of data. Several strategies have been compared and their performances regarding the interaction gain have been discussed [19], [36].

6.3. Search, linking and navigation

6.3.1. Detecting fake news and tampered images in social networks

Participants: Cédric Maigrot, Ewa Kijak, Vincent Claveau.

Social networks make it possible to share information rapidly and massively. Yet, one of their major drawback comes from the absence of verification of the piece of information, especially with viral messages. This is the issue addressed by the participants to the Verification Multimedia Use task of Mediaeval 2016. They used several approaches and clues from different modalities (text, image, social information).

One promising approach is to examine if the image (if any) has been doctored. In recent work [23], we study context-aware methods to localize tamperings in images from social media. The problem is defined as a comparison between image pairs: an near-duplicate image retrieved from the network and a tampered version. We propose a method based on local features matching, followed by a kernel density estimation, that we compare to recent similar approaches. The proposed approaches are evaluated on two dedicated datasets containing a variety of representative tamperings in images from social media, with difficult examples. Context-aware methods are proven to be better than blind image forensics approach. However, the evaluation allows to analyze the strengths and weaknesses of the contextual-based methods on realistic datasets.

In further work [9], [22], we explore the interest of combining and merging these approaches in order to evaluate the predictive power of each modality and to make the most of their potential complementarity.

6.3.2. A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking

Participants: Christian Raymond, Guillaume Gravier, Vedran Vukotić.

With the recent resurgence of neural networks and the proliferation of massive amounts of unlabeled data, unsupervised learning algorithms became very popular for organizing and retrieving large video collections in a task defined as video hyperlinking. Information stored as videos typically contain two modalities, namely an audio and a visual one, that are used conjointly in multimodal systems by undergoing fusion. Multimodal autoencoders have been long used for performing multimodal fusion. In this work, we start by evaluating different initial, single-modal representations for automatic speech transcripts and for video keyframes. We progress to evaluating different autoencoding methods of performing multimodal fusion in an offline setup. The best performing setup is then evaluated in a live setup at TRECVID's 2016 video hyperlinking task. As in offline evaluations, we show that focusing on crossmodal translations as a way of performing multimodal fusion yields improved multimodal representations and that our simple system, trained in an unsupervised manner, with no external information information, defines the new state of the art in a live video hyperlinking setup. We conclude by performing an analysis on data gathered after the live evaluations at TRECVID 2016 and express our thoughts on the overall performance of our proposed system [8].

6.3.3. A study on multimodal video hyperlinking with visual aggregation

Participants: Mateusz Budnik, Mikail Demirdelen, Guillaume Gravier.

Video hyperlinking offers a way to explore a video collection, making use of links that connect segments having related content. Hyperlinking systems thus seek to automatically create links by connecting given anchor segments to relevant targets within the collection. In 2018, we pursued our long-term research effort towards multimodal representations of video segments in a hyperlinking system based on bidirectional deep neural networks, which achieved state-of-the-art results in the TRECVid 2016 evaluation. A systematic study of different input representations was done with a focus on the aggregation of the representation of multiple keyframes. This includes, in particular, the use of memory vectors as a novel aggregation technique, which provides a significant improvement over other aggregation methods on the final hyperlinking task. Additionally, the use of metadata was investigated leading to increased performance and lower computational requirements for the system [35].

6.3.4. Opinion mining in social networks

Participants: Anne-Lyse Minard, Christian Raymond, Vincent Claveau.

As part of the DeFT text-mining challenge, we participated in the elaboration of a task on fine-grained opinion mining in tweets [34] and to the analysis of the participants' results. We have also proposed systems [33] for each sub-task: (i) tweet classification according to the topic of the tweet, (ii) tweet classification according to their polarity, (iii) detection of the polarity markers and target of opinion in tweets. For the two first tasks, the approaches we proposed rely on a combination of boosting, decision trees and Recurrent Neural Networks. For the last task, we experimented with RNN coupled with a CRF layer. All of these systems performed very well and ranked in the best performing systems for each of the task.

6.3.5. Biomedical Information Extraction in social networks

Participants: Anne-Lyse Minard, Christian Raymond, Vincent Claveau.

This year, we participated in SMM4H challenge about extracting medical information from social networks. Faour tasks were proposed: (i) detection of posts mentioning a drug name, (ii) classification of posts describing medication intake, (iii) classification of adverse drug reaction mentioning posts, (iv) Automatic detection of posts mentioning vaccination behavior. In [24], we presented the systems developed by IRISA to participate to these four tasks. For these tweet classification tasks, we adopt a common approach based on recurrent neural networks (BiLSTM). Our main contributions are the use of certain features, the use of Bagging in order to deal with unbalanced datasets, and on the automatic selection of difficult examples. These techniques allow us to reach 91.4, 46.5, 47.8, 85.0 as F1-scores for Tasks 1 to 4, ranking us among the 3 first participants for each task.

6.3.6. Information Extraction in the biomedical domain

Participants: Clément Dalloux, Vincent Claveau, N. Grabar [STL-CNRS].

Automatic detection of negated content is often a pre-requisite in information extraction systems, especially in the biomedical domain. Following last year work, we propose two main contributions in this field [43]. We first introduced a new corpora built with excerpts from clinical trial protocols in French and Brazilian Portuguese, describing the inclusion criteria for patient recruitment. The corpora are manually annotated for marking up the negation cues and their scope. Secondly, two supervised learning approaches are been proposed for the automatic detection of negation. Besides, one of the approaches is validated on English data from the state of the art: the approach shows very good results and outperforms existing approaches, and it also yields comparable results on the French data.

We also have developed other data-sets (annotated corpora). Indeed, textual corpora are extremely important for various NLP applications as they provide information necessary for creating, setting and testing these applications and the corresponding tools. They are also crucial for designing reliable methods and reproducible results. Yet, in some areas, such as the medical area, due to confidentiality or to ethical reasons, it is complicated and even impossible to access textual data representative of those produced in these areas. We propose the CAS corpus [14] built with clinical cases, such as they are reported in the published scientific literature in French. We describe this corpus, containing over 397,000 word occurrences, and its current annotations (PoS, lemmas, negation, uncertainty).

As part of this work, we also developed software available as web-services on http://allgo.inria.fr (see the Software section).

6.3.7. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking

Participants: Yannis Avrithis, F. Radenovic [Univ. Prague], Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Ondra Chum [Univ. Prague].

In this work [27] we address issues with image retrieval benchmarking on standard and popular Oxford 5k and Paris 6k datasets. In particular, annotation errors, the size of the dataset, and the level of challenge are addressed: new annotation for both datasets is created with an extra attention to the reliability of the ground truth. Three new protocols of varying difficulty are introduced. The protocols allow fair comparison between different methods, including those using a dataset pre-processing stage. For each dataset, 15 new challenging queries are introduced. Finally, a new set of 1M hard, semi-automatically cleaned distractors is selected. An extensive comparison of the state-of-the-art methods is performed on the new benchmark. Different types of methods are evaluated, ranging from local-feature-based to modern CNN based methods. The best results are achieved by taking the best of the two worlds. Most importantly, image retrieval appears far from being solved.

6.3.8. Unsupervised object discovery for instance recognition

Participants: Oriane Siméoni, Yannis Avrithis, Ahmet Iscen [Univ. Prague], Giorgos Tolias [Univ. Prague], Ondra Chum [Univ. Prague].

Severe background clutter is challenging in many computer vision tasks, including large-scale image retrieval. Global descriptors, that are popular due to their memory and search efficiency, are especially prone to corruption by such a clutter. Eliminating the impact of the clutter on the image descriptor increases the chance of retrieving relevant images and prevents topic drift due to actually retrieving the clutter in the case of query expansion. In this work, we propose a novel salient region detection method. It captures, in an unsupervised manner, patterns that are both discriminative and common in the dataset. Saliency is based on a centrality measure of a nearest neighbor graph constructed from regional CNN representations of dataset images. The descriptors derived from the salient regions improve particular object retrieval, most noticeably in a large collections containing small objects [28].

MIMETIC Project-Team

7. New Results

7.1. Outline

In 2018, MimeTIC has maintained his activity in motion analysis, modelling and simulation. In motion analysis, we focused our efforts on two 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.

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 trade-off 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) developing new automatic methods for users' performance recognition and evaluation, especially online action detection, 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 defined Toric Space for camera placement to drone toric space to control a group of drones filming the action of an actor to ensure covering cinematographic distinct viewpoints. We also developed original VR-based staging and cinematography methods to make these processes be more interactive and immersive.

7.2. Motion analysis

7.2.1. Biomechanics for motion analysis-synthesis

Participants: Charles Pontonnier [contact], Georges Dumont, Franck Multon, Antoine Muller, Pierre Puchaud.

Based on a former PhD thesis (of Antoine Muller), we aim at democratizing the use of musculoskeletal analysis for a wide range of users. We proposed contributions enabling better performances of such analyses and preserving accuracy, as well as contributions enabling an easy subject-specific model calibration [47], [48]. In order to control the whole analysis process, we propose a global approach of all the analysis steps: kinematics, dynamics and muscle forces estimation. For all of these steps, quick analysis methods have been proposed. Particularly, a quick muscle force sharing problem resolution method [26] has been proposed, based on interpolated data and improvements have been proposed [25]. Moreover, the Music Toolbox is now proposed as an opensource software.

The determination of maximal torque envelopes method that we defined for the elbow torque analysis have been used for the shoulder [44]. It is important, in order to calibrate muscular models, to be able to identify force parameters in a musculoskeletal.

7.2.2. Interactions between walkers

Participants: Anne-Hélène Olivier [contact], Armel Crétual, Richard Kulpa, Sean Lynch.

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. In the context of Sean Lynch's PhD, which was defended in October 2018 [12], we aimed at manipulating the nature of the visual information available to the participants to understand which information about the other walker are important to avoid a collision. We presented at IEEE VR 2018, our work on the influence of global and local information appearances [46] as well as on the influence of mutual gaze in the interaction [39].

In the context of transportation research, we developed a new collaboration with Ifsttar (LEPSIS, LESCOT) involving questions about interaction between pedestrians on a narrow sidewalk [50], [42].

We also provide lot of efforts to investigate, in collaboration with Julien Pettré from Inria Rainbow team, the process involved in the selection of interactions within our neighbourhood. Considering the complex case of multiple interactions, we first 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. Results showed that pedestrians are able to interact with two other walkers simultaneously, rather than treating each interaction in sequence. These results are currently in press in Frontiers in Psychology [22]. Second, we performed experiments involving 40 people to understand how collective behaviour emerges [31]. Third, in virtual conditions, we also coupled the analysis of gaze behaviour and the trajectory and showed that human gaze, during navigation, is attracted by other walkers presenting the higher risk of future collision [21], [32].

Finally, we continue working on the applications of studying human behaviour for application in human-moving robot interactions. 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. In collaboration with Philippe Souères and Christian Vassallo (LAAS, Toulouse), our work analyzed the behavior of human walkers crossing the trajectory of a mobile robot that was programmed to reproduce this human avoidance strategy. In contrast with a previous study, which showed that humans mostly prefer to give the way to a non-reactive robot, we observed similar behaviors between human-human avoidance and human-robot avoidance when the robot replicates the human interaction rules. This result highlight the importance of controlling robots in a human-like way in order to ease their cohabitation with humans [28]. In collaboration with Jose Grimaldo da Silva and Thierry Fraichard (Inria Grenoble), we designed a shared-effort model during interaction between a moving robot and a human relying on walker-walker collision avoidance data [34].

7.2.3. Biomechanical analysis of tennis serve

Participants: Richard Kulpa [contact], Benoit Bideau, Pierre Touzard.

In the context of the exclusive collaboration with the FFT (French Tennis Federation), we made new experiments on top-level young French players (between 12 up to 18 years old) to quantify the relation between motor technical errors and their impact on the risk of injury. We thus concurrently captured the kinematics of their tennis serve and the muscular activities of the upper-body. We recently validated that the Waiter's serve implies higher risk of injuries [27]. It is a movement that was know by the coaches as not productive and risky but it was never validated. Moreover, we evaluated the strategies of pacing use during five-set matches in the top tennis tournaments [20].

7.3. Virtual human simulation

7.3.1. Novel Distance Geometry based approaches for Human Motion Retargeting

Participants: Franck Multon [contact], Ludovic Hoyet, Antonio Mucherino, Zhiguang Liu.

Since September 2016, Antonio Mucherino has a half-time Inria detachment in the MimeTIC team (ended Sept 2018), 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 [24], and we co-supervised Antonin Bernardin for his Master thesis in 2017, 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 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.

Concurrently, we proposed a pose transfer algorithm from a source character to a target character, without using skeleton information. Previous work mainly focused on retargeting skeleton animations whereas the contextual meaning of the motion is mainly linked to the relation- ship between body surfaces, such as the contact of the palm with the belly. In the context of the Inria PRE program, we propose a new context-aware motion retargeting framework [38], based on deforming a target character to mimic a source character poses using harmonic mapping. We also introduce the idea of Context Graph: modeling local interactions between surfaces of the source character, to be preserved in the target character, in order to ensure fidelity of the pose. In this approach, no rigging is required as we directly manipulate the surfaces, which makes the process totally automatic. Our results demonstrate the relevance of this automatic rigging-less approach on motions with complex contacts and interactions between the character's surface.

7.3.2. Investigating the Impact of Training for Example-Based Facial Blendshape Creation.

Participant: Ludovic Hoyet [contact].

In collaboration with Technicolor and Trinity College Dublin, we explored how certain training poses can influence the Example-Based Facial Rigging (EBFR) method [33]. We analysed the output of EBFR given a set of training poses to see how well the results reproduced our ground truth actor scans compared to a pure Deformation Transfer approach (Figure 5). We found that, while the EBFR results better matched the ground truth overall, there were certain cases that didn't see any improvement. While some of these results may be explained by lack of sufficient training poses for the area of the face in question, we found that certain lip poses weren't improved by training despite a large number of mouth training poses supplied. Our initial goal for this project was to identify what facial expressions are important to use as training when using Example-Based Facial Rigging to create facial rigs. This preliminary work has indicated certain parts of the face that might require more attention when automatically creating blendshapes, which still require to be further investigated, e.g., to identify a subset of facial expressions that would be considered the "ideal" subset to use for training the EBFR algorithm.

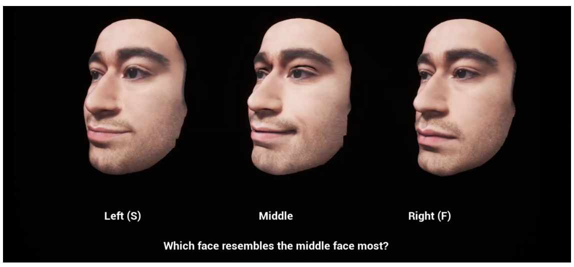


Figure 5. An example of stimuli shown to participants comparing the facial rigs created without training, and those with training.

7.4. Human motion in VR

7.4.1. Motion recognition and classification

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

Being able to interactively detect and recognize actions based on skeleton data, in unsegmented streams, has become an important computer vision topic. It raises three scientific problems in relation with variability. The first one is the temporal variability that occurs when subjects perform gestures with different speeds. The second one is the inter-class spatial variability, which refers to disparities between the displacement amounts induced by different classes (i.e. long vs. short movements). The last one is the intra-class spatial variability caused by differences in style and gesture amplitude. Hence, we designed an original approach that better considers these three issues [15]. To address temporal variability we introduce the notion of curvilinear segmentation. It consists in extracting features, not on temporally-based sliding windows, but on segments in which the accumulated curvilinear displacement of skeleton joints equals a specific amount. Second, to tackle inter-class spatial variability, we define several competing classifiers with their dedicated curvilinear windows. Last, we address intraclass spatial variability by designing a fusion system that takes the decisions and confidence scores of every competing classifier into account. Extensive experiments on four challenging skeleton-based datasets demonstrate the relevance and efficiency of the proposed approach.

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

Participant: Richard Kulpa [contact].

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 expert's 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 [23]. We applied our method on tennis serve and karate katas.

7.4.3. Studying the Sense of Embodiment in VR Shared Experiences

Participants: Rebecca Fribourg, Ludovic Hoyet [contact].

To explore how the sense of embodiment is influenced by the fact of sharing a virtual environment with another user, we conducted an experiment where users were immersed in a virtual environment while being embodied in an anthropomorphic virtual representation of themselves [36], in collaboration with Hybrid Inria team. In particular, two situations were studied: either users were immersed alone, or in the company of another user (see Figure 6). During the experiment, participants performed a virtual version of the well-known whac-a-mole game, therefore interacting with the virtual environment, while sitting at a virtual table. Our results show that users were significantly more “efficient” (i.e., faster reaction times), and accordingly more engaged, in performing the task when sharing the virtual environment, in particular for the more competitive tasks. Also, users experienced comparable levels of embodiment both when immersed alone or with another user. These results are supported by subjective questionnaires but also through behavioural responses, e.g. users reacting to the introduction of a threat towards their virtual body. Taken together, our results show that competition and shared experiences involving an avatar do not influence the sense of embodiment, but can increase user engagement. Such insights can be used by designers of virtual environments and virtual reality applications to develop more engaging applications.

Figure 6. Setup of the experiment: each user was able to interact in the virtual environment with his own avatar, while the physical setup provided both a reference frame and passive haptic feedback. From left to right: experimental conditions Alone, Mirror and Shared; Physical setup of the experiment.

7.4.4. Biofidelity in VR

Participants: Simon Hilt, Charles Pontonnier, Georges Dumont [contact].

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 impact of uncertainties on friction coefficients on haptic feedback. The coefficients 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. This year, we conducted experiments aiming at comparing gesture, recorded by an optoelectronic setup, and muscular activities, recorded by EMG, between real and virtual (with haptic feedback) manipulation.

7.5. Digital storytelling

7.5.1. Film Editing Patterns: Thinking like a Director

Participant: Marc Christie [contact].

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 [29] 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.

7.5.2. Directing Cinematographic Drones

Participants: Marc Christie [contact], Quentin Galvane.

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 aesthetic camera motions. The work was published in the Transactions on Graphics journal and was accepted for presentation at SIGGRAPH [18].

In addition we have focused on full automated and non-reactive path-planning for cinematographic drones. Most existing tools typically require the user to specify and edit the camera path, for example by providing a complete and ordered sequence of key viewpoints. In our contribution, we propose a higher level tool designed to enable even novice users to easily capture compelling aerial videos of large-scale outdoor scenes. Using a coarse 2.5D model of a scene, the user is only expected to specify starting and ending viewpoints and designate a set of landmarks, with or without a particular order. Our system automatically generates a diverse set of candidate local camera moves for observing each landmark, which are collision-free, smooth, and adapted to the shape of the landmark. These moves are guided by a landmark-centric view quality field, which combines visual interest and frame composition. An optimal global camera trajectory is then constructed that chains together a sequence of local camera moves, by choosing one move for each landmark and connecting them with suitable transition trajectories. This task is formulated and solved as an instance of the Set Traveling Salesman Problem. The work was published and presented at SIGGRAPH [30].

7.5.3. Automated Virtual Staging

Participants: Marc Christie [contact], Quentin Galvane, Fabrice Lamarche, Amaury Louarn.

While the topic of virtual cinematography has essentially focused on the problem of computing the best viewpoint in a virtual environment given a number of objects placed beforehand, the question of how to place the objects in the environment with relation to the camera (referred to as staging in the film industry) has received little attention. This work first proposes a staging language for both characters and cameras that extends existing cinematography languages with multiple cameras and character staging. Second, we propose techniques to operationalize and solve staging specifications given a 3D virtual environment. The novelty holds in the idea of exploring how to position the characters and the cameras simultaneously while maintaining a number of spatial relationships specific to cinematography. We demonstrate the relevance of our approach through a number of simple and complex examples [45].

7.5.4. VR Staging and Cinematography

Participants: Marc Christie [contact], Quentin Galvane.

Creatives in animation and film productions have forever been exploring the use of new means to prototype their visual sequences before realizing them, by relying on hand-drawn storyboards, physical mockups or more recently 3D modelling and animation tools. However these 3D tools are designed in mind for dedicated animators rather than creatives such as film directors or directors of photography and remain complex to control and master. In this work we propose a VR authoring system which provides intuitive ways of crafting visual sequences, both for expert animators and expert creatives in the animation and film industry. The proposed system is designed to reflect the traditional process through (i) a storyboarding mode that enables rapid creation of annotated still images, (ii) a previsualisation mode that enables the animation of the characters, objects and cameras, and (iii) a technical mode that enables the placement and animation of complex camera rigs (such as cameras cranes) and light rigs. Our methodology strongly relies on the benefits of VR manipulations to re-think how content creation can be performed in this specific context, typically how to animate contents in space and time. As a result, the proposed system is complimentary to existing tools, and provides a seamless back-and-forth process between all stages of previsualisation. We evaluated the tool with professional users to gather experts' perspectives on the specific benefits of VR in 3D content creation [37].

7.5.5. Improving Camera tracking technologies

Participants: Marc Christie [contact], Xi Wang.

Robustness of indirect SLAM techniques to light changing conditions remains a central issue in the robotics community. With the change in the illumination of a scene, feature points are either not extracted properly due to low contrasts, or not matched due to large differences in descriptors. In this work, we propose a multi-layered image representation (MLI) in which each layer holds a contrast enhanced version of the current image in the tracking process in order to improve detection and matching. We show how Mutual Information can be used to compute dynamic contrast enhancements on each layer. We demonstrate how this approach dramatically improves the robustness in dynamic light changing conditions on both synthetic and real environments compared to default ORB-SLAM. This work focalises on the specific case of SLAM relocalisation in which a first pass on a reference video constructs a map, and a second pass with a light changed condition relocalizes the camera in the map [41], [40].

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's activity ranges from theoretical results to algorithmic design and software contributions in the fields of sparse representations, inverse problems, and dimension reduction.

7.1.1. Computational Representation Learning: Algorithms and Theory

Participants: Rémi Gribonval, Hakim Hadj Djilani, Cássio Fraga Dantas, Jeremy Cohen.

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.

Multilayer sparse matrix products for faster computations. 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, with reduced sample complexity. Besides the principle and its application to image denoising [105], we demonstrated the potential of the approach to speedup linear inverse problems [104], and a comprehensive journal paper was published in 2016 [107]. Pioneering identifiability results have been obtained in the Ph.D. thesis of Luc Le Magoarou [108].

We further explored the application of this technique to obtain fast approximations of Graph Fourier Transforms [106], and studied their approximation error [109]. In a journal paper published this year [16] 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 FAμST software library (see Section 6) was first released as Matlab code primarily for reproducibility of the experiments of [107]. A C++ version is being developed to provide transparent interfaces of FAμST data-structures with both Matlab and Python.

Kronecker product structure for faster computations. In parallel to the development of FAuST, we have proposed another approach to structured dictionary learning that also aims at speeding up both sparse coding and dictionary learning. We used the fact that for tensor data, a natural set of linear operators are those that operate on each dimension separately, which correspond to rank-one multilinear operators. These rank-one operators may be cast as the Kronecker product of several small matrices. Such operators require less memory and are computationally attractive, in particular for performing efficient matrix-matrix and matrix-vector operations. In our proposed approach, dictionaries are constrained to belong to the set of low-rank multilinear operators, that consist of the sum of a few rank-one operators. A special case of the proposed structure is the widespread separable dictionary, named SuKro, which was evaluated experimentally last year on an image denoising application [81]. The general approach, coined HOSUKRO for High Order Sum of Kronecker products, has been shown this year to reduce empirically the sample complexity of dictionary learning, as well as theoretical complexity of both the learning and the sparse coding operations [27].

Combining faster matrix-vector products with screening techniques. We combined accelerated matrix-vector multiplications offered by FAμST / HOSUKRO matrix approximations with dynamic screening [57], that safely eliminates inactive variables to speedup iterative convex sparse recovery algorithms. First, we showed how to obtain safe screening rules for the exact problem while manipulating an approximate dictionary [80]. We then adapted an existing screening rule to this new framework and define a general procedure to leverage the advantages of both strategies. This year we completed a comprehensive preprint submitted for publication in a journal [49] that includes new techniques based on duality gaps to optimally switch from a coarse dictionary approximation to a finer one. Significant complexity reductions were obtained in comparison to screening rules alone [28].

7.1.2. 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 [76], 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 [77], 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. This year, a condensed version of these results has been prepared which is now accepted for publication [14].

7.1.3. Algorithmic exploration of large-scale Compressive Learning via Sketching

Participants: Rémi Gribonval, 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 Keriven (ENS Paris), Nicolas Tremblay (GIPSA-Lab, Grenoble), Phil Schniter & Evan Byrne (Ohio State University, USA), Laurent Jacques & Vincent Schellekens (Univ Louvain, Belgium), Florimond Houssiau & Y.-A. de Montjoye (Imperial College London, UK)

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 [58], [61], [59], [60] to fit a mixture of isotropic Gaussians to data vectors by computing a low-dimensional sketch of the data. The sketch represents empirical generalized 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 Ph.D. thesis of Nicolas Keriven [97] consolidated extensions to non-isotropic Gaussians, with a new algorithm called CL-OMP [96] and large-scale experiments demonstrating its potential for speaker verification

[95]. A journal paper was published this year [15], with an associated toolbox for reproducible research (see SketchMLBox, Section 6).

Sketching for Compressive Clustering and beyond. In 2016 we started a new endeavor to extend the sketched learning approach beyond Gaussian Mixture Estimation.

First, we showed empirically that sketching can be adapted to compress a training collection while allowing large-scale *clustering*. The approach, called "Compressive K-means", uses CL-OMP at the learning stage [98]. This year, we showed that in the high-dimensional setting one can substantially speedup both the sketching stage and the learning stage by replacing Gaussian random matrices with fast random linear transforms in the sketching procedure [23].

An alternative to CL-OMP for cluster recovery from a sketch is 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 [62]. During his first year of Ph.D., Antoine Chatalic visited the group of Phil Schiter to further investigate this topic, and a journal paper is in preparation.

We also demonstrated that sketching can be used in blind source localization and separation, by learning mixtures of alpha-stable distributions [32], see details in Section 7.5.3 .

Finally, sketching provides a potentially privacy-preserving data analysis tool, since the sketch does not explicitly disclose information about individual datum. A conference paper establishing theoretical privacy guarantees (with the *differential privacy* framework) and exploring the utility / privacy tradeoffs of Compressive K-means has been submitted for publication.

7.1.4. Theoretical results on Low-dimensional Representations, Inverse problems, and Dimension Reduction

Participants: Rémi Gribonval, Clément Elvira.

Main collaboration: Mike Davies (University of Edinburgh, UK), Gilles Puy (Technicolor R&I France, Rennes), Yann Traonmilin (Institut Mathématique de Bordeaux), Nicolas Keriven (ENS Paris), Gilles Blanchard (Univ Postdam, Germany), Cédric Herzet (SIMSMART project-team, IRMAR / Inria Rennes), Charles Soussen (Centrale Supelec, Gif-sur-Yvette), Mila Nikolova (CMLA, Cachan)

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). We published a comprehensive paper [20] 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 [121]. These results have been surveyed in a book chapter published this year [46]. In the context of nonlinear inverse problems, we showed that the notion of RIP is still relevant with proper adaptation [42].

Optimal convex regularizers for linear inverse problems. The ℓ^1-norm is a good convex regularization for the recovery of sparse vectors from under-determined linear measurements. No other convex regularization seems to surpass its sparse recovery performance. We explored possible explanations for this phenomenon by defining several notions of " best " (convex) regularization in the context of general low-dimensional recovery and showed that indeed the ℓ^1-norm is an optimal convex sparse regularization within this framework [43]. A journal paper is in preparation with extensions concerning nuclear norm regularization for low-rank matrix recovery and further structured low-dimensional models.

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 is under revision for a journal [88].

Recovery guarantees for algorithms with continuous dictionaries. We established theoretical guarantees on sparse recovery guarantees for a greedy algorithm, orthogonal matching pursuit (OMP), in the context of continuous dictionaries [40], e.g. as appearing in the context of sparse spike deconvolution. Analyses based on discretized dictionary fail to be conclusive when the discretization step tends to zero, as the coherence goes to one. Instead, our analysis is directly conducted in the continuous setting amd exploits specific properties of the positive definite kernel between atom parameters defined by the inner product between the corresponding atoms. For the Laplacian kernel in dimension one, we showed in the noise-free setting that OMP exactly recovers the atom parameters as well as their amplitudes, regardless of the number of distinct atoms [40]. A journal paper is in preparation describing a full class of kernels for which such an analysis holds, in particular for higher dimensional parameters.

On Bayesian estimation and proximity operators. There are two major routes to address the ubiquitous family of inverse problems appearing in signal and image processing, such as denoising or deblurring. The first route is Bayesian modeling: prior probabilities are used to model both the distribution of the unknown variables and their statistical dependence with the observed data, and estimation is expressed as the minimization of an expected loss (e.g. minimum mean squared error, or MMSE). The other route is the variational approach, popularized with sparse regularization and compressive sensing. It consists in designing (often convex) optimization problems involving the sum of a data fidelity term and a penalty term promoting certain types of unknowns (e.g., sparsity, promoted through an L1 norm).

Well known relations between these two approaches have lead to some widely spread misconceptions. In particular, while the so-called Maximum A Posterori (MAP) estimate with a Gaussian noise model does lead to an optimization problem with a quadratic data-fidelity term, we disprove through explicit examples the common belief that the converse would be true. In previous work we showed that for denoising in the presence of additive Gaussian noise, for any prior probability on the unknowns, the MMSE is the solution of a penalized least squares problem, with all the apparent characteristics of a MAP estimation problem with Gaussian noise and a (generally) different prior on the unknowns [89]. In other words, the variational approach is rich enough to build any MMSE estimator associated to additive Gaussian noise via a well chosen penalty.

This year, we achieved generalizations of these results beyond Gaussian denoising and characterized noise models for which the same phenomenon occurs. In particular, we proved that with (a variant of) Poisson noise and any prior probability on the unknowns, MMSE estimation can again be expressed as the solution of a penalized least squares optimization problem. For additive scalar denoising, the phenomenon holds if and only if the noise distribution is log-concave, resulting in the perhaps surprising fact that scalar Laplacian denoising can be expressed as the solution of a penalized least squares problem. [51] Somewhere in the proofs appears an apparently new characterization of proximity operators of (nonconvex) penalties as subdifferentials of convex potentials [50].

7.1.5. Algorithmic Exploration of Sparse Representations for Neurofeedback

Participant: Rémi Gribonval.

Claire Cury, Pierre Maurel & Christian Barillot (VISAGES Inria project-team, Rennes)

In the context of the HEMISFER (Hybrid Eeg-MrI and Simultaneous neuro-feedback for brain Rehabilitation) Comin Labs project (see Section 9.1.1.1), 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 [118]. This year we focused on directly estimating neuro-feedback scores rather than brain activity. Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) both allow measurement of brain activity for neuro-feedback (NF), respectively with high temporal resolution for EEG and high spatial resolution for fMRI. Using simultaneously fMRI and EEG for NF training is very promising to devise brain rehabilitation protocols, however performing NF-fMRI is costly, exhausting and time consuming, and cannot be repeated too many times for the same subject. We proposed a technique to predict NF scores from EEG recordings only, using a training phase where both EEG and fMRI NF are available. A conference paper has been submitted.

7.1.6. Sparse Representations as Features for Heart Sounds Classification

Participant: Nancy Bertin.

Main collaborations: Roilhi Frajo Ibarra Hernandez, Miguel Alonso Arevalo (CICESE, Ensenada, Mexico)

A heart sound signal or phonocardiogram (PCG) is the most simple, economical and non-invasive tool to detect cardiovascular diseases (CVD), the main cause of death worldwide. During the visit of Roilhi Ibarra, we proposed a pipeline and benchmark for binary heart sounds classification, based on his previous work on a sparse decomposition of the PCG [91]. We improved the feature extraction architecture, by combining features derived from the Gabor atoms selected at the sparse representation stage, with Linear Predictive Coding coefficients of the residual. We compared seven classifiers with two different approaches in presence of multiple hearts beats in the recordings: feature averaging (proposed by us) and cycle averaging (state-of-the-art). The feature sets were also tested when using an oversampling method for balancing. The benchmark identified systems showing a satisfying performance in terms of accuracy, sensitivity, and Matthews correlation coefficient, with best results achieved when using the new feature averaging strategy together with oversampling. This work was accepted for publication in an international conference [30].

7.1.7. An Alternative Framework for Sparse Representations: Sparse “Analysis” Models

Participants: Rémi Gribonval, Nancy Bertin, Clément Gaultier.

Main collaborations: Srdan Kitic (Orange, 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 [87], [117] [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 [99], [101], [100], [55]. 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], a book chapter 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, was published this year [45]. Second, we continued extending the cosparse framework on audio restoration problems [85], [84], [82], especially improvements on our released real-time declipping algorithm (A-SPADE - see Section 6.2) and extension to multichannel data [29].

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. 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 [63], defended in October 2016 and co-supervised by R. Gribonval, and awarded with the GDR ISIS/GRETSI/EEA thesis prize, see Section 5.1.1), 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, [67], [65], [66]. We characterized waveforms with optimum PAPR [68], [64] as well as the tradeoffs between PAPR and Power Spectral Density properties of a wavelet modulation scheme [70]. Our design of new adaptive multi-carrier waveform systems able to cope with frequency-selective channels while minimizing PAPR gave rise to a patent [69] and was published this year [22], [13].

7.3. Emerging activities on high-dimensional learning with neural networks

Participants: Rémi Gribonval, Himalaya Jain, Pierre Stock.

Main collaborations: Patrick Perez (Technicolor R & I, Rennes), Gitta Kutyniok (TU Berlin, Germany), Morten Nielsen (Aalborg University, Denmark), Felix Voigtlaender (KU Eichstätt, Germany), Herve Jegou and Benjamin Graham (FAIR, Paris)

dictionary learning, large-scale indexing, sparse deep networks, normalization, sinkhorn, regularization

Many of the data analysis and processing pipelines that have been carefully engineered by generations of mathematicians and practitioners can in fact be implemented as deep networks. Allowing the parameters of these networks to be automatically trained (or even randomized) allows to revisit certain classical constructions. Our team has started investigating the potential of such approaches both from an empirical perspective and from the point of view of approximation theory.

Learning compact representations for large-scale image search. The PhD thesis of Himalaya Jain [11] was dedicated to learning techniques for the design of new efficient methods for large-scale image search and indexing. A first step was to propose techniques for approximate nearest neighbor search exploiting quantized sparse representations in learned dictionaries [92]. The thesis then explored structured binary codes, computed through supervised learning with convolutional neural networks [93]. This year, we integrated these two components in a unified end-to-end learning framework where both the representation and the index are learnt [31]. These results have led to a patent application.

Equi-normalization of Neural Networks. Modern neural networks are over-parameterized. In particular, each rectified linear hidden unit can be modified by a multiplicative factor by adjusting input and out- put weights, without changing the rest of the network. Inspired by the Sinkhorn-Knopp algorithm, we introduced a fast iterative method for minimizing the l2 norm of the weights, equivalently the weight decay regularizer. It provably converges to a unique solution. Interleaving our algorithm with SGD during training improves the test accuracy. For small batches, our approach offers an alternative to batch- and group- normalization on CIFAR-10 and ImageNet with a ResNet-18. This work has been submitted for publication.

Approximation theory with deep networks. We study the expressivity of sparsely connected deep networks. Measuring a network's complexity by its number of connections with nonzero weights, or its number of neurons, we consider the class of functions which error of best approximation with networks of a given complexity decays at a certain rate. Using classical approximation theory, we showed that this class can be endowed with a norm that makes it a nice function space, called approximation space. We established that the presence of certain "skip connections" has no impact of the approximation space, and studied the role of the network's nonlinearity (also known as activation function) on the resulting spaces, as well as the benefits of depth. For the popular ReLU nonlinearity (as well as its powers), we related the newly identified spaces to classical Besov spaces, which have a long history as image models associated to sparse wavelet decompositions. The sharp embeddings that we established highlight how depth enables sparsely connected networks to approximate functions of increased "roughness" (decreased Besov smoothness) compared to shallow networks and wavelets. A journal paper is in preparation.

7.4. Emerging activities on Nonlinear Inverse Problems

Compressive sensing, compressive learning, audio inpainting, phase estimation

7.4.1. 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 [73], 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 [71], head pose estimation in images [79], sound source separation and localization [72], and virtually-supervised acoustic space learning (see Section 7.6.1). 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.

7.4.2. Audio Inpainting and Denoising

Participants: Rémi Gribonval, Nancy Bertin, Clément Gaultier.

Main collaborations: Srdan Kitic (Orange, 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 [54]), we proposed over the last three years a series of algorithms leveraging the competitive cosparse approach, which offers a very appealing trade-off between reconstruction performance and computational time [100], [102] [6]. The work on cosparse audio declipping which was awarded the Conexant best paper award at the LVA/ICA 2015 conference [102] resulted in a software release in 2016. In 2017, this work was extended towards advanced (co)sparse decompositions, including several forms of structured sparsityand towards their application to the denoising task.In particular, we investigated the incorporation of the so-called "social" structure constraint [103] into problems regularized by a cosparse prior [84], [85], and exhibited a common framework allowing to tackle both denoising and declipping in a unified fashion [82].

In 2018, a new algorithm for joint declipping of multichannel audio was derived and published [29]. Extensive experimental benchmarks were conducted, questioning the previous state-of-the-art habits in degradation levels (usually moderate to inaudible) and evaluation (small datasets, SNR-based performance criteria) and setting up new standards for the task (large and diverse datasets, severe saturation, perceptual quality evaluation) as well as guidelines for the choice of the best variant (sparse or cosparse, with or without structural time-frequency constraints...) depending on the data and operational conditions. These new results will be included in an ongoing journal paper, to be submitted in 2019.

7.5. 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 [122] 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 Sections 7.5.1 , 7.5.2); more exploratory work towards new approaches diverging away from local Gaussian modeling (Section 7.5.3) ; formulating and addressing a larger class of problems related to localization and separation, in the contexts of robotics (Section 7.5.4) and virtual reality (Section 7.5.2). Eventually, one of these new tracks, audio scene analysis with machine learning, evolved beyond the "localization and separation" paradigm, and is the subject of a new axis of research presented in Section 7.6 .

7.5.1. Towards Real-world Localization and Separation

Participants: Nancy Bertin, Frédéric Bimbot, Rémi Gribonval, Ewen Camberlein, Romain Lebarbenchon, Mohammed Hafsati.

Main collaborations: Emmanuel Vincent (MULTISPEECH Inria project-team, Nancy)

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, submitted to a special issue of Speech Communication, was published this year [12].

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). In addition, evolutions of the MBSSLocate software initiated during this project led to a successful participation in the IEEE-AASP Challenge on Acoustic Source Localization and Tracking (LOCATA), and to industrial transfer (see Section 8.1.1).

7.5.2. Separation for Remixing Applications

Participants: Nancy Bertin, Rémi Gribonval, Mohammed Hafsati.

Main collaborations: Nicolas Epain (IRT b<>com, Rennes)

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 [86], 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. Simulation results showed that separating sources in the HOA domain results in a 5 to 15 dB increase in signal-to-distortion ratio, compared to the microphone domain. These results led to a conference paper submission in 2018. We continued extending our methods to hybrid acquisition scenarios, where the separation of HOA signals can be informed by complementary close-up microphonic signals. Future work will include subjective evaluation of the developed workflows.

7.5.3. Beyond the Local Complex Gaussian Model

Participant: Antoine Deleforge.

Main collaboration: Nicolas Keriven (ENS Paris), 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 [110]. We provided a theoretical analysis of some limitations of the classical LCG-based multichannel Wiener filter in [21]. In [32] 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 [120][9] or more recently deep neural networks [119], it becomes desirable to use probabilistic models enforcing strong magnitude priors. In [75], we explored deterministic magnitude models. An approximate and tractable probabilistic version of this referred to as BEADS (Bayesian Expansion Approximating the Donut Shape) was presented this year [33]. The source prior considered is a mixture of isotropic Gaussians regularly placed on a zero-centered complex circle.

7.5.4. Applications to Robot Audition

Participants: Nancy Bertin, Antoine Deleforge.

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. Results obtained in the previous years [116], [114], [115] were encompassed and extended in two journal papers published this year [17], [18]. In particular, we obtained new results on the use of interaural level difference as the only input feature of the servo, a counter-intuitive result outside the robotic context. We also showed the robustness, low-complexity and independence to Head Related Transfer Function (HRTF) of the approach on humanoid robots.

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). The dataset was made publicly available at dregon.inria.fr and was presented together with the results in [37].

7.6. Towards comprehensive audio scene analysis

Source localization and separation, machine learning, room geometry, room properties, multichannel audio classification

By contrast to the previous lines of work and results on source localization and separation, which are mostly focused on the *sources*, the following emerging activities consider the audio scene and its analysis in a wider sense, including the environment around the sources, and in particular the *room* they are included in, and their properties. This inclusive vision of the audio scene allows in return to revisit classical audio processing tasks, such as localization, separation or classification.

7.6.1. Virtually-Supervised Auditory Scene Analysis

Participants: Antoine Deleforge, Nancy Bertin, Diego Di Carlo, Clément Gaultier, Rémi Gribonval.

Main collaborations: Ivan Dokmanic (University of Illinois at Urbana-Champaign, Coordinated Science Lab, USA), 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.

Recently, two directions for advanced audio geometry estimation have emerged and were investigated in our team. The first one is physics-driven [45]. This approach implicitly 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. However, it relies 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 [72], [74]. 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) [83], [94]. 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 [94]. 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 [83]. The VAST datasets and approaches made the bed of several new works in 2018, including real-world source localization on a wider range of settings (LOCATA test data on various microphone arrays) and echo estimation (see below).

7.6.2. Room Properties: Estimating or Learning Early Echoes

Participants: Antoine Deleforge, Nancy Bertin, Diego Di Carlo.

Main collaborations: Ivan Dokmanic (University of Illinois at Urbana-Champaign, Coordinated Science Lab, USA), Robin Scheibler (Tokyo Metropolitan University, Tokyo, Japan), Helena Peic-Tukuljac (EPFL, Switzerland).

In [35] we showed that the knowledge of early echoes improved sound source separation performances, which motivates the development of (blind) echo estimation techniques. Echoes are also known to potentially be a key to the room geometry problem [78]. In 2018, two different approaches to this problem were explored.

In [34] we proposed an analytical method for early echoes estimation. This method builds on the framework of finite-rate-of-innovation sampling. The approach operates directly in the parameter-space of echo locations and weights, and enables near-exact blind and off-grid echo retrieval from discrete-time measurements. It is shown to outperform conventional methods by several orders of magnitude in precision, in an ideal case where the room impulse response is limited to a few weighted Diracs. Future work will include alternative initialization schemes and convex relaxations, extensions to sparse-spectrum signals and noisy measurements, and applications to dereverberation and audio-based room shape reconstruction.

As a concurrent approach exploration, the PhD thesis of Diego Di Carlo aims at applying the VAST framework to the blind estimation of acoustic echoes, or other room properties (such as reverberation time, acoustic properties at the boundaries, etc.) This year, we focused on identifying promising couples of inputs and outputs for such an approach, especially by leveraging the notions of relative transfer functions between microphones, the room impulse responses, the time-difference-of-arrivals, the angular spectra, and all their mutual relationships. In a simple yet common scenario of 2 microphones close to a reflective surface and one source (which may occur, for instance, when the sensors are placed on a table such as in voice-based assistant devices), we introduced the concept of microphone array augmentation with echoes (MIRAGE) and showed how estimation of early-echo characteristics with a learning-based approach is not only possible but can in fact benefit source localization. In particular, it allows to retrieve 2D direction of arrivals from 2 microphones only, an impossible task in anechoic settings. These first results were submitted to an international conference. Future work will consider extension to more realistic and more complex scenarios (including more microphones, sources and reflective surfaces) and the estimation of other room properties such as the acoustic absorption at the boundaries, or ultimately, the room geometry.

7.6.3. Multichannel Audio Event and Room Classification

Participants: Marie-Anne Lacroix, Nancy Bertin.

Main collaborations: Pascal Scalart, Romuald Rocher (GRANIT Inria project-team, Lannion)

Typically, audio event detection and classification is tackled as a “pure” single-channel signal processing task. By constrast, audio source localization is the perfect example of multi-channel task “by construction”. In parallel, the need to classify the type of scene or room has emerged, in particular from the rapid development of wearables, the “Internet of things” and their applications. The PhD of Marie-Anne Lacroix, started in September 2018, combines these ideas with the aim of developing multi-channel, room-aware or spatially-aware audio classification algorithms for embedded devices. The PhD topic includes low-complexity and low-energy stakes, which will be more specifically tackled thanks to the GRANIT members area of expertise. During the first months of the PhD, we gathered existing data and identified the need for new simulations or recordings, and combined ideas from existing single-channel classification techniques with traditional spatial features in order to design a baseline algorithm for multi-channel joint localization and classification of audio events, currently under development.

7.7. 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 [56], in particular the System & Contrast model [1].

7.7.1. Tensor-based Representation of Sectional Units in Music

Participant: Frédéric Bimbot.

This work was primarily carried out by Corentin Guichaoua, former PhD student with Panama, now with IRMA (CNRS UMR 7501, Strasbourg).

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 [90] 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 time-scales 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 % on that task [41]

7.7.2. Modeling music by Polytopic Graphs of Latent Relations (PGLR)

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 [111] defined as the shortest displacement of notes, in semitones, between a pair of chords. Other chord representations and relations are possible, as studied in [113] 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 [112] and are related to 6 main redundant sequences which can be viewed as canonical multiscale structural patterns.

In parallel, recent work has also been dedicated to modeling melodic and rhythmic motifs in order to extend the polytopic model to multiple musical dimensions.

Results obtained in this framework illustrate the efficiency of the proposed model in capturing structural information within musical data and support the view that musical content can be delinearised in order to better describe its structure. Extensive results will be included in Corentin Louboutin's PhD, which is planned to be defended early 2019.

7.7.3. Exploring Structural Dependencies in Melodic Sequences using Neural Networks

Participants: Nathan Libermann, Frédéric Bimbot.

This work is carried out in the framework of a PhD, co-directed by Emmanuel Vincent (Inria-Nancy).

In order to be able to generate structured melodic phrases and section, we explore various schemes for modeling dependencies between notes within melodies, using deep learning frameworks.

A a first set of experiments, we have considered a GRU-based sequential learning model, studied under different learning scenarios in order to better understand the optimal architectures in this context that can achieve satisfactory results. By this means, we wish to explore different hypotheses relating to temporal non-invariance relationships between notes within a structural segment (motif, phrase, section).

We have defined three types of recursive architectures corresponding to different ways to exploit the local history of a musical note, in terms of information encoding and generalization capabilities.

These experiments have been conducted on the Lakh MIDI dataset and more particularly on a subset of 8308 monophonic 16-bar melodic segments. The obtained results indicate a non-uniform distribution of modeling capabilities prediction of recurrent networks, suggesting the utility of non-ergodic models for the generation of melodic segments [38].

Ongoing work is extending these findings to the design of specific NN architectures, sto account for this non-invariance of information across musical segments.

7.7.4. Graph Signal Processing for Multiscale Representations of Music Similarity

Participants: Valentin Gillot, Frédéric Bimbot.

“Music Similarity” is a multifaceted concept at the core of Music Information Retrieval (MIR). Among the wide range of possible definitions and approaches to this notion, a popular one is the computation of a so-called content-based similarity matrix (S), in which each coefficient is a similarity measure between descriptors of short time frames at different instants within a music piece or a collection of pieces.

Matrix S can be seen as the adjacency matrix of an underlying graph, embodying the local and non-local similarities between parts of the music material. Considering the nodes of this graph as a new set of indices for the original music frames or pieces opens the door to a “delinearized” representation of music, emphasizing its structure and its semiotic content.

Graph Signal Processing (GSP) is an emerging topic devoted to extend usual signal processing tools (Fourier analysis, filtering, denoising, compression, ...) to signals “living” on graphs rather than on the time line, and to exploit mathematical and algorithmic tools on usual graphs, in order to better represent and manipulate these signals. Toy applications of GSP concepts on music content in music resequencing and music inpainting are illustrating this trend.

From exploratory experiments, first observations point towards the following hypotheses :

- local and non-local structures of a piece are highlighted in the adjacency matrix built from a simple time-frequency representation of the piece,
- the first eigenvectors of the graph Laplacian provide a rough structural segmentation of the piece,
- clusters of frames built from the eigenvectors contain similar, repetitive sound sequences.

The goal of Valentin Gillot's PhD is to consolidate these hypotheses and investigate further the topic of Graph Signal Processing for music, with more powerful conceptual tools and experiments at a larger scale.

The core of the work will consist in designing a methodology and implement an evaluation framework so as to (i) compare different descriptors and similarity measures and their capacity to capture relevant structural information in music pieces or collection of pieces, (ii) explore the structure of musical pieces by refining the frame clustering process, in particular with a multi-resolution approach, (iii) identify salient characteristics of graphs in relation to mid-level structure models and (iv) perform statistics on the typical properties of the similarity graphs on a large corpus of music in relation to music genres and/or composers.

By the end of the PhD, we expect the release of a specific toolbox for music composition, remixing and repurposing using the concepts and algorithms developed during the PhD.

RAINBOW Project-Team

7. New Results

7.1. Optimal and Uncertainty-Aware Sensing

7.1.1. Visual Tracking for Motion Capture and virtual reality

Participants: Guillaume Cortes [Hybrid], Eric Marchand.

Considering the visual tracking system for motion proposed last year, we studied a novel approach for Mobile Spatial Augmented Reality on Tangible objects [14]. MoSART is dedicated to mobile interaction with tangible objects in single or collaborative situations. It is based on a novel 'all-in-one' Head-Mounted Display (AMD) including a projector (for the SAR display) and cameras (for the scene registration). Equipped with the HMD the user is able to move freely around tangible objects and manipulate them at will. The system tracks the position and orientation of the tangible 3D objects and projects virtual content over them. The tracking is a feature-based stereo optical tracking providing high accuracy and low latency. A projection mapping technique is used for the projection on the tangible objects which can have a complex 3D geometry. Several interaction tools have also been designed to interact with the tangible and augmented content, such as a control panel and a pointer metaphor, which can benefit as well from the MoSART projection mapping and

7.1.2. Deformable Object 3D Tracking based on Depth Information and Physical Model

Participants: Agniva Sengupta, Eric Marchand, Alexandre Krupa.

In the context of the iProcess project (see Section 9.3.3.2), we have developed a method for tracking rigid objects of complex shapes. This year, we started to elaborate a method to track deformable objects using a depth camera (RGB-D sensor). This method is based on the assumption that a coarse mesh representing the model of the object is known and that a simple volumetric tetrahedral mesh has been computed offline, representing the internal physical model of the object. To take into account the deformation of the object, a corotational Finite Element Method (FEM) is considered as the physical model. Given the sequential pointcloud of the object undergoing deformation, we have developed an algorithm that fits the deformable model to the observed pointcloud. The FEM simulation is done using the SOFA library and our approach was tested for the tracking of simulated deformation of objects. For the moment, the method succeeds to accurately track the object deformation, given that we know the point of application of force (causing the deformation) and the force direction vector. Online estimation of the direction vector of this force is currently a work in progress.

7.1.3. General Model-based Tracker

Participants: Souriya Trinh, Fabien Spindler, Eric Marchand, François Chaumette.

We have extended our model-based visual tracking method by considering as new potential measurement the depth map provided by a RGB-D sensor [75]. The method has been adapted to be fully modular and can combine edge, texture, and depth features. It has been released in the new version of ViSP.

7.1.4. Reflectance and Illumination Estimation for Realistic Augmented Reality

Participants: Salma Jiddi, Eric Marchand.

Photometric registration consists in blending real and virtual scenes in a visually coherent way. To achieve this goal, both reflectance and illumination properties must be estimated. These estimates are then used, within a rendering pipeline, to virtually simulate the real lighting interaction with the scene.

We have been interested in indoor scenes where light bounces off of objects with different reflective properties (diffuse and/or specular). In these scenarios, existing solutions often assume distant lighting or limit the analysis to a single specular object [63]. We address scenes with various objects captured by a moving RGB-D camera and estimate the 3D position of light sources. Furthermore, using spatio-temporal data, our algorithm recovers dense diffuse and specular reflectance maps. Finally, using our estimates, we demonstrate photo-realistic augmentations of real scenes (virtual shadows, specular occlusions) as well as virtual specular reflections on real world surfaces.

We also consider the problem of estimating the 3D position and intensity of multiple light sources using an approach based on cast shadows on textured real surfaces [62], [86]. We separate albedo/texture and illumination using lightness ratios between pairs of points with the same reflectance property but subject to different lighting conditions. Our selection algorithm is robust in presence of challenging textured surfaces. Then, estimated illumination ratios are integrated, at each frame, within an iterative process to recover position and intensity of light sources responsible of cast shadows.

7.1.5. Multi-Layered Image Representation for Robust SLAM

Participant: Eric Marchand.

Robustness of indirect SLAM techniques to light changing conditions remains a central issue in the robotics community. With the change in the illumination of a scene, feature points are either not extracted properly due to low contrasts, or not matched due to large differences in descriptors. We proposed a multi-layered image representation (MLI) that computes and stores different contrast-enhanced versions of an original image [76]. Keypoint detection is performed on each layer, yielding better robustness to light changes. An optimization technique is also proposed to compute the best contrast enhancements to apply in each layer in order to improve detection and matching. We extend the MLI approach [77] and we show how Mutual Information can be used to compute dynamic contrast enhancements on each layer. We demonstrate how this approach dramatically improves the robustness in dynamic light changing conditions on both synthetic and real environments compared to default ORB-SLAM. This work focuses on the specific case of SLAM relocalization in which a first pass on a reference video constructs a map, and a second pass with a light changed condition relocalizes the camera in the map.

7.1.6. Trajectory Generation for Optimal State Estimation

Participants: Marco Cognetti, Marco Ferro, Paolo Robuffo Giordano.

This activity addresses the general problem of *active sensing* where the goal is to analyze and synthesize optimal trajectories for a robotic system that can maximize the amount of information gathered by the (few) noisy outputs (i.e., sensor readings) while at the same time reducing the negative effects of the process/actuation noise. Indeed, the latter is far from being negligible for several robotic applications (a prominent example are aerial vehicles). Last year we developed a general framework for solving *online* the active sensing problem by continuously replanning an optimal trajectory that maximize a suitable norm of the Constructibility Gramian (CG), while also coping with a number of constraints including limited energy and feasibility. This approach, however, did not consider the presence of process noise which, as explained, can have a significant effect in many robotic systems of interest (e.g., UAVs). This year we have then extended this work to the case of a non-negligible process noise in [56], where we showed how to generate optimal trajectories able to still maximize the amount of information collected while moving, but by properly weighting (and attenuating) the negative effects of process noise in the execution of the planned trajectory. We are actually working towards the extension of this machinery to the case of realization of a robot task (e.g., reaching and grasping for a mobile manipulators), and to the mutual localization problem for a multi-robot group.

7.1.7. Cooperative Localization using Interval Analysis

Participants: Ide Flore Kenmogne Fokam, Vincent Drevelle, Eric Marchand.

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. Uncertainty in the landmarks positions have to be considered, but this can lead to pessimism in the computed solution. Hence we derived a quantifier-free expression of the pose solution-set to improve the vision-based position domain computation [66]. Image and range based cooperative localization of UAVs has been studied, first in the case of two robots sharing their measurements [65]. Then, scaling to the case of multiple robots as also been addressed, by sharing the computed position domains [64], [67].

7.2. Advanced Sensor-Based Control

7.2.1. Model Predictive Control for Visual Servoing of a UAV

Participants: Bryan Penin, François Chaumette, Paolo Robuffo Giordano.

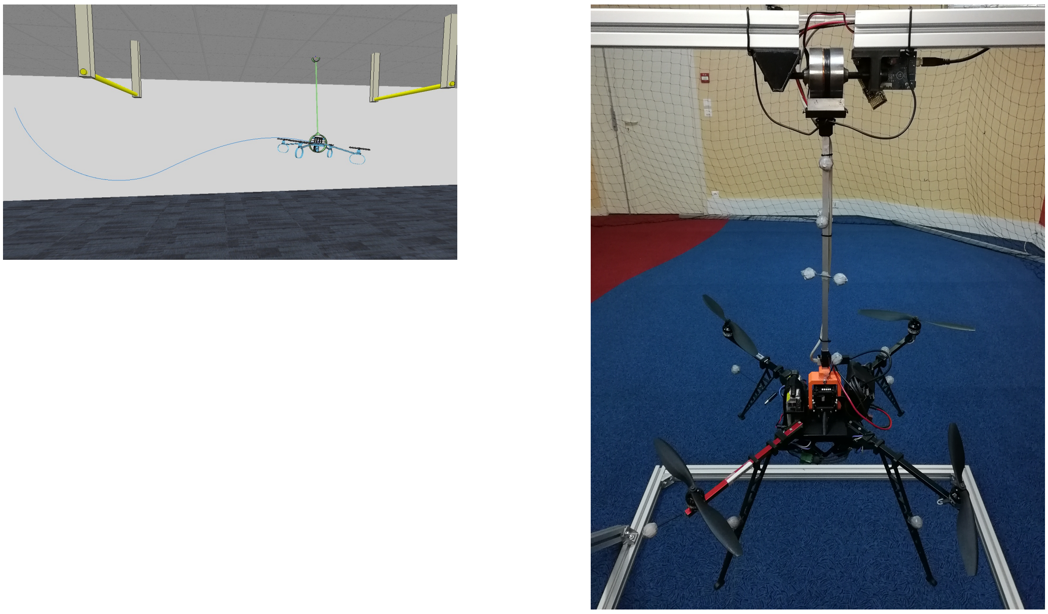
Visual servoing is a well-known class of techniques meant to control the pose of a robot from visual input by considering an error function directly defined in the image (sensor) space. These techniques are particularly appealing since they do not require, in general, a full state reconstruction, thus 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 have developed a series of online trajectory re-planning (MPC-like) schemes for explicitly dealing with this kind of constraints during flight. In particular, in [33], the problem of aggressive flight when tracking a target has been considered, with the additional (and complex) constraint of avoiding occlusions w.r.t. obstacles in the scene. A suitable optimization framework has been devised to be solved online during flight for continuously replanning the future UAV trajectory subject to the mentioned sensing constraints as well as actuation constraints. An experimental validation with the quadrotor UAVs available in the team has also been provided. In [34], we have instead considered the problem of planning a trajectory from a start to a goal location for a UAV equipped with an onboard camera, by assuming that measurements of environment landmarks (needed to recover the UAV state from visual input) may be intermittent due to occlusions by obstacles. The goal is then to plan a trajectory that can minimize the negative effects of “missing measurements” by keeping the state uncertainty limited despite the temporary loss of measurements. This planning problems has been solved by exploiting a bi-directional RRT algorithm for joining the start and goal locations, and an experimental validation has also been performed.

7.2.2. 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 “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. During last year we have considered in [17] the modeling, control and trajectory planning problem for 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), see Fig. 9 (a). We are currently finalizing a real prototype (Fig. 9 (b)) for obtaining an experimental validation of the whole approach.



(a) (b)

Figure 9. UAVs in Physical Interaction with the Environment. a) The simulated MonkeyRotor performing a hook-to-hook maneuver. b) The prototype currently under finalization.

7.2.3. *Trajectory Generation for Minimum Closed-Loop State Sensitivity*

Participants: Quentin Delamare, Paolo Robuffo Giordano.

The goal of this research activity is to propose a new point of view in addressing the control of robots under parametric uncertainties: rather than striving to design a sophisticated controller with some robustness guarantees for a specific system, we propose to attain robustness (for any choice of the control action) by suitably shaping the reference motion trajectory so as to minimize the *state sensitivity* to parameter uncertainty of the resulting closed-loop system. In [70], we have explored this novel idea by showing how to properly define and evaluate the *state sensitivity matrix* and its gradient w.r.t. the desired trajectory parameters. This then allows setting up an optimization problem in which the desired trajectory is optimized so as to minimize a suitable norm of the state sensitivity. The machinery has been applied to two case studies involving a unicycle and a planar quadrotor with successful results (monte-carlo statistical analysis). We are currently considering extensions of this initial idea (e.g., by also considering a notion of *input sensitivity*), as well as an experimental validation of the approach.

7.2.4. *Visual Servoing for Steering Simulation Agents*

Participants: Axel Lopez Gandia, Eric Marchand, François Chaumette, Julien Pettré.

This research activity is dedicated to the simulation of human locomotion, and more especially to the simulation of the visuomotor loop that controls human locomotion in interaction with the static and moving obstacles of its environment. Our approach is based on the principles of visual servoing for robots. To simulate visual perception, an agent perceives its environment through a virtual camera located the position of its head. The visual input is processed by each agent in order to extract the relevant information for controlling its motion. In particular, the optical flow is computed to give the agent access to the relative motion of visible objects around it. Some features of the optical flow are finally extracted to estimate the risk of collision with obstacle. We have established the mathematical relations between those visual features and the agent's self motion. Therefore, when necessary, the agent motion is controlled and adjusted so as to cancel the visual features indicating a risk of future collision. We are now in the process of evaluating our motion control technique and exploring relevant applications, as well as preparing a publication summarizing this work.

7.2.5. Study of human locomotion to improve robot navigation

Participants: Florian Berton, Julien Bruneau, Julien Pettré.

This research activity is dedicated to the study of human gaze behaviour during locomotion. This activity is directly linked to the previous one on simulation, as human locomotion study results will serve as an input for the design of novel models for simulation. In this activity, we are first interested in collective pedestrian dynamics, i.e., how humans move in crowds, how they interact locally and how this results into the emergence of specific patterns at larger scales [52]. Virtual Reality is one main experimental tools in our approach, so as to control and reproduce easily situations we expose participants to, as well as to explore the nature of the visual cues human use to control their locomotion [22], [68]. We are also interested in the study of the activity of the gaze during locomotion that, in addition to the classical study of kinematics motion parameters, provides information on the nature of visual information acquired by humans to move, and the relative importance of visual elements in their surroundings [54], [26]. We directly exploit our experimental result to propose relevant navigation control techniques for robot to make them more adapted to move among humans [41], [58].

7.2.6. Direct Visual Servoing

Participants: Quentin Bateux, Eric Marchand.

We proposed a deep neural network-based method to perform high-precision, robust and real-time 6 DOF positioning tasks by visual servoing [53]. A convolutional neural network is fine-tuned to estimate the relative pose between the current and desired images and a pose-based visual servoing control law is considered to reach the desired pose. This approach efficiently and automatically creates a dataset used to train the network. We show that this enables the robust handling of various perturbations (occlusions and lighting variations). We then propose the training of a scene-agnostic network by providing both the desired and current images to a deep network for generating the camera motion. The method is validated on a 6 DOF robot.

7.2.7. Visual Servoing using Wavelet and Shearlet Transforms

Participants: Lesley-Ann Duflot, Alexandre Krupa.

We pursued our work on the elaboration of a direct visual servoing method in which the signal control inputs are the coefficients of a multiscale image representation [4]. In particular, we considered the use of multiscale image representations that are based on discrete wavelet and shearlet transforms. This year, 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 [18]. We also considered this control framework in 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 of this application was to automatically retrieve the region of the sample that corresponds to an initial optical biopsy for diagnosis purpose. Experimental results demonstrated the efficiency of our approach that uses the wavelet coefficients of the OCT image as input of the control law to perform this task [61].

7.2.8. Visual Servoing from the Trifocal Tensor

Participants: Kaixiang Zhang, François Chaumette.

In visual servoing, three images are usually available at each iteration of the control loop: the very first one, the current one, and the desired one. That is why the trifocal tensor defined from this set of images is a potential candidate for providing visual features to be used as inputs of the control scheme. We have first modeled the interaction matrix related to the components of the trifocal tensor. We have then designed a set of reduced visual features with good decoupling properties, from which a thorough Lyapunov-based stability analysis has been developed [78].

7.2.9. 3D Steering of Flexible Needle by Ultrasound Visual Servoing

Participants: Jason Chevrie, Marie Babel, Alexandre Krupa.

Needle insertion procedures under ultrasound guidance are commonly used for diagnosis and therapy. However, it is often critical to accurately reach a targeted region due to the deflection of the flexible needle and the presence of intra-operative tissue motions. Therefore this year we improved our robotic framework dedicated to 3D steering of flexible needle that is based on ultrasound visual servoing. We developed a new control approach that both steers the flexible needle toward a desired target and compensates the tissue self-motion during the needle insertion. In our approach, the target to be reached by the needle is tracked in 2D ultrasound images and the needle tip position and orientation are measured by an electromagnetic tracker. Tissue motion compensation is performed using force feedback to reduce targeting error and forces applied to the tissue. The method also uses a mechanics-based interaction model that is updated online to provide the current shape of the deformable needle. In addition, a novel control law using task functions was proposed to fuse motion compensation, needle steering via manipulation of its base and steering of the needle tip in order to reach the target. Validation of the tracking and steering algorithms were performed in gelatin phantom and bovine liver on which periodical perturbation motions (magnitude of 15 mm) were applied to simulate physiological motions. Experimental results demonstrated that our approach can reach a moving target with an average targeting error of 1.2 mm and 2.5 mm in resp. gelatin and liver, which is accurate enough for common needle insertion procedures [12].

7.2.10. Robotic Assistance for Ultrasound Elastography by Visual Servoing, Force Control and Teleoperation

Participants: Pedro Alfonso Patlan Rosales, Alexandre Krupa.

Ultrasound elastography is an image modality that unveils elastic parameters of a tissue, which are commonly related with certain pathologies. It is performed by applying continuous stress variation on the tissue in order to estimate a strain map from successive ultrasound images. Usually, this stress variation is performed manually by the user through the manipulation of an ultrasound probe and it results therefore in an user-dependent quality of the strain map. To improve the ultrasound elastography imaging and provide quantitative measurements, we developed an assistant robotic palpation system that automatically applies the motion to a 2D or 3D ultrasound probe that is needed to generate in real-time the elastography images during teleoperation [5]. This year, we have extended our robotic framework by developing a method that provides to the user the capability to physically feel the stiffness of the observed tissue of interest via a haptic device. This work has been submitted to the ICRA 2019 conference.

7.2.11. Deformation Servoing of Soft Objects

Participant: Alexandre Krupa.

This year, we started a new research activity whose objective is to provide robotic control approaches that improve the dexterity of robots interacting with deformable objects. The goal is to control one or several robots interacting with a soft object in such a way to reach a desired configuration of object deformation. Nowadays, most of the existing deformation control methods require accurate models of the object and/or environment in order to perform such tasks. Contrarily to these methods, we want to propose model-free methods that rely only on visual observation provided by a RGB-D sensor to control the deformation of soft objects without a priori knowledge of their material mechanical parameters and without a priori knowledge of their environment. In a preliminary study, we compared the model-based method based on physics simulation (Finite Element Model) and the model-free method of the state of the art. We also developed a first approach based on visual servoing that uses in the robot control law an online estimation of the interaction matrix that links the variation of the object deformation to the velocity of the robot end-effector. These different approaches have been implemented in simulation and are currently tested on a robotic arm (Adept Viper 650) interacting with a soft object (sponge). The first results are encouraging since they showed that our model-free visual servoing approach based on online estimation of the interaction matrix provides similar results than the model-based approach based on physics simulation.

7.2.12. Multi-Robot Formation Control

Participants: Paolo Robuffo Giordano, Fabrizio Schiano.

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. 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. In [71], we considered the localization problem for multi-robots (that is, the problem of reconstructing the relative poses from the available bearing measurements), by recasting it as a nonlinear observability problem: this rigorous analysis led us to introduce the notion of *Dynamic Bearing Observability Matrix*, which in a sense extends the classical Bearing Rigidity Matrix to explicitly account for the robot motion. It was then possible to show that the scale factor of the formation is, indeed, observable by processing the bearing measurements and (known) agent motion, a result confirmed experimentally by employing a EKF on a group of quadrotor UAVs. This and more results on bearing-based formation control and localization for quadrotor UAVs are summarized in [7].

7.2.13. Coupling Force and Vision for Controlling Robot Manipulators

Participants: François Chaumette, Paolo Robuffo Giordano, Alexander Oliva.

The goal of this recent activity is about coupling visual and force information for advanced manipulation tasks. To this end, we plan to exploit the recently acquired Panda robot (see Sect. 6.6.4), a state-of-the-art 7-dof manipulator arm with torque sensing in the joints, and the possibility to command torques at the joints or forces at the end-effector. Thanks to this new robot, we plan to study how to optimally combine the torque sensing and control strategies that have been developed over the years to also include in the loop the feedback from a vision sensor (a camera). In fact, the use of vision in torque-controlled robot is quite limited because of many issues, among which the difficulty of fusing low-rate images (about 30 Hz) with high-rate torque commands (about 1 kHz), the delays caused by any image processing and tracking algorithms, and the unavoidable occlusions that arise when the end-effector needs to approach an object to be grasped. Our aim is therefore to advance the state-of-the-art in the field of torque-controlled manipulator arms by also including in the loop in an explicit way the use of a vision sensor. We will probably rely on estimation strategies for coping with the different rates of the two sensing modalities, and to online trajectory replanning strategies for dealing with constraints of the system (e.g., limited fov of the camera, of the fact that visibility of the target object is lost when closing in for grasping).

7.3. Haptic Cueing for Robotic Applications

7.3.1. Haptic Guidance of a Biopsy Needle

Participants: Hadrien Gurnel, Alexandre Krupa.

The objective of this work is to provide assistance during manual needle steering for biopsies or therapy purposes (see Section 9.1.6). At the difference of our work presented in Section 7.2.9 where a robotic system is used to autonomously actuate the needle, we propose in this study another way of assistance for needle insertion. The principle is to provide haptic cue feedback to the clinician in order to help him during his manual gesture by the application of repulsive or attractive forces. The proposed solution is based on a shared robotic control, where the clinician and a haptic device, both holding the base of the needle, cooperate together. In a preliminary study, we elaborated 5 different haptic-guidance strategies to assist the needle pre-positioning and pre-orienting on a pre-defined insertion point, and with a pre-planned desired incidence angle. From this pre-operative information and intra-operative measurements of the location of the needle, haptic cues are generated to guide the clinician toward the desired needle position and orientation. These 5 different haptic guides were recently tested by 2 physicians, both experts in needle manipulation and compared to the reference gesture performed without assistance. The results have been submitted to the IPCAI 2019 conference. Future work will consist in evaluating the different haptic guides from an user-experience study involving more participants.

7.3.2. Wearable Haptics

Participants: Marco Aggravi, Claudio Pacchierotti, Paolo Robuffo Giordano.

We worked on a wearable haptic device for the forearm and its application in robotic teleoperation [8]. The device is able to provide skin stretch, pressure, and vibrotactile stimul, see Fig. 10 . Two servo motors, housed in a 3D printed lightweight platform, actuate an elastic fabric belt, wrapped around the arm. When the two servo motors rotate in opposite directions, the belt is tightened (or loosened), thereby compressing (or decompressing) the arm. On the other hand, when the two motors rotate in the same direction, the belt applies a shear force to the arm skin. Moreover, the belt houses four vibrotactile motors, positioned evenly around the arm at 90 degrees from each other. The device weights 220 g for 115×122×50 mm of dimensions, making it wearable and unobtrusive. We carried out a perceptual characterization of the device as well as two human-subjects teleoperation experiments in a virtual environment, employing a total of 34 subjects.

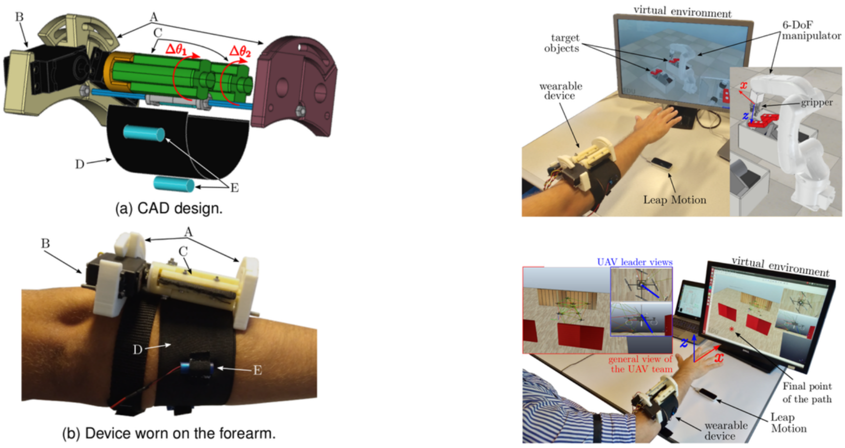


Figure 10. The proposed wearable device for the arm and its evaluation. The device consists of a static platform (A) that accommodates two servomotors (B) and two pulleys (C), a fabric belt (D), and four vibrotactile motors (E).

In the first experiment, participants were asked to control the motion of a robotic manipulator for grasping an object; in the second experiment, participants were asked to teleoperate the motion of a quadrotor fleet along a given path. In both scenarios, the wearable haptic device provided feedback information about the status of the slave robot(s) and of the given task. Results showed the effectiveness of the proposed device. Performance on completion time, length trajectory, and perceived effectiveness when using the wearable device improved of 19.8%, 25.1%, and 149.1% than when wearing no device, respectively. Finally, all subjects but three preferred the conditions including wearable haptics.

7.3.3. Mid-Air Haptic Feedback

Participants: Claudio Pacchierotti, Thomas Howard.

GUIs have been the gold standard for more than 25 years. However, they only support interaction with digital information indirectly (typically using a mouse or pen) and input and output are always separated. Furthermore, GUIs do not leverage our innate human abilities to manipulate and reason with 3D objects. Recently, 3D interfaces and VR headsets use physical objects as surrogates for tangible information, offering limited malleability and haptic feedback (e.g., rumble effects). In the framework of project H-Reality, we are working to develop novel mid-air haptics paradigm that can convey the information spectrum of touch sensations in the real world, motivating the need to develop new, natural interaction techniques. Moreover, we want to use robotic manipulators to enlarge the workspace of mid-air haptic systems, using depth cameras and visual servoing techniques to follow the motion of the user's hand.

7.3.4. Haptic Cueing in Telemanipulation

Participants: Firas Abi Farraj, Paolo Robuffo Giordano, Claudio Pacchierotti.

Robotic telemanipulators are already widely used in nuclear decommissioning sites for handling radioactive waste. However, currently employed systems are still extremely primitive, making the handling of these materials prohibitively slow and ineffective. As the estimated cost for the decommissioning and clean-up of nuclear sites keeps rising, it is clear that one would need faster and more effective approaches. Towards this goal, we presented the user evaluation of a recently proposed haptic-enabled shared-control architecture for telemanipulation [51]. An autonomous algorithm regulates a subset of the slave manipulator degrees of freedom (DoF) in order to help the human operator in grasping an object of interest. The human operator can then steer the manipulator along the remaining null-space directions with respect to the main task by acting on a grounded haptic interface. The haptic cues provided to the operator are designed in order to inform about the feasibility of the user's commands with respect to possible constraints of the robotic system. This work compared this shared-control architecture against a classical 6-DOF teleoperation approach in a real scenario by running experiments with 10 subjects. The results clearly show that the proposed shared-control approach is a viable and effective solution for improving currently-available teleoperation systems in remote telemanipulation tasks.

7.3.5. Haptic Feedback for an Augmented Wheelchair Driving Experience

Participants: Louise Devigne, Marie Babel, François Pasteau.

Smart powered wheelchairs can increase mobility and independence for people with disability by providing navigation support. For rehabilitation or learning purposes, it would be of great benefit for wheelchair users to have a better understanding of the surrounding environment while driving. Therefore, a way of providing navigation support is to communicate information through a dedicated and adapted feedback interface. We have then proposed a framework in which feedback is provided by sending forces through the wheelchair controller as the user steers the wheelchair. This solution is based on a low complex optimization framework able to perform smooth trajectory correction and to provide obstacle avoidance. The impact of the proposed haptic guidance solution on user driving performance was assessed during this pilot study for validation purposes through an experiment with 4 able-bodied participants. Results of this pilot study showed that the number of collisions significantly decreased while force feedback was activated, thus validating the proposed framework [60].

7.3.6. Virtual Shadows to Improve Self Perception in CAVE

Participants: Guillaume Cortes [Hybrid], Eric Marchand.

In immersive projection systems (IPS), the presence of the user's real body limits the possibility to elicit a virtual body ownership illusion. But, is it still possible to embody someone else in an IPS even though the users are aware of their real body ? In order to study this question, we propose to consider using a virtual shadow in the IPS, which can be similar or different from the real user's morphology. We have conducted an experiment ($N = 27$) to study the users' sense of embodiment whenever a virtual shadow was or was not present. Participants had to perform a 3D positioning task in which accuracy was the main requirement. The results showed that users widely accepted their virtual shadow (agency and ownership) and felt more comfortable when interacting with it (compare to no virtual shadow). Yet, due to the awareness of their real body, the users have less acceptance of the virtual shadow whenever the shadow gender differs from their own. Furthermore, the results showed that virtual shadows increase the users' spatial perception of the virtual environment by decreasing the inter-penetrations between the user and the virtual objects. Taken together, our results promote the use of dynamic and realistic virtual shadows in IPS and pave the way for further studies on "virtual shadow ownership" illusion.

7.4. Shared Control Architectures

7.4.1. Shared Control for Remote Manipulation

Participants: Firas Abi Farraj, Paolo Robuffo Giordano, Claudio Pacchierotti, Rahaf Rahal, Mario Selvaggio.

As teleoperation systems become more sophisticated and flexible, the environments and applications where they can be employed become less structured and predictable. This desirable evolution toward more challenging robotic tasks requires an increasing degree of training, skills, and concentration from the human operator. For this reason, researchers started to devise innovative approaches to make the control of such systems more effective and intuitive. In this respect, shared control algorithms have been investigated as one the main tools to design complex but intuitive robotic teleoperation system, helping operators in carrying out several increasingly difficult robotic applications, such as assisted vehicle navigation, surgical robotics, brain-computer interface manipulation, rehabilitation. This approach makes it possible to share the available degrees of freedom of the robotic system between the operator and an autonomous controller. The human operator is in charge of imparting high level, intuitive goals to the robotic system; while the autonomous controller translates them into inputs the robotic system can understand. How to implement such division of roles between the human operator and the autonomous controller highly depends on the task, robotic system, and application. Haptic feedback and guidance have been shown to play a significant and promising role in shared control applications. For example, haptic cues can provide the user with information about what the autonomous controller is doing or is planning to do; or haptic force can be used to gradually limit the degrees of freedom available to the human operator, according to the difficulty of the task or the experience of the user. The dynamic nature of haptic guidance enables us to design very flexible robotic system, which can easily and rapidly change the division of roles between the user and autonomous controller.

Along this general line of research, we worked at different approaches:

- We proposed novel haptic guidance methods for a dual-arm telerobotic manipulation system [36] which are able to deal with several different constraints, such as collisions, joint limits, and singularities. We combined the haptic guidance with shared-control algorithms for autonomous orientation control and collision avoidance meant to further simplify the execution of grasping tasks. In addition, a human subject study was carried out to assess the effectiveness and applicability of the proposed control approaches both in simulated and real scenarios. Results showed that the proposed haptic-enabled shared-control methods significantly improve the performance of grasping tasks with respect to the use of classic teleoperation with neither haptic guidance nor shared control. Live demos of some of these approaches have been shown to the general public at the Maker Faire 2018 in Rome.
- In the framework of the RoMANS H2020 project, we worked together with CEA to implement an intuitive and effective shared-control teleoperation system with haptic feedback using the CEA robotic hand at the slave side and the Haption glove at the master side. The system was tested at CEA in an object grasping, manipulation, and sorting scenario. A video of the experiment is available at: https://youtu.be/M-tpVP9Fakc.
- Finally, in [38] we reported the results of a collaborative project involving LAAS-CNRS as leader, where we implemented an aerial-ground comanipulator system, denoted *Tele-MAGMaS*, where a fixed-based manipulator arm cooperates with a UAV equipped with an onboard gripper for carrying together (and manipulating) a long bar. The system has been demonstrated live during the Hannover Fair in 2017.

7.4.2. Shared Control for Mobile Robot Navigation

Participant: Paolo Robuffo Giordano.

Besides manipulators, we also considered shared control algorithms for mobile robot navigation. In [25], we have presented (and experimentally validated) an online trajectory planning approach that allows a human operator to act on the trajectory to be tracked by a mobile robot (a quadrotor UAV in the experiments) in conjunction with the robot autonomy which can locally modify the planned trajectory for avoiding obstacles of staying close to points of interest. This “shared planning approach” is quite general and its application to other robotic systems is under investigation.

7.4.3. Shared Control of a Wheelchair for Navigation Assistance

Participants: Louise Devigne, Marie Babel.

Power wheelchairs allow people with motor disabilities to have more mobility and independence. However, driving safely such a vehicle is a daily challenge particularly in urban environments while encountering *negative* obstacles, dealing with uneven grounds, etc. Indeed, differences of elevation have been reported to be one of the most challenging environmental barrier to negotiate while driving a wheelchair with tipping and falling being is the most common accidents power wheelchair users encounter. It is thus our actual challenge to design assistive solutions for power wheelchair navigation in order to improve safety while navigating in such environments. To this aim, we proposed a first shared-control algorithm which provides assistance while navigating with a wheelchair in an environment consisting of negative obstacles [80].

7.4.4. 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 the dynamic and kinematic effects of the passive caster wheels, particularly during maneuvers in narrow rooms and direction changes. In order to prevent their nasty behaviour, we proposed a caster wheel behavior model based on experimental measurements. The study has been realised for the three existing types of wheelchair, which present different kinematic behaviors, i.e. front caster type, rear caster type and mid-wheel drive. The orientation of the caster wheels has been measured experimentally for different initial orientations, velocities and user mass, according to a predefined experimental design. The repeatability of the motions has been studied, and from these measurements, their behavior has been modeled. By using this model with the wheelchair kinematic expressions, we were able to calculate the real trajectory of the wheelchair to enhance an existing driving assistance for powered wheelchair [79].

7.4.5. Wheelchair Autonomous Navigation for Fall Prevention

Participants: Solenne Fortun, Marie Babel.

The Prisme project (see Section 9.1.7) 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. This year, we designed the related control framework based on sensor-based servoing principles and validated it in simulation. Next step will consist in realizing tests within the Rehabilitation Center of Pôle Saint Hélier.

7.4.6. Robot-Human Interactions during Locomotion

Participants: Julien Legros, Javad Amirian, Fabien Grzeskowiak, Ceilidh Hoffmann, Marie Babel, Jean Bernard Hayet, Julien Pettré.

This research activity is dedicated to the design of robot navigation techniques to make them capable of safely moving through a crowd of people. We are following two main research paths. The first one is dedicated to the prediction of crowd motion based on the state of the crowd as sensed by a robot. The second one is dedicated to the creation of a virtual reality platform that enables robots and humans to share a common virtual space where robot control techniques can be tested with no physical risk of harming people, as they remain separated in the physical space. We are currently developing these ideas, which should bring good results in the near future.

SIROCCO Project-Team

7. New Results

7.1. Visual Data Analysis

Scene depth, Scene flows, 3D modelling, Light-fields, 3D point clouds

7.1.1. 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. We have introduced the concept of super-ray, which is a grouping of rays within and across views, 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 for correcting refocusing angular aliasing. The concept of super-rays has been extended to video light fields addressing problems of temporal tracking of super-rays using sparse scene flows[15].

7.1.2. Scene depth estimation from light fields

Participants: Christian Galea, Christine Guillemot, Xiaoran Jiang, Jinglei Shi.

While there exist scene depth and scene flow estimation methods, these methods, mostly designed for stereo content or for pairs of rectified views, do not effectively apply to new imaging modalities such as light fields. We have focused on the problem of *scene depth estimation* for every viewpoint of a dense light field, exploiting information from only a sparse set of views [17]. This problem is particularly relevant for applications such as light field reconstruction from a subset of views, for view synthesis, for 3D modeling and for compression. Unlike most existing methods, the proposed algorithm computes disparity (or equivalently depth) for every viewpoint taking into account occlusions. In addition, it preserves the continuity of the depth space and does not require prior knowledge on the depth range. Experiments show that, both for synthetic and real light fields, our algorithm achieves competitive performance compared to state-of-the-art algorithms which exploit the entire light field and usually generate the depth map for the center viewpoint only. Figure 2 shows the estimated depth map for a synthetic light field in comparison with the ground truth. The estimated depth maps allow us to construct accurate 3D point clouds of the captured scene [16]. This work is now pursued considering deep learning solutions.

7.1.3. Scene flow estimation from light fields

Participants: Pierre David, Christine Guillemot.

Temporal processing of dynamic 3D scenes requires estimating the displacement of the objects in the 3D space, i.e., so-called scene flows. Scene flows can be seen as 3D extensions of optical flows by also giving the variation in depth along time in addition to the optical flow. Estimating dense scene flows in light fields pose obvious problems of complexity due to the very large number of rays or pixels. This is even more difficult when the light field is sparse, i.e., with large disparities, due to the problem of occlusions. We have addressed the complexity problem by designing a sparse estimation method followed by a densification step that avoids the difficulty of computing matches in occluded areas. The developments in this area are also made difficult due to the lack of test data, i.e., there is no publicly available synthetic video light fields with the corresponding ground truth scene flows. In order to be able to assess the performance of the proposed method, we have therefore created synthetic video light fields from the MPI Sintel dataset. This video light field data set has been produced with the Blender software by creating new production files placing multiple cameras in the scene, controlling the disparity between the set of views.

Figure 2. Estimated depth map (middle) for the light field 'Buddha' in comparison with the ground truth (right).

7.2. Signal processing and learning methods for visual data representation and compression

Sparse representation, data dimensionality reduction, compression, scalability, rate-distortion theory

7.2.1. *Multi-shot single sensor light field camera using a color coded mask*

Participant: Christine Guillemot.

In collaboration with the University of Linkoping (Prof. J. unger, Dr. E. Miandji), we have proposed a compressive sensing framework for reconstructing a light field from a single-sensor consumer camera capture with color coded masks [19]. The proposed *camera architecture* captures incoherent measurements of the light field via a controllable color mask placed in front of the sensor. To enhance the incoherence, hence the reconstruction quality, we propose to utilize multiple shots where, for each shot, the mask configuration is changed to create a new random pattern. To reduce computations and increase the incoherence, we also perform a random sampling of the spatial domain. The compressive sensing framework relies on a dictionary trained over a light field data set. Numerical simulations show significant improvements compared with a similar coded aperture system for light field capture.

7.2.2. *Compressive 4D light field reconstruction*

Participants: Christine Guillemot, Fatma Hawary.

Exploiting the assumption that light field data is sparse in the Fourier domain, we have also developed a new method for reconstructing a 4D light field from a random set of measurements [14]. The reconstruction algorithm searches for these bases (i.e., their frequencies) which best represent the 4D Fourier spectrum of the sampled light field. The method has been further improved by introducing an orthogonality constraint on the residue, in the same vein as orthogonal matching pursuit but in the Fourier transform domain, as well as a refinement for non integer frequencies. The method achieves a very high reconstruction quality, in terms of PSNR (more than 1dB gain compared to state-of-the-art algorithms).

7.2.3. *Light fields dimensionality reduction with low-rank models*

Participants: Elian Dib, Christine Guillemot, Xiaoran Jiang.

We have further investigated low-rank approximation methods exploiting data geometry for dimensionality reduction of light fields. While our first solution was considering global low-rank models based on homographies, we have recently developed local low-rank models exploiting disparity. The local support of the approximation is given by super-rays (see section 7.1.1). The super-rays group super-pixels which are consistent across the views while being constrained to be of same shape and size. The corresponding super-pixels in all views are found thanks to disparity compensation. In order to do so, a novel method has been proposed to estimate the disparity for each super-ray using a low rank prior, so that the super-rays are constructed to yield the lowest approximation error for a given rank. More precisely, the disparity for each super-ray is found in order to align linearly correlated sub-aperture images in such a way that they can be approximated by the considered low rank model. The rank constraint is expressed as a product of two matrices, where one matrix contains basis vectors (or eigen images) and where the other one contains weighting coefficients. The eigen images are actually splitted into two sets, one corresponding to light rays visible in all views and a second one, very sparse, corresponding to occluded rays (see Fig. 3). A light field compression algorithm has been designed encoding the different components of the resulting low rank approximation.

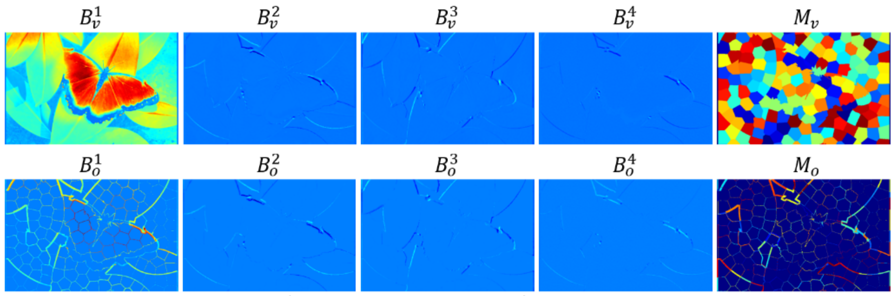


Figure 3. Eigen-images and segmentation maps for visible and occluded sets of pixels.

7.2.4. Graph-based transforms for light fields and omni-directional image compression

Participants: Christine Guillemot, Thomas Maugey, Mira Rizkallah, Xin Su.

Graph-based transforms are interesting tools for low-dimensional embedding of light field data. This embedding can be learned with a few eigenvectors of the graph Laplacian. However, the dimension of the data (e.g., light fields) has obvious implications on the storage footprint of the Laplacian matrix and on the eigenvectors computation complexity, making graph-based *non separable* transforms impractical for such data. To cope with this difficulty, in [21], we have first developed *local super-rays based separable (spatial followed by angular)* weighted and unweighted transforms to jointly capture light fields correlation spatially and across views. While separable transforms on super-rays allow us to significantly decrease the eigenvector computation complexity, the basis functions of the spatial graph transforms to be applied on the super-ray pixels of each view are often not compatible, resulting in decreased correlation of the coefficients across views, hence in a loss of performance of the angular transform, compared to the non-separable case. We have therefore developed a graph construction optimization procedure which seeks to find the eigen-vectors having the best alignment with those computed on a reference frame while still approximately diagonalizing their respective Laplacians. Fig.4 shows the second eigenvector of different super-pixels belonging to the same super-ray before and after optimization. A rate-distortion optimized graph partitioning algorithm has also been developed [20] for coding 360° videos signals, to achieve a good trade-off between distortion, smoothness of the signal on each subgraph, and the coding cost of the graph partition.

7.2.5. Neural networks for learning image transforms and predictors

Participants: Thierry Dumas, Christine Guillemot, Aline Roumy.

We have explored the problem of learning transforms for image compression via autoencoders. Learning a transform is equivalent to learning an autoencoder, which is of its essence unsupervised and therefore more difficult than classical supervised learning. In compression, the learning has in addition to be performed under a rate-distortion criterion, and not only a distortion criterion. Usually, the rate-distortion performances of image compression are tuned by varying the quantization step size. In the case of autoencoders, this in principle

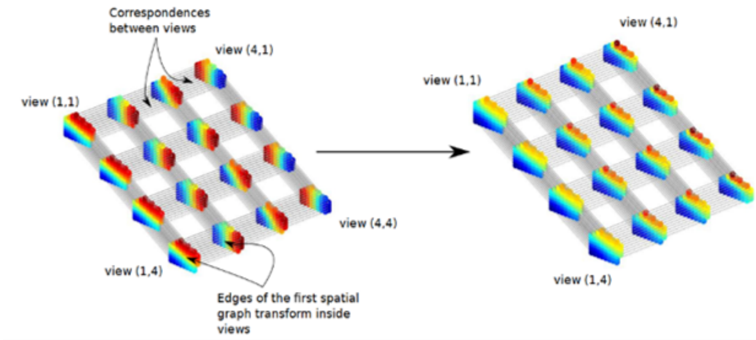


Figure 4. Second eigenvector of super-pixels forming a super-ray before and after optimization.

would require learning one transform per rate-distortion point at a given quantization step size. We have shown in [12] that comparable performances can be obtained with a unique learned transform. The different rate-distortion points are then reached by varying the quantization step size at test time. This approach saves a lot of training time.

Another important operator in compression algorithm is the predictor that aims at capturing spatial correlation. We have developed a set of neural network architectures, called Prediction Neural Networks Set (PNNS), based on both fully-connected and convolutional neural networks, for intra image prediction. It is shown that, while fully-connected neural networks give good performances for small block sizes, convolutional neural networks provide better predictions in large blocks with complex textures. Thanks to the use of masks of random sizes during training, the neural networks of PNNS well adapt to the available context that may vary, depending on the position of the image block to be predicted. Unlike the H.265 intra prediction modes, which are each specialized in predicting a specific texture, the proposed PNNS can model a large set of complex textures.

7.2.6. Cloud-based predictors and neural network temporal predictors video compression

Participants: Jean Begaint, Christine Guillemot.

Video codecs are primarily designed assuming that rigid, block-based, two-dimensional displacements are suitable models to describe the motion taking place in a scene. However, translational models are not sufficient to handle real world motion such as camera zoom, shake, pan, shearing or changes in aspect ratio. Building upon the region-based geometric and photometric model proposed in [5] to exploit correlation between images in the cloud, we have developed a region-based inter-prediction scheme for video compression. The proposed predictor is able to estimate multiple homography models in order to predict complex scene motion. We also introduce an affine photometric correction to each geometric model. Experiments on targeted sequences with complex motion demonstrate the efficiency of the proposed approach compared to the state-of- the-art HEVC video codec [11]. To further improve the accuracy of the temporal predictor, we have explored the use of deep neural networks for frame prediction and interpolation, and preliminary results have shown gains going up to 5% compared with the latest HEVC video codec.

7.3. Algorithms for inverse problems in visual data processing

Inpainting, view synthesis, super-resolution

7.3.1. View synthesis in light fields and stereo set-ups

Participants: Simon Evain, Christine Guillemot, Matthieu Hog, Xiaoran Jiang.

We have developed a lightweight convolutional neural network architecture able to perform view synthesis with occlusion handling in a stereo context, from one single, unlabelled and unannotated image, beyond state-of-the-art performance and with only a small amount of data required for training. In particular, it is able, at training and at test time, to estimate the disparity map corresponding to the problem at hand, and to evaluate a confidence in its prediction when using said disparity map for the synthesis. Knowing this confidence measure, it is then able to refine the value of the pixels wrongly estimated, with a refinement network component. The end result is a prediction built from a geometrical analysis of the scene, and completed in wrongly predicted areas by occlusion handling. Since 3D scene information is extracted in the course of the analysis, multiple new views can then be generated by interpolation.

Finally, in collaboration with Technicolor (N. Sabater and M. Hog), we have explored a novel way using recurrent neural networks to solve the problem of view synthesis in light fields. In particular, we proposed a novel solution using Long Short Term Memory Networks on a plane sweep volume. The approach has the advantage of having very few parameters and can be run on arbitrary sequence length. We have shown that the approach yields results that are competitive with the state of the art for dense light fields. Experimental results also show promising results when run on wider baselines.

7.3.2. Light field inpainting and restoration

Participants: Pierre Allain, Christine Guillemot, Laurent Guillo.

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 occluding mask. We have developed a method that is computationally fast while giving coherent disparity in the masked region, allowing us to inpaint a light field of 81 views in a few seconds [10].

We have also developed a novel light field denoising algorithm using a vector-valued regularization operating in the 4D ray space. More precisely, the method performs a PDE-based anisotropic diffusion along directions defined by local structures in the 4D ray space. It does not require prior estimation of disparity maps. The local structures in the 4D light field are extracted using a 4D tensor structure. We use a diffusivity coefficient derived from the amount of local variations in the 4D space to control the smoothing along directions, surfaces, or volumes in the 4D ray space. The diffusivity coefficient is computed as a function of the 4 eigenvalues of the 4D structure tensor. Experimental results show that the proposed denoising algorithm performs well compared to state of the art methods, while keeping tractable complexity, even with high noise levels (see Fig.5).

7.3.3. High dynamic range light fields capture

Participant: Christine Guillemot.

In collaboration with Trinity College Dublin (Prof. A. Smolic, Dr. M. Le Pendu), we have proposed a method for capturing *High Dynamic Range (HDR) light fields* with dense viewpoint sampling. Analogously to the traditional HDR acquisition process, several light fields are captured at varying exposures with a plenoptic camera. The raw data are de-multiplexed to retrieve all light field viewpoints for each exposure. We then perform a soft detection of saturated pixels. Considering a matrix which concatenates all the vectorized views, we formulate the problem of recovering saturated areas as a Weighted Low Rank Approximation (WLRA) where the weights are defined from the soft saturation detection. The proposed WLRA method [18], extending the matrix completion algorithm of [7] to nonbinary weights, is shown to better handle the transition between the saturated and non-saturated areas. While the Truncated Nuclear Norm (TNN) minimization, traditionally used for single view HDR imaging, does not generalize to light fields, the proposed WLRA method successfully recovers the parallax in the over-exposed areas.

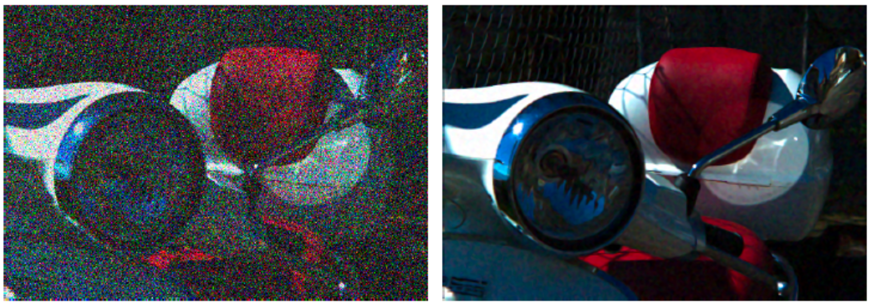


Figure 5. Illustration of denoising results, with additive white Gaussian noise of standard deviation $\sigma = 100$.

7.4. Distributed processing and robust communication

Information theory, stochastic modelling, robust detection, maximum likelihood estimation, generalized likelihood ratio test, error and erasure resilient coding and decoding, multiple description coding, Slepian-Wolf coding, Wyner-Ziv coding, information theory, MAC channels

7.4.1. Information theoretic bounds for sequential massive random access to large database of correlated data

Participants: Thomas Maugey, Mai Quyen Pham, Aline Roumy.

Massive random access is a new source coding paradigm that we proposed. It allows us to extract arbitrary sources from an appropriately compressed database purely by bit extraction. We studied the sequential aspect of this problem where the clients successively access to one source after the other. Theoretical bounds have been derived, and it was shown that the extraction can be done at the same rate as if the database was decoded and the requested sources were re-encoded. As for the storage, a reasonable overhead is required. In [26], we derived the optimal storage and transmission rate regions to the case of more general sources, which occur in practical scenarios. For the lossless source coding problem, we considered non i.i.d. sources (i.e., with memory, but also non necessary ergodic). We also showed that, in the case source statistics are unknown, the rate is increased by a factor that vanishes as the length of the data goes to infinity. Lossy compression is another context of interest, in particular for the application to video. Therefore, we derived achievable storage and transmission rate regions under a distortion constraint for i.i.d. [26] and correlated [13] Gaussian sources. Similarly, the transmission rate-distortion region is the same as if re-encoding of the requested sources was allowed. We are currently extending this work, by studying the constraints of the successive user requests and their influence on the transmission-storage rates performance.

7.4.2. Correlation model selection for interactive video communication

Participants: Navid Mahmoudian Bidgoli, Thomas Maugey, Aline Roumy.

One application of the sequential massive random access problem is interactive video communication 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. A question of interest is therefore how to select a model among a candidate set of models that incurs the lowest extra rate cost to the system. To answer this question, one should evaluate the effect on the transmission rate of using at the decoder a wrong model distribution. This question is related to an open problem in information theory called the mismatch capacity. So, we did not tackle the question for any type of code as in the case of the mismatch capacity. In contrast, we focused on a type of code of practical interest: the linear codes. More precisely, we proposed a criterion to select the model when a linear block code is used for compression. We showed that, experimentally, the proposed bound is an accurate estimate of the effect of using a wrong model.

7.4.3. Compression of spatio-temporally correlated and massive georeferenced data

Participants: Thomas Maugey, Aline Roumy.

Another application of the sequential massive random access problem is interactive compression of spatio-temporally correlated sources. For example, highly instrumented smart cities are facing problems of management and storage of a large volume of data coming from an increasing number of sources. In [23] different compression schemes have been proposed that are able to exploit not only the temporal but also the spatial correlation between data sources. A special focus was made on a scheme where some sensors are used as references to predict the remaining sources. Finally, an adaptation of the scheme was proposed to offer interactivity and free selection of some sources by a client. This work was been done in collaboration with the Inria I4S project-team (A. Criniere), IFFSTAR (J. Dumoulin) and the L2S (M. Kieffer).

7.4.4. ICON 3D - Interactive COding for Navigation in 3D scenes

Participants: Navid Mahmoudian Bidgoli, Thomas Maugey.

In the context of the ICON3D project, in collaboration with I3S-Nice (F. Payan), we have proposed a novel prediction tool for improving the compression performance of texture atlases of 3D meshes. This algorithm, called Geometry-Aware (GA) intra coding, takes advantage of the topology of the associated 3D meshes, in order to reduce the redundancies in the texture map. For texture processing, the general concept of the conventional intra prediction, used in video compression, has been adapted to utilize neighboring information on the 3D surface. We have also studied how this prediction tool can be integrated into a complete coding solution. In particular, a new block scanning strategy, as well as a graph-based transform for residual coding have been proposed. Experimental results show that the knowledge of the mesh topology can significantly improve the compression efficiency of texture atlases.