RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2017

Section New Results

Edition: 2018-02-19

ASAP Project-Team

6. New Results

6.1. Theory of Distributed Systems

6.1.1. Simulation of Partial Replication in Distributed Transactional Memory

Participant: François Taïani.

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

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

6.1.2. Distributed Universal Constructions: a Guided Tour

Participant: Michel Raynal.

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

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

Participant: Michel Raynal.

This work introduced a signature-free distributed algorithm which builds an atomic read/write shared memory on top of a fully connected peer-to-peer n-process asynchronous message-passing system in which up to t<n/3 processes may commit Byzantine failures. From a conceptual point of view, this algorithm is designed to be as close as possible to the algorithm proposed by [42], which builds an atomic register in an n-process asynchronous message-passing system where up to t<n/2 processes may crash. The proposed algorithm is particularly simple. It does not use cryptography to cope with Byzantine processes, and is optimal from a t-resilience point of view (t<n/3). A read operation requires $O(n)$ messages, and a write operation requires $O(n^2)$ messages. This work was done in collaboration with Achour Mostéfaoui, Matoula Petrolia and Claude Jard from the University of Nantes and was published in Theory of Computing Systems [19].

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

Participant: Michel Raynal.

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

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

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

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

Participant: Michel Raynal.

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

6.1.6. Long-Lived Tasks

Participant: Michel Raynal.

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

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

6.1.7. Which Broadcast Abstraction Captures k-Set Agreement?

Participant: Michel Raynal.

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

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

Participant: Michel Raynal.

We introduced a new algorithm that reduces multivalued consensus to binary consensus in an asynchronous message-passing system made up of n processes where up to t may commit Byzantine failures. This algorithm has the following noteworthy properties: it assumes $t < n/3t < n/3$ (and is consequently optimal from a resilience point of view), uses $O(n^2)$ messages, has a constant time complexity, and uses neither signatures nor additional computational power (such as random numbers, failure detectors, additional scheduling assumption, or additional synchrony assumption). The design of this reduction algorithm relies on two new all-to-all communication abstractions. The first one allows the non-faulty processes to reduce the number of proposed values to c, where c is a small constant. The second communication abstraction allows each non-faulty process to compute a set of (proposed) values satisfying the following property: if the set of a non-faulty process is a singleton containing value v, the set of any non-faulty process contains v. Both communication abstractions have an $O(n^2)$ message complexity and a constant time complexity. The reduction of multivalued Byzantine consensus to binary Byzantine consensus is then a simple sequential use of these communication abstractions. To the best of our knowledge, this is the first asynchronous message-passing algorithm that reduces multivalued consensus to binary consensus with $O(n^2)$ messages and constant time complexity (measured with the longest causal chain of messages) in the presence of up to $t < n/3t < n/3$ Byzantine processes, and without using cryptography techniques. Moreover, this reduction algorithm uses a single instance of the underlying binary consensus, and tolerates message re-ordering by Byzantine processes. This work, done in collaboration with Achour Mostefaoui from LS2N (Nantes), appeared in Acta Informatica [20].

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

Participant: Michel Raynal.

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

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

Participants: Michel Raynal, François Taïani.

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

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

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

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

Participant: George Giakkoupis.

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

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

6.2. Network and Graph Algorithms

6.2.1. Tight bounds on vertex connectivity under sampling

Participant: George Giakkoupis.

A fundamental result by Karger (SODA 1994) states that for any λ-edge-connected graph with n nodes, independently sampling each edge with probability $p = \Omega(\log(n)/\lambda)$ results in a graph that has edge connectivity $\Omega(\lambda p)$, with high probability. In [15], we proved the analogous result for vertex connectivity, when either vertices or edges are sampled. We showed that for any k-vertex-connected graph G with n nodes, if each node is independently sampled with probability $p = \Omega(\sqrt{\log(n)/k})$, then the subgraph induced by the sampled nodes has vertex connectivity $\Omega(kp^2)$, with high probability. If edges are sampled with probability $p = \Omega(\log(n)/k)$then the sampled subgraph has vertex connectivity $\Omega(kp)$, with high probability. Both bounds are existentially optimal.

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

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

Participant: George Giakkoupis.

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

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

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

6.3. Scalable Systems

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

Participants: Anne-Marie Kermarrec, François Taïani.

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

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

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

Participants: Resmi Ariyattu Chandrasekharannair, François Taïani.

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

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

Participants: Simon Bouget, David Bromberg, François Taïani.

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

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

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

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

Participants: Simon Bouget, David Bromberg, François Taïani.

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

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

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

Participant: Davide Frey.

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

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

6.3.6. Designing Overlay Networks for Decentralized Clouds

Participant: Marin Bertier.

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

COAST Project-Team

5. New Results

5.1. Design and Analysis of Collaborative Editing Approaches

Participants: Matthieu Nicolas, Victorien Elvinger, Hoai Le Nguyen, Quentin Laporte Chabasse, Claudia-Lavinia Ignat [contact], Gérald Oster, François Charoy, Olivier Perrin.

Since the Web 2.0 era, the Internet is a huge content editing place on which users collaborate. Such shared content can be edited by thousands of people. However, current consistency maintenance algorithms seem not to be adapted to massive collaborative updating involving large amount of contributors and a high velocity of changes. This year we designed new optimistic replication algorithms for maintaining consistency for complex data such as wikis. We also designed a peer-to-peer web-based real-time collaborative editor relying on our proposed algorithms as well as a mechanism that balances awareness and disturbance in this kind of systems. We also started to study collaborative editing user behavior.

Wikis are one of the most important tools of Web 2.0 allowing users to easily edit shared data. However, wikis offer limited support for merging concurrent contributions on the same pages. Users have to manually merge concurrent changes and there is no support for an automatic merging. Real-time collaborative editing reduces the number of conflicts as the time frame for concurrent work is very short. We proposed extending wiki systems with real-time collaboration and designed an automatic merging solution adapted for rich content wikis [2]. Our merging solution is based on an operational transformation approach for which we defined operations with high-level semantic capturing user intentions when editing wiki content such as move, merge and split. Our solution is the first one that deals with high level operations, existing approaches being limited to operations of insert, delete and update on textual documents.

Existing real-time collaborative editors rely on a central authority that stores user data which is a perceived privacy threat.We designed MUTE [8], a peer-to-peer web-based real-time collaborative editor that eliminates the disadvantages of central authority based systems. Users share their data with the collaborators they trust without having to store their data on a central place. MUTE features high scalability and supports offline and ad-hoc collaboration. MUTE relies on LogootSplit, a CRDT-based consistency maintenance algorithm for strings [15]. MUTE collaborative editor will be integrated in the virtual desktop of OpenPaaS::NG project [8].

When people work collaboratively on a shared document, they have two contradictory requirements on their editors that may affect the efficiency of their work. On the one hand, users would like to be aware of other users work on a particular part of the document. On the other hand, users would like to focus their attention on their own current work, with as little disturbance from the concurrent activities as possible. We designed a mechanism that lets users handle a balance between disturbance and awareness of concurrent updates [10]. Users can define focus regions and concentrate on the work in these regions without being disturbed by work of other users. Occasionally, users can preview concurrent updates and select a number of these updates to be integrated into the local copy.

We are interested in analysing user behavior during collaborative editing. This year we studied concurrency and conflicts in asynchronous collaboration [7]. We chose to study collaboration traces of distributed version control systems such as Git. We analysed Git repositories of four projects: Rails, IkiWiki, Samba and Linux Kernel. We analyzed the collaboration process of these projects at specific periods revealing how changes integration evolves during project development. We also analyzed how often users decide to rollback to previous document version when the integration process results in conflict. Finally, we studied the mechanism adopted by Git to consider changes made on two continuous lines as conflicting.

5.2. Trust-based Collaboration

Participants: Quang Vinh Dang, Claudia-Lavinia Ignat, Francois Charoy, Olivier Perrin, Mohammed Riyadh Abdmeziem, Hoang Long Nguyen.

Trust between users is an important factor for the success of a collaboration. Users might want to collaborate only with those users they trust. We are interested in assessing users trust according to their behaviour during collaboration in a large scale environment. In order to compute the trust score of users according to their contributions during a collaborative editing task, we need to evaluate the quality of the content of a document that has been written collaboratively. We investigated how to automatically assess the quality of Wikipedia articles in order to provide guidance for both authors and readers of Wikipedia. Most existing approaches for quality classification of Wikipedia articles rely on traditional machine learning with manual feature engineering, which requires a lot of expertise and effort and is language dependent. We proposed an approach that addresses the trade-off between accuracy, time complexity and language independence for the prediction models [5]. Our approach relying on Recurrent Neural Networks (RNN) eliminates disadvantages of feature engineering, i.e. it learns directly from raw data without human intervention and is language-neutral. Experimental results on English, French and Russian Wikipedia datasets show that our approach outperforms state-of-the-art solutions.

Rating prediction is a key task of e-commerce recommendation mechanisms. Recent studies in social recommendation enhance the performance of rating predictors by taking advantage of user relationships. However, these prediction approaches mostly rely on user personal information which is a privacy threat. We proposed dTrust [6], a simple social recommendation approach that avoids using user personal information. It relies uniquely on the topology of an anonymized trust-user-item network that combines user trust relations with user rating scores for items. This topology is fed into a deep feed-forward neural network. Experiments on real-world data sets showed that dTrust outperforms state-of-the-art in terms of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) scores for both warm-start and cold-start problems.

One dimension of our work is dedicated to ensure consistency of the key server. We design Trusternity, which is a secure, scalable auditing mechanism using a blockchain to replace the gossiping mechanism of transparent log system. We have implemented Trusternity as a proof-of-concept, and we have led some evaluation about the detection of malicious behavior on the blockchain network.

Securing P2P collaborative system remains a critical issue for its widespread adoption. One of our goals is to ensure that communication between collaborating partner is secure from end to end. We need to encrypt exchange of operations among partners. For that we propose to rely on group keys management. One of the issue is that the composition of the partnership can change and this require to change the group key. Since we don't want a central server to manage keys, that would break the p2p nature of our approach we need to propose group key management protocols that are resilient to change in groups, even in group of large size. [3]

5.3. Cloud Provisioning for Elastic BPM

Participants: François Charoy, Samir Youcef, Guillaume Rosinosky.

Cloud computing provider do not help consumer to use optimally the available resources. For this, several approaches have been proposed [24] that take benefit from the elasticity of the Cloud, starting and stoppping virtual machines on demand. They suffer from several shortcomings. Often they consider only one objective, the reduction of the cost, or a level of quality of service. We proposed to optimize two conflicting objectives, the number of migrations of tenants that is helpful to reach the optimal cost and the cost incurred considering a set of resources. Our approach allows to take into account the multi-tenancy property and the Cloud computing elasticity, and is efficient as shown by an extensive experimentation based on real data from Bonita BPM customers [9].

5.4. Risk Management for the Deployment of a Business Process in a Multi-Cloud Context

Participants: Amina Ahmed-Nacer, Claude Godart, Samir Youcef.

The lack of trust in cloud organizations is often seen as braking forces to SaaS developments. This work proposes an approach which supports a trust model and a business process model in order to allow the orchestration of trusted business process components in the cloud.

The contribution is threefold and consists in a method, a model and a framework. The method categorizes techniques to transform an existing business process into a risk-aware process model that takes into account security risks related to cloud environments. These techniques are partially described in the form of constraints to automatically support process transformation. The model formalizes the relations and the responsibilities between the different actors of the cloud. This allows to identify the different information required to assess and quantify security risks in cloud environments.

The framework is a comprehensive approach that decomposes a business process into fragments that can automatically be deployed on multiple clouds. The framework also integrates a selection algorithm that combines the security information of cloud offers and of the process with other quality of service criteria to generate an optimized configuration. It is implemented in a tool to assess cloud providers and decompose processes.

Rooted in past years work, we are contributing this year at the methodological and framework levels in two directions:

- At the methodological level, while our risk computing model rested previously only on data provided by cloud providers (provider-side risk model), we are developing a risk model integrating client-side knowledge (client-side risk model) [4].
- Additionally are developing a simulation tool for supporting designer decision with the ability to balance risk with cost when selecting the best cloud configuration.

5.5. Scheduling and Resource Allocation in Business Processes

Participants: Khalid Benali, Abir Ismaili-Alaoui.

Business Process Management (BPM) is concerned with continuously enhancing business processes by adapting a systematic approach that enables companies to increase the performance of their existing business processes and achieve their business goals. Business processes are generaly considered as blind and stateless, which mean that in each business process execution results from past process instances are not taken into consideration.

The main objective of our current research is to exploit the data generated from previous instances in order to enhance business processes in regards with several aspects, such as improvement of process business logical correctness, optimization of business process modeling issues, or improvemment of resource allocation and scheduling procedure in order to particularly optimize costs and time (among other factors).

We focus currently on this last aspect, i.e. scheduling and resource allocation in business processes. Business Processes may contain automatic tasks and non automatic tasks, so managing resources depends on the type of those resources (human or machine) In this context, our work use machine learning techniques to analyze data generated from previous business process execution to improve business process scheduling. This step ensure the assignment of the most critical business process instance task to a qualified (and may be costly) human resource while minimizing global execution costs through assignement of “dummy” tasks to machine agents.

CTRL-A Project-Team

6. New Results

6.1. Programming support for Autonomic Computing

6.1.1. Reactive languages

Participants: Gwenaël Delaval, Eric Rutten.

Our work in reactive programming for autonomic computing systems is focused on the specification and compilation of declarative control objectives, under the form of contracts, enforced upon classical mode automata as defined in synchronous languages. The compilation involves a phase of Discrete Controller Synthesis in order to obtain an imperative executable code. The programming language Heptagon / BZR (see Section Software and Platforms) integrates our research results [7].

Recent work concerns exploring new possibilities offered by logics-numeric control. We target the problem of the safe control of reconfigurations in component-based software systems (see also Section 6.1.2 for the component-based aspects), where strategies of adaptation to variations in both their environment and internal resource demands need to be enforced. In this context, the computing system involves software components that are subject to control decisions. We approach this problem under the angle of Discrete Event Systems (DES), involving properties on events observed during the execution (e.g., requests of computing tasks, work overload), and a state space representing different configurations such as activity or assemblies of components. We consider in particular the potential of applying novel logico-numerical control techniques to extend the expressivity of control models and objectives, thereby extending the application of DES in component-based software systems. We elaborate methodological guidelines for the application of logico-numerical control based on a case-study, and validate the result experimentally.

This work is in cooperation with the Sumo team at Inria Rennes and University of Liverpool, and is published in the CCTA 2017 conference [15].

6.1.2. Component-based approaches

Participants: Gwenaël Delaval, Eric Rutten.

Our work in component-based programming for autonomic computing systems as examplified by e.g., FRACTAL, considers essentially the problem of specifying the control of components assembly reconfiguration, with an approach based on the integration within such a component-based framework of a reactive language as in Section 6.1.1 [6].

Dynamic reconfiguration is a key capability of Component-based Software Systems to achieve self-adaptation as it provides means to cope with environment changes at runtime. The space of configurations is defined by the possible assemblies of components, and navigating this space while achieving goals and maintaining structural properties is managed in an autonomic loop. The natural architectural structure of component-based systems calls for hierarchy and modularity in the design and implementation of composites and their managers, and requires support for coordinated multiple autonomic loops. [1] [12].

In recent work, we leverage the modularity capability to strengthen the Domain-Specific Language (DSL) Ctrl-F, targeted at the design of autonomic managers in component-based systems. Its original definition involved discrete control-theoretical management of reconfigurations, providing assurances on the automated behaviors. The objective of modularity is two-fold: from the design perspective, it allows designers to seamlessly decompose a complex system into smaller pieces of reusable architectural elements and adaptive behaviours. From the compilation point of view, we provide a systematical and generative approach to decompose control problems described in the architectural level while relying on mechanisms of modular Discrete Control Synthesis (DCS), which allows us to cope with the combinatorial complexity that is inherent to DCS problems. We show the applicability of our approach by applying it to the self-adaptive case study of the existing RUBiS/Brownout eBay-like web auction system.

This work is done in cooperation with Inria teams ASCOLA in Nantes and SPIRALS in Lille, and is published in the Journal of Systems and Software [12] and the SeAC 2017 - 2nd Workshop on Self-Aware Computing, a satellite of the ICAC'17 conference [14].

We are also considering integration at the DSL level of expressivity extensions, for which the compilation and controller synthesis is relying on the ReaX tool developed at Inria Rennes, in the Sumo team as mentioned in Section 6.1.1 [15].

6.1.3. Rule-based systems

Participants: Adja Sylla, Gwenaël Delaval, Eric Rutten.

This work concerns a high-level language for safe rule-based programming in the LINC transactional rule-based platform developed at CEA [17]. Rule based middlewares such as LINC enable high level programming of distributed adaptive systems behaviours. LINC also provides the systems with transactional guarantees and hence ensures their reliability at runtime. However, the set of rules may contain design errors (e.g. conflicts, violations of constraints) that can bring the system in unsafe safe or undesirables states, despite the guarantees provided by LINC. On the other hand, automata based languages such as Heptagon/BZR enable formal verification and especially synthesis of discrete controllers to deal with design errors. Our work studies these two languages and combines their execution mechanisms, from a technical perspective. We target applications to the domain of Internet of Things and more particularly smart building, office or home (see Section 6.2.2.1).

This work is in cooperation with CEA LETI/DACLE, it is the topic of the PhD of Adja Sylla at CEA, co-advised with M. Louvel [11], and aspects on Software Engineering and Software Architecture for Multiple Autonomic Loops are published in the ICCAC 2017 conference [18].

6.1.4. A Language for the Smart Home

Participant: Gwenaël Delaval.

This work is about the design of the CCBL programming language (Cascading Contexts Based Language), an end-user programming language dedicated to Smart Home. CCBL has been proposed to avoid the problems encountered by end-users programming with ECA (Event Conditions Actions), which is the dominant approach in the field. This language has been evaluated by means of a user-based experiment where 21 adults (11 experimented programmers and 10 non-programmers) have been asked to express four increasingly complex behaviors using both CCBL and ECA. It has been shown that significantly less errors were made using CCBL than using ECA. From this experiment, some categorization and explanation of the errors made when using ECA have been proposed, with explanations about why users avoid these errors when programming with CCBL. Finally, error reporting for CCBL have been explored by identifying two specific errors and by developing a solution based on Heptagon and ReaX to detect them in CCBL programs.

This work is done in cooperation with the IIHM team of LIG (Alexandre Demeure), in the framework of a LIG « projet émergence » and was the topic of the MSc internship of Lénaïg Terrier [22].

6.2. Design methods for reconfiguration controller design in computing systems

We apply the results of the previous axes of the team's activity to a range of infrastructures of different natures, but sharing a transversal problem of reconfiguration control design. From this very diversity of validations and experiences, we draw a synthesis of the whole approach, towards a general view of Feedback Control as MAPE-K loop in Autonomic Computing [20] [19].

6.2.1. High-Performance Computing

Participants: Soguy Mak Kare Gueye, Gwenaël Delaval, Stéphane Mocanu, Bogdan Robu, Eric Rutten.

6.2.1.1. Towards a Control-Theory based approach for cluster overload avoidance

This work is addressing the problem of automated resource management in an HPC infrastructure, using techniques from Control Theory to design a controller that maximizes cluster utilization while avoiding overload. We put in place a mechanism for feedback (Proportional Integral, PI) to system software, through a maximum number of jobs to be sent to the cluster, in response to system information about the current number jobs processed.

This work is done in cooperation with the Datamove team of Inria/LIG, and Gipsa-lab. It was the topic of the internship of Emmanuel Stahl for the Grenoble INP ENSE^3 engineering school, [21].

6.2.1.2. Reconfiguration control in DPR FPGA

6.2.1.2.1. DPR FPGA and discrete control for reconfiguration

Implementing self-adaptive embedded systems, such as UAV drones, involves an offline provisioning of the several implementations of the embedded functionalities with different characteristics in resource usage and performance in order for the system to dynamically adapt itself under uncertainties. FPGA-based architectures offer for support for high flexibility with dynamic partial reconfiguration (DPR) features. We propose an autonomic control architecture for self-adaptive and self-reconfigurable FPGA-based embedded systems. The control architecture is structured in three layers: a mission manager, a reconfiguration manager and a scheduling manager. In this work we focus on the design of the reconfiguration manager. We propose a design approach using automata-based discrete control. It involves reactive programming that provides formal semantics, and discrete controller synthesis from declarative objectives.

This work is in the framework of the ANR project HPeC (see Section 8.2.1), and is published in the AHS 2017 conference [16].

6.2.1.2.2. Mission management and stochastic control

In the Mission Management workpackage of the ANR project HPeC, a concurrent control methodology is constructed for the optimal mission planning of a U.A.V. in stochastic environnement. The control approach is based on parallel ressource sharing Partially Observable Markov Decision Processes modeling of the mission. The parallel POMDP are reduced to discrete Markov Decision Models using Bayesian Networks evidence for state identification. The control synthesis is an iterative two step procedure : first MDP are solved for the optimisation of a finite horizon cost problem ; then the possible ressource conflicts between parallel actions are solved either by a priority policy or by a QoS degradation of actions, e.g., like using a lower resolution version of the image processing task if the ressource availability is critical.

6.2.2. IoT

Participants: Neïl Ayeb, Adja Sylla, Gwenaël Delaval, Stéphane Mocanu, Eric Rutten.

6.2.2.1. Control of smart buildings

A smart environment is equipped with numerous devices (i.e., sensors, actuators) that are possibly distributed over different locations (e.g., rooms of a smart building). These devices are automatically controlled to achieve different objectives related, for instance, to comfort, security and energy savings. Controlling smart environment devices is not an easy task. This is due to: the heterogeneity of devices, the inconsistencies that can result from communication errors or devices failure, and the conflicting decisions including those caused by environment dependencies.

Our work proposes a design framework for the reliable and environment aware management of smart environment devices. The framework is based on the combination of the rule based middleware LINC and the automata based language Heptagon/BZR (H/BZR). It consists of: an abstraction layer for the heterogeneity of devices, a transactional execution mechanism to avoid inconsistencies and a controller that, based on a generic model of the environment, makes appropriate decisions and avoids conflicts. A case study with concrete devices, in the field of building automation, is presented to illustrate the framework.

This work is in the framework of the cooperation with CEA (see Section 7.1), and is published in the Springer Journal of Internet Services and Applications, with recognized editors from the Middleware community [13]

6.2.2.2. Device management

The research topic is targeting an adaptative and decentralized management for the IoT. It will contribute design methods for processes in virtualized gateways in order to enhance IoT infrastructures.

More precisely, it concerns Device Management in the case of large numbers of connected sensors and actuators, as can be found in Smart Home and Building, Smart Electricity grids, and industrial frameworks as in Industry 4.0.

In contrast with a centralized management of such large sets of devices, for the autonomic management of their adaptations, upgrades and other commands, the objective its to target a distributed management, enabling local decisions, by proposing an appropriate middleware framework. These local adjustments will be processed using context data. The context is a synchronized (i.e., always up-to-date with reality) description of concepts and relations. Technically, the context data information are extracted from multiple sources such as IT environment, user environment and physical environment.

This work is in the framework of the Inria/Orange labs joint laboratory (see Section 7.2.1), and supported by the CIFRE PhD thesis grant of Neïl Ayeb, starting dec. 2017.

6.2.2.3. Security in SCADA industrial systems

We focus mainly on vulnerability search, automatic attack vectors synthesis and intrusion detection. Model checking techniques are used for vulnerability search and automatic attack vectors construction. Intrusion detection is mainly based on process-oriented detection with a technical approach from run-time monitoring. The LTL formalism is used to express safety properties which are mined on an attack-free dataset. The resulting monitors are used for fast intrusion detections.

A demonstrator of attack/defense scenarion in SCADA systems will be built on the existing G-ICS lab (hosted by ENSE3/Grenoble-INP).

This work is in the framework of the ANR project Sacade on cybersecurity of industrial systems (see Section 8.2.2),

MIMOVE Team

7. New Results

7.1. Living with Interpersonal Data: Observability and Accountability in the Age of Pervasive ICT

Participants: Murray Goulden (University of Nottingham), Peter Tolmie (University of Nottingham), Richard Mortier (University of Cambridge), Tom Lodge (University of Nottingham), Anna-Kaisa Pietilainen (Google), Renata Teixeira

The Internet of Things, alongside existing mobile digital technologies, herald a world in which pervasive sensing constantly captures data about us. Simultaneous with this technology programme are moves by policymakers to shore up the digital economy, through the legislating of new models of data management. These moves seek to give individuals control and oversight of their personal data. Within shared settings the consequences of these changes are the large-scale generation of interpersonal data, generated by and acting on the group rather than individual. We consider how such systems create new forms of observability and hence accountability amongst members of the home, and draw on the work of Simmel (1906) and Goffman (1971) to explore how these demands are managed. Such management mitigates the more extreme possibilities for domestic monitoring posited by these systems, yet without careful design there remains a considerable danger of unanticipated negative consequences.

7.2. Predicting the effect of home Wi-Fi quality on QoE

Participants: Diego da Hora (Telecom Paris Tech), Karel van Doorselaer (Technicolor), Koen van Oost (Technicolor), Renata Teixeira

Poor Wi-Fi quality can disrupt home users' internet experience, or the Quality of Experience (QoE). Detecting when Wi-Fi degrades QoE is extremely valuable for residential Internet Service Providers (ISPs) as home users often hold the ISP responsible whenever QoE degrades. Yet, ISPs have little visibility within the home to assist users. Our goal is to develop a system that runs on commodity access points (APs) to assist ISPs in detecting when Wi-Fi degrades QoE. Our first contribution is to develop a method to detect instances of poor QoE based on the passive observation of Wi-Fi quality metrics available in commodity APs (e.g., PHY rate). We use support vector regression to build predictors of QoE given Wi-Fi quality for popular internet applications. We then use K-means clustering to combine per-application predictors to identify regions of Wi-Fi quality where QoE is poor across applications. We call samples in these regions as poor QoE samples. Our second contribution is to apply our predictors to Wi-Fi metrics collected over one month from 3,479 APs of customers of a large residential ISP. Our results show that QoE is good on the vast majority of samples of the deployment, still we find 11.6% of poor QoE samples. Worse, approximately 21% of stations have more than 25% poor QoE samples. In some cases, we estimate that Wi-Fi quality causes poor QoE for many hours, though in most cases poor QoE events are short.

7.3. Narrowing the gap between QoS metrics and Web QoE using Above-the-fold metrics

Participants: Diego da Hora (Telecom Paris Tech), Alemnew Sheferaw Asrese (Aalto University), Vassilis Christophides, Renata Teixeira, Dario Rossi (Telecom Paris Tech)

Page load time (PLT) is still the most common application Quality of Service (QoS) metric to estimate the Quality of Experience (QoE) of Web users. Yet, recent literature abounds with proposals for alternative metrics (e.g., Above The Fold, SpeedIndex and variants) that aim at better estimating user QoE. The main purpose of this work is thus to thoroughly investigate a mapping between established and recently proposed objective metrics and user QoE. We obtain ground truth QoE via user experiments where we collect QoS metrics over 3,000 Web accesses annotated with explicit user ratings in a scale of 1 to 5, which we make available to the community. In particular, we contrast domain expert models (such as ITU-T and IQX) fed with a single QoS metric, to models trained using our ground-truth dataset over multiple QoS metrics as features. Results of our experiments show that, albeit very simple, ex- pert models have a comparable accuracy to machine learning approaches. Furthermore, the model accuracy improves considerably when building per-page QoE models, which may raise scalability concerns as we discuss.

7.4. Performance Modeling of the Middleware Overlay Infrastructure of Mobile Things

Participants: Georgios Bouloukakis, Nikolaos Georgantas, Valérie Issarny.

Internet of Things (IoT) applications consist of diverse Things (sensors and devices) in terms of hardware resources. Furthermore, such applications are characterized by the Things' mobility and multiple interaction types, such as synchronous, asynchronous, and streaming. Middleware IoT protocols consider the above limitations and support the development of effective applications by providing several Quality of Service features. These features aim to enable application developers to tune an application by switching different levels of response times and delivery success rates. However, the profusion of the developed IoT protocols and the intermittent connectivity of mobile Things, result to a non-trivial application tuning. In this work, we model the performance of the middleware overlay infrastructure using Queueing Network Models. To represent the mobile Thing's connections/disconnections, we model and solve analytically an ON/OFF queueing center. We apply our approach to streaming interactions with mobile peers. Finally, we validate our model using simulations. The deviations between the performance results foreseen by the analytical model and the ones provided by the simulator are shown to be less than 5%.

7.5. USNB: Enabling Universal Online Social Interactions

Participants: Rafael Angarita, Nikolaos Georgantas, Valérie Issarny.

Online social network services (OSNSs) have become an integral part of our daily lives. At the same time, the aggressive market competition has led to the emergence of multiple competing siloed OSNSs that cannot interoperate. As a consequence, people face the burden of creating and managing multiple OSNS accounts and learning how to use them to stay connected. This work is concerned with relieving users from such a burden by enabling universal online social interactions. The contributions of this work span: (1) a model of the universal social network bus (USNB) for OSNS interoperability; (2) a prototype for universal online social interactions that builds upon the proposed model; and (3) a preliminary experimental evaluation involving 50 participants. Results show that people are positive about the solution as they are able to reach out a larger community of users independently of the OSNSs they use.

7.6. Opportunistic Multiparty Calibration for Robust Participatory Sensing

Participants: Françoise Sailhan, Valérie Issarny, Otto Tavares Nascimento.

While bringing massive-scale sensing at low cost, mobile participatory sensing is challenged by the low accuracy of the sensors embedded in and/or connected to the smartphones. The mobile measurements that are collected need to be corrected so as to accurately match the phenomena being observed. This paper addresses this challenge by introducing a multi-hop, multiparty calibration method that operates in the background in an automated way. Using our method, sensors that are within a relevant sensing (and communication) range coordinate so that the observations of the participating (previously) calibrated sensors serve calibrating the other participants. As a result, our method is particularly well suited for participatory sensing within crowd meetings, as as for instance within public spaces. Our solution leverages multivariate linear regression, together with robust regression so as to discard the measurements that are of too low quality for being meaningful. To the best of our knowledge, we are the first to introduce a multiparty calibration algorithm, while previous work in the area focused on pairwise calibration. This work further introduces a supporting prototype implemented over Android, and related experiment in the context of noise sensing. We show that the proposed multiparty calibration system enhances the accuracy of the mobile noise sensing application.

7.7. Extracting usage patterns of home IoT devices

Participants: Vassilis Christophides, Gevorg Poghosyan (Insight Centre for Data Analytics), Ioannis Pefkianakis (Hewlett Packard Labs), Pascal Le Guyadec (Technicolor)

We have initially investigated how data analytics for Machine-to-Machine (M2M) data (connectivity, performance, usage) produced by connected devices in residential Intranet of Things, could support novel *home automation services* that enrich the living experience in smart homes. We have investigated new data mining techniques that go beyond binary association rule mining for traditional market basket analysis, considered by previous works. We design a multidimensional pattern mining framework, which collects raw data from operational home gateways, it discretizes and annotates the raw data, it produces traffic usage logs which are fed in a multidimensional association rule miner, and finally it extracts home residents' habits. Using our analysis engine, we extract complex device co-usage patterns of 201 residential broadband users of an ISP, subscribed to a n-play service. Such fine-grained device usage patterns provide valuable insights for emerging use cases, such as adaptive usage of home devices (aka horizontal integration of things). Such use cases fall within the wider area of human-cognizant Machine-to-Machine communication aiming to predict user needs and complete tasks without users initiating the action or interfering with the service. While this is not a new concept, according to Gartner cognizant computing is a natural evolution of a world driven not by devices but collections of applications and services that span across multiple devices, in which human intervention becomes as little as possible, by analyzing past human habits. To realize this vision, we are interested in co-usage patterns featuring spatio-temporal information regarding the context under which devices have been actually used in homes. For example, a network extender which is currently turned off, could be turned on at a certain day period (e.g., evening) when it has been observed to be highly used along with other devices (e.g., a laptop or a tablet). Alternatively, the identification of frequent co-usage of particular devices at a home (say iPhone with media player), could be used by a things' recommender to advertise the same set of devices at another home (say another iPhone user could be interested in a media player).

MYRIADS Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Fog Computing

Participants: Guillaume Pierre, Arif Ahmed, Ali Fahs, Alexandre Van Kempen, Salsabil Amri, Vinothkumar Nagasayanan, Berenger Nguyen Nhon.

Fog computing aims to extend datacenter-based cloud platforms with additional compute, networking and storage resources located in the immediate vicinity of the end users. By bringing computation where the input data was produced and the resulting output data will be consumed, fog computing is expected to support new types of applications which either require very low network latency (e.g., augmented reality applications) or which produce large data volumes which are relevant only locally (e.g., IoT-based data analytics).

Fog computing architectures are fundamentally different from those of traditional cloud platforms: to provide computing resources in physical proximity of any end user, fog computing platforms must necessarily rely on very large numbers of small Points-of-Presence connected to each other with commodity networks whereas clouds are typically organized with a handful of extremely powerful data centers connected by dedicated ultra-high-speed networks. This geographical spread also implies that the machines used in any Point-of-Presence may not be datacenter-grade servers but much weaker commodity machines.

We investigated the challenges of efficiently deploying Docker containers in fog platforms composed of tiny single-board computers such as Raspberry PIs. This operation can be painfully slow, in the order of multiple minutes depending on the container's image size and network condition. We showed that this bad performance is not only due to hardware limitations, but it is largely due to inefficiencies in the way Docker implements the container's image download operation. We proposed a number of optimization techniques which , when combined together, make container deployment up to 4 times faster than the vanilla Docker implementation. A publication on this topic is under submission.

Although fog computing infrastructures are fundamentally distributed, their management part still remains centralized: a single node (or small group of nodes) is in charge of maintaining the list of available server machines, monitoring them, distributing software to them, deciding which server must take care of which task, etc. We therefore aim to reduce the discrepancy between the broadly distributed compute/storage resources and the – currently – extremely centralized control of these resources, by focusing first on the resource scheduling function. This project has just started, and we expect to obtain the first results in 2018.

7.1.2. Edge Cloud

Participants: Anne-Cécile Orgerie, Cédric Tedeschi, Matthieu Simonin, Ehsan Ahvar, Genc Tato.

Myriads is involved in the Discovery project, whose goal is to design, develop and experiment a software stack for a distributed cloud platform where resources are directly injected into the backbone of the network [60]. To this end, we designed a novel family of overlay network to operate messaging and routing on top of such a distributed utility computing platform. The big picture of these overlays was described in a workshop [47].

7.1.3. Community Clouds

Participant: Jean-Louis Pazat.

In this work we consider an infrastructure based on devices (such as Internet boxes and NAS) owned and operated by end-users. A typical use-case is the sharing of CPU and storage capabilities by a community of users. This sharing is operated by hosting services to local and remote users. The devices of this distributed infrastructure have heterogeneous capabilities and no guaranteed availability. It is therefore challenging to ensure to the guest service a minimal hosting service level, such as availability or QoS.

We consider services build as an application based on micro-services. Such an application is deployed on the infrastructure by instantiating its constituant micro-services on some devices. One micro-services may rely on others micro-services to enable its own service. The performance of the resulting application is therefore highly dependent from the placement for each micro-service instance. Device parameters like CPU capabilities or network bandwidth and latency have a significant impact on the resulting response time of the micro-service, hence the application.

We explore solutions to adapt the placement of the micro-services to the capabilities of the infrastructure. As a first step, we are studying a static system where these capabilities are not variating. The placement decision can be expressed as the solution of an NP-Complete optimization problem. We have shown that a solution for this problem can be found with reasonably good precision using a meta-heuristic called Particle Swarm Optimization. The next step will be to study how this solution can be adapted in a dynamic system by considering the variations of the CPU and Network parameters and the availability of the devices.

This work in done in the context of Bruno Stevant's PhD thesis co-advised by Jean-Louis Pazat (Bruno Stevant is a member of OCIF team).

7.1.4. Evaluation of Data Stream Processing Frameworks in Clouds

Participants: Christine Morin, Deborah Agarwal, Subarna Chatterjee.

We address the problem of selecting a correct stream processing framework for a given application to be executed within a specific physical infrastructure. For this purpose, we have performed a thorough comparative analysis of three data stream processing platforms – Apache Flink, Apache Storm, and Twitter Heron (the enhanced version of Apache Storm), that are chosen based on their potential to process both streams and batches in real-time. For the comparative performance analysis of the chosen platforms, we have experimented using 8-node clusters on Grid5000 experimentation testbed and have selected a wide variety of applications ranging from a conventional benchmark (word count application) to sensor-based IoT application (air quality monitoring application) and statistical batch processing application (flight delay analysis application). The work focuses to analyze the performance of the frameworks in terms of the volume and throughput of data streams that each framework can possibly handle. The impact of each framework on the operating system is analyzed by experimenting and studying the resource utilization of the platforms in terms of CPU utilization, memory consumption. The energy consumption of the platforms is also studied to understand the suitability of the platforms towards green computing. Last, but not the least, the fault tolerance of the frameworks is also studied and analyzed. Lessons learnt from this work will precisely enlighten IaaS cloud end-users to wisely choose the correct streaming platform in order to run a particular application within a given set of VMs and will assist the cloud-providers to rationally allocate VMs equipped with a particular stream processing framework to PaaS cloud-users for running a specific streaming application. A paper has been submitted to an international conference in November 2017.

7.1.5. Stream Processing for Maritime Surveillance

Participants: Pascal Morillon, Christine Morin, Matthieu Simonin, Cédric Tedeschi.

In the context of maritime surveillance, and of the Sesame Project, we started the design and implementation of a platform dedicated to the batch and real-time processing of AIS messages sent by ships to inform about their identity, position and destination among other pieces of information.

Having use cases in mind such as detecting ships entering a protected areas, or ships having suspect behaviors, we designed a software architecture able to process AIS messages and produce synthetic data so as to answer these questions.

First experiments using a preliminary version of this platform have been conducted over the Grid'5000 platform using an archive of one-month of the AIS messages collected globally during March 2017. In particular, we've been able to index these messages using ElasticSearch [0] and visualize them using Kibana [0].

[0] https://www.elastic.co/fr/
[0] https://www.elastic.co/products/kibana

The architecture has been described in a poster presented at BiDS'17 [56].

7.1.6. Adaptive deployment for multi-cloud applications

Participants: Nikos Parlavantzas, Manh Linh Pham.

This work builds on the Adapter system, developed in the context of the PaaSage European project (2012-2016). The Adapter is part of the PaaSage open-source platform, a holistic solution for supporting the automatic deployment and execution of multi-cloud applications. Specifically, the Adapter is responsible for dynamic, cross-cloud application adaptation, taking into account adaptation costs and benefits in making deployment decisions. In 2017, we improved the Adapter and performed a comprehensive evaluation using experiments in a multi-cloud environment. The results demonstrate that Adapter supports automated multi-cloud adaptation while optimizing the performance and cost of the application. The results are described in an article currently under submission.

7.1.7. Application configuration and reconfiguration in multi-cloud environments

Participant: Nikos Parlavantzas.

Current approaches to cloud application configuration and reconfiguration are typically platform dependent, error prone and provide little support for optimizing application performance and resource utilisation. To address these limitations, we are combining the use of software product lines (SPLs) with performance prediction and automatic adaptation techniques. This work is performed in the context of the thesis of Carlos Ruiz Diaz, a PhD student at the University of Guadalajara, co-advised by Nikos Parlavantzas. The work has produced an SPL-based framework supporting initial configuration and dynamic adaptation in a systematic, platform-independent way.

In 2017, we extended the framework with a proactive adaptation solution that performs vertical VM scaling based on predictions of resource utilisation and performance. The solution targets multi-tier applications deployed on IaaS clouds. Experimental results demonstrate that the solution maintains expected application performance while reducing resource waste [46].

7.1.8. Adaptive resource management for high-performance, multi-sensor systems

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

In the context of our collaboration with Thales Research and Technology and Baptiste Goupille-Lescar's PhD work, we are applying cloud resource management techniques to high-performance, multi-sensor, embedded systems with real-time constraints. The objective is to increase the flexibility and efficiency of resource allocation in such systems, enabling the execution of dynamic sets of applications with strict QoS requirements.

In 2017, we focused on an industrial use case concerning the operation of a multi-function surface active electronically scanned array (AESA) radar. We developed a simulation environment using an industrial high-precision AESA simulator and the Ptolemy II simulation framework, and we are using this environment to explore and evaluate different dynamic application placement solutions [57].

7.2. Greening Clouds

ICT (Information and Communications Technologies) ecosystem now approaches 6% of world electricity consumption and this ICT energy use will continue grow fast because of the information appetite of Big Data, big networks and big infrastructures as Clouds that unavoidably leads to big power.

7.2.1. Energy Models

Participants: Ehsan Ahvar, Loic Guegan, Anne-Cécile Orgerie, Martin Quinson.

Cloud computing allows users to outsource the computer resources required for their applications instead of using a local installation. It offers on-demand access to the resources through the Internet with a pay-as-you-go pricing model.However, this model hides the electricity cost of running these infrastructures.

The costs of current data centers are mostly driven by their energy consumption (specifically by the air conditioning, computing and networking infrastructure). Yet, current pricing models are usually static and rarely consider the facilities' energy consumption per user. The challenge is to provide a fair and predictable model to attribute the overall energy costs per virtual machine and to increase energy-awareness of users. We aim at proposing such energy cost models without heavily relying on physical wattmeters that may be costly to install and operate.

Another goal consists in better understanding the energy consumption of computing and networking resources of Clouds in order to provide energy cost models for the entire infrastructure including incentivizing cost models for both Cloud providers and energy suppliers. These models will be based on experimental measurement campaigns on heterogeneous devices. Inferring a cost model from energy measurements is an arduous task since simple models are not convincing, as shown in our previous work. We aim at proposing and validating energy cost models for the heterogeneous Cloud infrastructures in one hand, and the energy distribution grid on the other hand. These models will be integrated into simulation frameworks in order to validate our energy-efficient algorithms at larger scale.

7.2.2. Exploiting Renewable Energy in Clouds

Participants: Benjamin Camus, Yunbo Li, Anne-Cécile Orgerie.

The development of IoT (Internet of Things) equipment, the popularization of mobile devices, and emerging wearable devices bring new opportunities for context-aware applications in cloud computing environments. The disruptive potential impact of IoT relies on its pervasiveness: it should constitute an integrated heterogeneous system connecting an unprecedented number of physical objects to the Internet. Among the many challenges raised by IoT, one is currently getting particular attention: making computing resources easily accessible from the connected objects to process the huge amount of data streaming out of them.

While computation offloading to edge cloud infrastructures can be beneficial from a Quality of Service (QoS) point of view, from an energy perspective, it is relying on less energy-efficient resources than centralized Cloud data centers. On the other hand, with the increasing number of applications moving on to the cloud, it may become untenable to meet the increasing energy demand which is already reaching worrying levels. Edge nodes could help to alleviate slightly this energy consumption as they could offload data centers from their overwhelming power load and reduce data movement and network traffic. In particular, as edge cloud infrastructures are smaller in size than centralized data center, they can make a better use of renewable energy.

We propose to investigate the end-to-end energy consumption of IoT platforms. Our aim is to evaluate, on concrete use-cases, the benefits of edge computing platforms for IoT regarding energy consumption. We aim at proposing end-to-end energy models for estimating the consumption when offloading computation from the objects to the edge or to the core Cloud, depending on the number of devices and the desired application QoS, in particular trading-off between performance (response time) and reliability (service accuracy).

7.2.3. Smart Grids

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

We propose exploiting Smart Grid technologies to come to the rescue of energy-hungry Clouds. Unlike in traditional electrical distribution networks, where power can only be moved and scheduled in very limited ways, Smart Grids dynamically and effectively adapt supply to demand and limit electricity losses (currently 10% of produced energy is lost during transmission and distribution).

For instance, when a user submits a Cloud request (such as a Google search for instance), it is routed to a data center that processes it, computes the answer and sends it back to the user. Google owns several data centers spread across the world and for performance reasons, the center answering the user's request is more likely to be the one closest to the user. However, this data center may be less energy efficient. This request may have consumed less energy, or a different kind of energy (renewable or not), if it had been sent to this further data center. In this case, the response time would have been increased but maybe not noticeably: a different trade-off between quality of service (QoS) and energy-efficiency could have been adopted.

While Clouds come naturally to the rescue of Smart Grids for dealing with this big data issue, little attention has been paid to the benefits that Smart Grids could bring to distributed Clouds. To our knowledge, no previous work has exploited the Smart Grids potential to obtain and control the energy consumption of entire Cloud infrastructures from underlying facilities such as air conditioning equipment (which accounts for 30% to 50% of a data center's electricity bill) to network resources (which are often operated by several actors) and to computing resources (with their heterogeneity and distribution across multiple data centers). We aim at taking advantage of the opportunity brought by the Smart Grids to exploit renewable energy availability and to optimize energy management in distributed Clouds.

7.2.4. Involving Users in Energy Saving

Participants: David Guyon, Christine Morin, Anne-Cécile Orgerie.

In a Cloud moderately loaded, some servers may be turned off when not used for energy saving purpose. Cloud providers can apply resource management strategies to favor idle servers. Some of the existing solutions propose mechanisms to optimize VM scheduling in the Cloud. A common solution is to consolidate the mapping of the VMs in the Cloud by grouping them in a fewer number of servers. The unused servers can then be turned off in order to lower the global electricity consumption.

Indeed, current work focuses on possible levers at the virtual machine suppliers and/or services. However, users are not involved in the choice of using these levers while significant energy savings could be achieved with their help. For example, they might agree to delay slightly the calculation of the response to their applications on the Cloud or accept that it is supported by a remote data center, to save energy or wait for the availability of renewable energy. The VMs are black boxes from the Cloud provider point of view. So, the user is the only one to know the applications running on her VMs.

We plan to explore possible collaborations between virtual machine suppliers, service providers and users of Clouds in order to provide users with ways of participating in the reduction of the Clouds energy consumption. This work will follow two directions: 1) to investigate compromises between power and performance/service quality that cloud providers can offer to their users and to propose them a variety of options adapted to their workload; and 2) to develop mechanisms for each layer of the Cloud software stack to provide users with a quantification of the energy consumed by each of their options as an incentive to become greener.

Our results were published in [40], [32], [31].

7.3. Securing Clouds

7.3.1. Security Monitoring in Clouds

Participants: Christine Morin, Jean-Louis Pazat, Louis Rilling, Anna Giannakou, Amir Teshome Wonjiga, Clément El Baz.

In the INDIC project we aim at making security monitoring a dependable service for IaaS cloud customers. To this end, we study three topics:

- defining relevant SLA terms for security monitoring,
- enforcing and verifying SLA terms,
- making the SLA terms enforcement mechanisms self-adaptable to cope with the dynamic nature of clouds.

The considered enforcement and verification mechanisms should have a minimal impact on performance.

In 2017, we did a thorough performance evaluation and security correctness analysis of the SAIDS approach, that we proposed in 2015, and that makes a network intrusion detection system (NIDS) deployed in a cloud operator infrastructure self-adaptable. In the performance evaluation we studied the performance impact of SAIDS on the cloud infrastructure operations related to the management of virtual machines (typically creation, migration, and deletion) as well as the scalability of SAIDS with respect to the number of NIDS devices managed. This performance evaluation was done on the Grid'5000 platform. The results showed that SAIDS adds very low overhead and is scalable. The security analysis was done both experimentally and based on a risk analysis. This analysis validated the security correctness of SAIDS. A full paper presenting SAIDS and its evaluation is submitted for publication in 2018. A demo of SAIDS was presented at FIC 2017, Lille, France in January 2017 and at the Inria Industry Days, Paris, France on October 17th, 2017.

Regarding SLA definition and enforcement, in 2017 we evaluated the verification method that we defined in 2016 and that enables a Cloud customer to verify that an NIDS located in the operator infrastructure is configured correctly according to the Service-Level Objectives (SLO) figuring in the SLA. The performance evaluation was done on the Grid'5000 platform and showed that the proposed verification method requires making a trade-off between verification speed and impact on the performance of the production applications deployed in the tenant's virtual machines. The security correctness analysis was based on a risk analysis and showed the constraints on the types of attacks that can be used for verification as well as the limitations due to the tools used in the prototype [55]. A full paper presenting the verification method and its evaluation is submitted for publication in 2018.

After the acquired experience on verifying security monitoring metrics, we started studying how to define relevant SLOs that are verifiable. We plan to get results in 2018 and submit a paper for publication in 2018 or 2019.

Finally, in October 2017 we started studying how security monitoring SLAs could take into account context changes like the evolution of threats and updates to the tenants' software.

Our work done as part of the INDIC project were presented in [59].

7.3.2. Risk assessment in clouds

Participant: Christine Morin.

Cloud providers have an incomplete view of their hosted virtual infrastructures managed by a Cloud Management System (CMS) and a Software Defined Network (SDN) controller. For various security reasons (e.g. isolation verification, modeling attack paths in the network), it is necessary to know which virtual machines can interact via network protocols. This requires building a connectivity graph between the virtual machines, that we can extract with the knowledge of the overall topology and the deployed network security policy. Existing methodologies for building such models for physical networks produce incomplete results. Moreover, they are not suitable for cloud infrastructures due to either their intrusiveness or lack of connectivity discovery. We propose a method to compute the connectivity graph, relying on information provided by both the CMS and the SDN controller. Connectivity can first be extracted from knowledge databases, then dynamically updated on the occurrence of cloud-related events. We realized an experimental evaluation of the proposed method to determine its correctness and performance in a realistic context, considering CPU and RAM consumption, the volume of data generated, and execution time for the different portions of the algorithm involved. Experiments were run on the Grid'5000 platform with OpenStack CMS and ONOS SDN controller. Our approach proves on a representative infrastructure to compute exact, complete and up-to- date connectivity graphs in reasonable time [42], [41].

7.3.3. Personal Data Management in Cloud-based IoT Systems

Participants: Christine Morin, Jean-Pierre Banâtre, Deborah Agarwal, Subhadeep Sarkar, Louis Rilling.

The Internet of Things (IoT), in today's digital world, encompasses billions of smart connected devices. These devices generate an unprecedented amount of data, which often bears sensitive personal information of individuals. In present service models, the data are processed and managed by service providers, beyond the visibility of the owner of the data. Although the EU General Data Protection Regulation (GDPR) strives to protect citizens and their data by regulation, citizens and service providers need technological advances to gain effective control over their data or to prove compliance with the new regulation. Our primary objective is to enforce, by design, the GDPR at the system level so as to preserve the privacy concerning personal data. We started off with enforcement of the data erasure facility as expressed in the GDPR. Data erasure corresponds to both automatic erasure of data after expiration of their retention period and ad-hoc on request of the data owner. Our first contribution, towards this, is design of a customizable privacy policy, which would allow the end users to express their preferences regarding the purpose of use, location of processing, retention period, sharing and storage policies concerning their personal data. We developed a XML-based policy expression language by defining the required data structures and vocabulary, which will facilitate the end-users to easily express their preferences. Next, we have investigated into the possible way of the implementation of the proposed solution and identified the exploitation of the operation system capabilities as an appropriate means to the cause. For this, we have potentially chosen the Sel4 (or may be some other capability-based microkernel) as our platform of operation. Finally, we have identified the different challenges towards implementation of our solution and did some groundwork towards proposing the solutions to the same. These challenges include efficient identification of replication of data, locating all replicas of a given data segment, and implementing erasure of data in a cross-domain service model.

7.4. Experimenting with Clouds

7.4.1. Simulation

Participants: Martin Quinson, Loic Guegan, Toufik Boubehziz, The Anh Pham.

We propose to combine two complementary experimental approaches: direct execution on testbeds such as Grid'5000, that are eminently believable but rather labor intensive, and simulations (using *e.g.* SimGrid) that are much more light-weighted, but requires are careful assessment. One specificity of the Myriads team is that we are working on these experimental methodologies *per se*, raising the standards of *good experiments* in our community. The Grid'5000 operational team is embedded in our research team, ensuring that our work remains aligned with the ground reality.

In 2017, our work was mostly centered on letting SimGrid become a *de facto* standard for the simulation of distributed platforms. We introduced a new programming interface, particularly adapted to the study of abstract algorithms. Beyond the engineering task, this requires to carefully capture the concepts that are important to the practitioners on distributed systems.

SimGrid is not limited to abstract algorithms, and can also be used to simulate real applications. This year, we published a journal article on the many challenges to overcome when designing a simulator of high performance systems. This work was published in the TPDS journal [20].

On the modeling side, our team worked this year toward the improvement of energy models, both for computational facilities and for the network. Despite the scarce availability of real testbeds that allow fine-grained energy measurements, we managed to provide a generic energy consumption model, published in [35], [43].

Finally, we restarted our efforts toward the formal verification of distributed systems. The model-checker that is integrated within SimGrid is already functional ([44]), but more work is necessary to make it efficient. We even found cases for which our reduction algorithm may miss defects in the verified system. This work will certainly motivate much more work in the future years.

7.4.2. Use cases

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

7.4.2.1. Simulation framework for studying between-herd pathogen spread in a region

In the context of the MIHMES project (2012-2017) and in collaboration with INRA researchers, we transformed a legacy application for simulating the spread of bovine viral diarrhea virus (BVDV) to a cloud-enabled application based on the DiFFuSE framework (Distributed framework for cloud-based epidemic simulations). Specifically, the original sequential code was first modified to add single-computer parallelism using OpenMP. We then decomposed the code into separate services that were deployed across multiple clouds and independently scaled. Using this service-based cloud-enabled simulation, we performed a set of experiments that demonstrated that applying DiFFuSE increases performance, allows exploring different cost-performance trade-offs, automatically handles failures, and supports elastic allocation of resources from multiple clouds [45].

7.4.2.2. FluxNet and AmeriFlux Data Analysis

The carbon flux datasets from AmeriFlux (Americas) and FLUXNET (global) are comprised of long-term time series data and other measurements at each tower site. There are over 800 flux towers around the world collecting this data. The non-time series measurements include information critical to performing analysis on the site's data. Examples include: canopy height, species distribution, soil properties, leaf area, instrument heights, etc. These measurements are reported as a variable group where the value plus information such as method of measurement and other information are reported together. Each variable group has a different number and type of parameters that are reported. The current output format is a normalized file. Users have found this file difficult to use.

Our earlier work in the DALHIS Inria associate team focused on building user interfaces to specify the data. This year we jointly worked on developing a Jupyter Notebook that would serve as a tool for users to read in and explore the data in a personalized tutorial type environment. We developed two notebooks and the next step is to start user testing on the notebooks.

REGAL Project-Team

5. New Results

5.1. Distributed Algorithms for Dynamic Networks and Fault Tolerance

Participants: Luciana Bezerra Arantes [correspondent], Sébastien Bouchard, Marjorie Bournat, João Paulo de Araujo, Swan Dubois, Denis Jeanneau, Jonathan Lejeune, Franck Petit [correspondent], Pierre Sens, Julien Sopena.

Nowadays, distributed systems are more and more heterogeneous and versatile. Computing units can join, leave or move inside a global infrastructure. These features require the implementation of *dynamic* systems, that is to say they can cope autonomously with changes in their structure in terms of physical facilities and software. It therefore becomes necessary to define, develop, and validate distributed algorithms able to managed such dynamic and large scale systems, for instance mobile *ad hoc* networks, (mobile) sensor networks, P2P systems, Cloud environments, robot networks, to quote only a few.

The fact that computing units may leave, join, or move may result of an intentional behavior or not. In the latter case, the system may be subject to disruptions due to component faults that can be permanent, transient, exogenous, evil-minded, etc. It is therefore crucial to come up with solutions tolerating some types of faults.

In 2017, we obtained the following results.

5.1.1. Algorithms for Dynamic and Large Systems

In [32] we propose VCube-PS, a new topic-based Publish/Subscribe system built on the top of a virtual hypercube like topology. Membership information and published messages to subscribers (members) of a topic group are broadcast over dynamically built spanning trees rooted at the message's source. For a given topic, delivery of published messages respects causal order. Performance results of experiments conducted on the PeerSim simulator confirm the efficiency of VCube-PS in terms of scalability, latency, number, and size of messages when compared to a single rooted, not dynamically, tree built approach.

We also explore in [20] scheduling challenges in providing probabilistic Byzantine fault tolerance in a hybrid cloud environment, consisting of nodes with varying reliability levels, compute power, and monetary cost. In this context, the probabilistic Byzantine fault tolerance guarantee refers to the confidence level that the result of a given computation is correct despite potential Byzantine failures. We formally define a family of such scheduling problems distinguished by whether they insist on meeting a given latency limit and trying to optimize the monetary budget or vice versa. For the case where the latency bound is a restriction and the budget should be optimized, we propose several heuristic protocols and compare between them using extensive simulations.

In [27], we a new resource reservation protocol in the context of delay-sensitive rescue mobile networks. The search for service providers (e.g., ambulance, fire truck, etc.) after a disaster, must take place within a short time. Therefore, service discovery protocol which looks for providers that can attend victims, respecting time constraints, is crucial. In such a situation, a commonly solution for ensuring network connectivity between victims and providers is ad hoc networks (MANET), composed by battery-operated mobile nodes of persons (victims or not). Using message aggregations techniques, we propose an new reservation protocol aiming at reducing the number of messages over the network and, therefore, node's battery consumption

5.1.2. Self-Stabilization

Self-stabilization is a generic paradigm to tolerate transient faults (*i.e.*, faults of finite duration) in distributed systems. In [14], we propose a silent self-stabilizing leader election algorithm for bidirectional arbitrary connected identified networks. This algorithm is written in the locally shared memory model under the distributed unfair daemon. It requires no global knowledge on the network. Its stabilization time is in $\Theta(n^3)$

steps in the worst case, where n is the number of processes. Its memory requirement is asymptotically optimal, *i.e.*, $\Theta(\log n)$ bits per processes. Its round complexity is of the same order of magnitude — *i.e.*, $\Theta(n)$ rounds — as the best existing algorithms designed with similar settings. To the best of our knowledge, this is the first self-stabilizing leader election algorithm for arbitrary identified networks that is proven to achieve a stabilization time polynomial in steps. By contrast, we show that the previous best existing algorithms designed with similar settings stabilize in a non polynomial number of steps in the worst case.

5.1.3. Mobile Agents

In [21], we consider systems made of autonomous mobile robots evolving in highly dynamic discrete environment *i.e.*, graphs where edges may appear and disappear unpredictably without any recurrence, stability, nor periodicity assumption. Robots are uniform (they execute the same algorithm), they are anonymous (they are devoid of any observable ID), they have no means allowing them to communicate together, they share no common sense of direction, and they have no global knowledge related to the size of the environment. However, each of them is endowed with persistent memory and is able to detect whether it stands alone at its current location. A highly dynamic environment is modeled by a graph such that its topology keeps continuously changing over time. In this paper, we consider only dynamic graphs in which nodes are anonymous, each of them is infinitely often reachable from any other one, and such that its underlying graph (*i.e.*, the static graph made of the same set of nodes and that includes all edges that are present at least once over time) forms a ring of arbitrary size.

In this context, we consider the fundamental problem of *perpetual exploration*: each node is required to be infinitely often visited by a robot. This paper analyzes the computability of this problem in (fully) synchronous settings, *i.e.*, we study the deterministic solvability of the problem with respect to the number of robots. We provide three algorithms and two impossibility results that characterize, for any ring size, the necessary and sufficient number of robots to perform perpetual exploration of highly dynamic rings.

5.1.4. Approach in the Plane

In [35] we study the task of *approach* of two mobile agents having the same limited range of vision and moving asynchronously in the plane. This task consists in getting them in finite time within each other's range of vision. The agents execute the same deterministic algorithm and are assumed to have a compass showing the cardinal directions as well as a unit measure. On the other hand, they do not share any global coordinates system (like GPS), cannot communicate and have distinct labels. Each agent knows its label but does not know the label of the other agent or the initial position of the other agent relative to its own. The route of an agent is a sequence of segments that are subsequently traversed in order to achieve approach. For each agent, the computation of its route depends only on its algorithm and its label. An adversary chooses the initial positions of both agents in the plane and controls the way each of them moves along every segment of the routes, in particular by arbitrarily varying the speeds of the agents. Roughly speaking, the goal of the adversary is to prevent the agents from solving the task, or at least to ensure that the agents have covered as much distance as possible before seeing each other. A deterministic approach algorithm is a deterministic algorithm that always allows two agents with any distinct labels to solve the task of approach regardless of the choices and the behavior of the adversary. The cost of a complete execution of an approach algorithm is the length of both parts of route travelled by the agents until approach is completed.

Let Δ and l be the initial distance separating the agents and the length of (the binary representation of) the shortest label, respectively. *Assuming that Δ and l are unknown to both agents, does there exist a deterministic approach algorithm whose cost is polynomial in Δ and l?*

Actually the problem of approach in the plane reduces to the network problem of rendezvous in an infinite oriented grid, which consists in ensuring that both agents end up meeting at the same time at a node or on an edge of the grid. By designing such a rendezvous algorithm with appropriate properties, as we do in this paper, we provide a positive answer to the above question.

Our result turns out to be an important step forward from a computational point of view, as the other algorithms allowing to solve the same problem either have an exponential cost in the initial separating distance and in the

labels of the agents, or require each agent to know its starting position in a global system of coordinates, or only work under a much less powerful adversary.

5.2. Large scale data distribution

Participants: Mesaac Makpangou, Sébastien Monnet, Pierre Sens, Marc Shapiro, Paolo Viotti, Sreeja Nair, Ilyas Toumlilt, Alejandro Tomsic, Dimitrios Vasilas.

5.2.1. Data placement and searches over large distributed storage

Distributed storage systems such as Hadoop File System or Google File System (GFS) ensure data availability and durability using replication. Persistence is achieved by replicating the same data block on several nodes, and ensuring that a minimum number of copies are available on the system at any time. Whenever the contents of a node are lost, for instance due to a hard disk crash, the system regenerates the data blocks stored before the failure by transferring them from the remaining replicas. In [33] we focused on the analysis of the efficiency of replication mechanism that determines the location of the copies of a given file at some server. The variability of the loads of the nodes of the network is investigated for several policies. Three replication mechanisms are tested against simulations in the context of a real implementation of a such a system: Random, Least Loaded and Power of Choice. The simulations show that some of these policies may lead to quite unbalanced situations. It is shown in this paper that a simple variant of a power of choice type algorithm has a striking effect on the loads of the nodes. Mathematical models are introduced and investigated to explain this interesting phenomenon. The analysis of these systems turns out to be quite complicated mainly because of the large dimensionality of the state spaces involved. Our study relies on probabilistic methods, mean-field analysis, to analyze the asymptotic behavior of an arbitrary node of the network when the total number of nodes gets large.

In the summary prefix tree (SPT), a trie data structure that supports efficient superset searches over DHT. Each document is summarized by a Bloom filter which is then used by SPT to index this document. SPT implements an hybrid lookup procedure that is well-adapted to sparse indexing keys such as Bloom filters. It also proposes a mapping function that permits to mitigate the impact of the skewness of SPT due to the sparsity of Bloom filters, especially when they contain only few words. To perform efficient superset searches, SPT maintains on each node a local view of the global tree. The main contributions are the following. First, the approximation of the superset relationship among keyword-sets by the descendance relationship among Bloom filters. Second, the use of a summary prefix tree (SPT), a trie indexing data structure, for keyword-based search over DHT. Third, an hybrid lookup procedure which exploits the sparsity of Bloom filters to offer good performances. Finally, an algorithm that exploits SPT to efficiently find descriptions that are supersets of query keywords.

5.2.2. Just-Right Consistency

Consistency is a major concern in the design of distributed applications, but the topic is still not well understood. It is clear that no single consistency model is appropriate for all applications, but how do developers find their way in the maze of models and the inherent trade-offs between correctness and availability? The Just-Right Consistency approach presented here offers some guidance. First, we classify the safety patterns that are of interest to maintain application correctness. Second, we show how two of these patterns are “AP-compatible” and can be guaranteed without impacting availability, thanks to an appropriate data model and consistency model. Then we address the last, “CAP-sensitive” pattern. In a restricted but common case it can be maintained efficiently in a mostly-available way. In the general case, we exhibit a static analysis logic and tool which ensures just enough synchronisation to maintain the invariant, and availability otherwise.

In summary, instead of pre-defining a consistency model and shoe-horning the application to fit it, and instead of making the application developer compensate for the imperfections of the data store in an *ad-hoc* way, we have a provably correct approach to tailoring consistency to the specific application requirements. This approach is supported by several artefacts developed by Regal and collaborators: Conflict-Free Replicated Data Types (CRDTs), the Antidote cloud database, and the CISE verification tool.

This paper is under submission.

5.3. Memory management in system software

Participants: Damien Carver, Jonathan Lejeune, Pierre Sens, Julien Sopena [correspondent], Gauthier Voron.

Recent years have seen the increasingly widespread use of **multicore** architectures and **virtualized environments**. This development has an impact on all parts of the system software. Virtual machine (VM) technology offers both isolation and flexibility but has side effects such as fragmentation of the physical resources, including memory. This fragmentation reduces the amount of available memory a VM can use. Many recent works study that a NUMA (Non Uniform Memory Access) architecture, common in large multi-core processors, highly impacts application performance. We focus on improving the memory and cache management in various virtualized environments such as Xen hypervisor or linux-containers targeting big data applications on multicore architectures.

While virtualization only introduces a small overhead on machines with few cores, this is not the case on larger ones. Most of the overhead on the latter machines is caused by the NUMA architecture they are using. In order to reduce this overhead, in [34] we show how NUMA placement heuristics can be implemented inside Xen. With an evaluation of 29 applications on a 48-core machine, we show that the NUMA placement heuristics can multiply the performance of 9 applications by more than 2.

We also study the memory arbitration between containers. In the Damien Carver's PhD thesis, we are designing ACDC [23] (Advanced Consolidation for Dynamic Containers), a kernel-level mechanisms that automatically provides more memory to the most active containers.

In the Francis Laniel's PhD thesis, we study a new architecture using Non Volatile RAM NVRAM. Although NVRAM are slower than classical RAM, they have better energetic features. We investigate solutions where RAM and NVRAM coexist in order to balance the energy consumption and performance according to the needs of the system.

SPIRALS Project-Team

7. New Results

7.1. A Domain-specific Language for The Control of Self-adaptive Component-based Architecture

In [12], together with Frederico Alvares (Inria Ascola) and Eric Rutten (Inria Ctrl-A), we have proposed Ctrl-F, a new domain-specific language for specifying reconfiguration policies in self-adaptable component-based software systems. Self-adaptive behaviors in the context of component-based architecture are generally designed based on past monitoring events, configurations (component assemblies) as well as behavioral programs defining the adaptation logics and invariant properties. The novelty of the proposed Ctrl-F language is to enable taking decisions on predictions on the possible futures of the system in order to avoid going into branches of the behavioral program leading to bad configurations. Ctrl-F is formally defined by a translation into *Finite State Automata* models. We use *Discrete Controller Synthesis* to automatically generate a controller to enforce correct self-adaptive behaviors. Ctrl-F is integrated with our FraSCAti middleware platform for distributed service and component oriented systems.

7.2. A New Interface for Mobile Cloud Robotics

In [35], together with Nathalie Mitton (Inria Fun), we have proposed OMCRI, a new interface for mobile cloud robotics. This interface enables to abstract from the heterogeneity of robotic platforms and to bring some resource management facilities to fleets of robots. This result is based on the expertise that we have developed in the management of resources for cloud computing environments, especially around the OCCI standard. To the best of our knowledge, OMCRI is the first interface that enables to concretize the vision of robotics as a service. This result has obtained the best award at the 2nd IEEE International Congress on Internet of Things (ICIOT 2017).

WHISPER Project-Team

7. New Results

7.1. Software engineering for infrastructure software

Work in 2017 on the Linux kernel has focused on the problem of kernel device driver porting and on kernel compilation as a validation mechanism in the presence of variability. We have also completed a study with researchers at Singapore Management University on the relationship between the code coverage of test cases and the number of post-release defects, focusing on a range of popular open-source projects. Finally, we have worked with researchers at the University of Frankfurt on the design of a transformation language targeting data representation changes.

Porting Linux device drivers to target more recent and older Linux kernel versions to compensate for the ever-changing kernel interface is a continual problem for Linux device driver developers. Acquiring information about interface changes is a necessary, but tedious and error prone, part of this task. To address these problems, we have proposed two tools, *Prequel* and *gcc-reduce*, to help the developer collect the needed information. Prequel provides language support for querying git commit histories, while gcc-reduce translates error messages produced by compiling a driver with a target kernel into appropriate Prequel queries. We have used our approach in porting 33 device driver files over up to 3 years of Linux kernel history, amounting to hundreds of thousands of commits. In these experiments, for 3/4 of the porting issues, our approach highlighted commits that enabled solving the porting task. For many porting issues, our approach retrieves relevant commits in 30 seconds or less. This work was published at USENIX ATC [16] and a related talk was presented at Linuxcon Europe. The Prequel tool and some of our experimental results are available at http://prequel-pql.gforge.inria.fr/. The complete tool suite is available at http://select-new.gforge.inria.fr/.

The Linux kernel is highly configurable, and thus, in principle, any line of code can be included or excluded from the compiled kernel based on configuration operations. Configurability complicates the task of a *kernel janitor*, who cleans up faults across the code base. A janitor may not be familiar with the configuration options that trigger compilation of a particular code line, leading him to believe that a fix has been compile-checked when this is not the case. We have proposed JMake, a mutation-based tool for signaling changed lines that are not subjected to the compiler. JMake shows that for most of the 12,000 file-modifying commits between Linux v4.3 and v4.4 the configuration chosen by the kernel allyesconfig option is sufficient, once the janitor chooses the correct architecture. For most commits, this check requires only 30 seconds or less. We furthermore characterize the situations in which changed code is not subjected to compilation in practice. This work was published at DSN [15] and a related talk was presented at Linuxcon Europe. JMake is available at http://jmake-release.gforge.inria.fr/.

Testing is a pivotal activity in ensuring the quality of software. Code coverage is a common metric used as a yardstick to measure the efficacy and adequacy of testing. However, does higher coverage actually lead to a decline in post-release bugs? Do files that have higher test coverage actually have fewer bug reports? The direct relationship between code coverage and actual bug reports has not yet been analysed via a comprehensive empirical study on real bugs. In an empirical study, we have examined these questions in the context of 100 large open-source Java software projects based on their actual reported bugs. Our results show that coverage has an insignificant correlation with the number of bugs that are found after the release of the software at the project level, and no such correlation at the file level. This work was done in collaboration with researchers at Singapore Management University and has been published in the IEEE Transactions on Reliability [12].

Data representation migration is a program transformation that involves changing the type of a particular data structure, and then updating all of the operations that somehow depend on that data structure according to the new type. Changing the data representation can provide benefits such as improving efficiency and improving the quality of the computed results. Performing such a transformation is challenging, because it requires applying data-type specific changes to code fragments that may be widely scattered throughout the

source code, connected by dataflow dependencies. Refactoring systems are typically sensitive to dataflow dependencies, but are not programmable with respect to the features of particular data types. Existing program transformation languages provide the needed flexibility, but do not concisely support reasoning about dataflow dependencies.

To address the needs of data representation migration, we have proposed a new approach to program transformation that relies on a notion of semantic dependency: every transformation step propagates the transformation process onward to code that somehow depends on the transformed code. Our approach provides a declarative transformation-specification language, for expressing type-specific transformation rules. Our approach further provides scoped rules, a mechanism for guiding rule application, and tags, a device for simple program analysis within our framework, to enable more powerful program transformations. Evaluation of our prototype based on our approach, targeting C and C++ software, shows that it can improve program performance and the precision of the computed results, and that it scales to programs of up to 3700 lines. This work was done in collaboration with researchers at the University of Frankfurt and was published at PEPM [18].

7.2. Trustworthy domain-specific compilers

This year, we concluded the correctness proof of the compiler back-end of the Lustre [32] synchronous dataflow language. Synchronous dataflow languages are widely used for the design of embedded systems: they allow a high-level description of the system and naturally lend themselves to a hierarchical design. Developed in collaboration with members of the Parkas team of Inria Paris (Tim Bourke, Lélio Brun, Marc Pouzet), the Gallium team of Inria Paris (Xavier Leroy) and Collège de France (Lionel Rieg), this work formalizes the compilation of a synchronous data-flow language into an imperative sequential language, which is eventually translated to Cminor [56], one of CompCert's intermediate languages. The proof has been developed and verified in the Coq theorem prover. This project illustrates perfectly our methodology: the design of synchronous dataflow languages is first governed by semantic considerations (Kahn process networks and the synchrony hypothesis) that are then reifed into syntactic artefacts. The implementation of a certified compiler highlights this dependency on semantics, forcing us to give as crisp a semantics as possible for the proof effort to be manageable. This work was published in a national conference [19] as well as in an international conference [13], both on the topic of language design and implementation.

Expanding upon these ideas, Darius Mercadier started his PhD with us in October. We are currently developing a synchronous dataflow language targeting verified and high-performance implementations of bitsliced algorithms, with application to cryptographical algorithms [40]. Our preliminary results [22] are encouraging.

7.3. Algebra of programming

We have pursued our study of the algebraic structures of programming languages, from a syntactic as well as semantics perspective. Tackling the semantics aspect, Pierre-Évariste Dagand published a journal article introducing the theory of ornaments [11] to a general audience of functional programmers. Ornaments amount to a domain-specific language, usually described in type theory, for describing structure-preserving changes in algebraic datatypes. Such descriptions can be used to improve code reuse as well as ease of refactoring in functional languages. This work is part of a wider effort by our community to foster the adoption of ornaments when programming with algebraic datatypes, be it in type theory [48] or general-purpose functional programming languages [65], [89]. Tackling the syntactic aspect and in collaboration with researchers at the University of Utrecht (Victor Miraldo, Wouter Swierstra), Pierre-Évariste Dagand has worked on a formalization of `diffs` for structured data [20]. This preliminary and foundational work aims at providing a typed specification to the problem of computing the difference of two pieces of structured data. Unlike previous approaches [43], following a type-theoretical approach allowed us to formalize the difference of two structure as a typed object. The task of computing the difference of two structured objects is then able to exploit this typing information to control the search space (which is otherwise gigantic). Having a typed difference also ensures that applying such a `diff` to a well-structured data results in either a failure (the difference is in conflict with the given file) or another well-structured data.

7.4. Developing infrastructure software using Domain Specific Languages

In terms of DSL design for domains where correctness is critical, our current focus is first on process scheduling for multicore architecture, and second on selfishness in distributed systems. Ten years ago, we developed Bossa, targeting process scheduling on unicore processors, and primarily focusing on the correctness of a scheduling policy with respect to the requirements of the target kernel. At that time, the main use cases were soft real-time applications, such as video playback. Bossa was and still continues to be used in teaching, because the associated verifications allow a student to develop a kernel-level process scheduling policy without the risk of a kernel crash. Today, however, there is again a need for the development of new scheduling policies, now targeting multicore architectures. As identified by Lozi *et al.* [61], large-scale server applications, having specific resource access properties, can exhibit pathological properties when run with the Linux kernel's various load balancing heuristics. We are working on a new domain-specific language, Ipanema, to enable verification of critical scheduling properties such as liveness and work-conservation; for the latter, we are exploring the use of the Leon theorem prover from EPFL [17]. A first version of the language has been designed and we expect to release a prototype of Ipanema working next year. The work around Ipanema is the subject of a very active collaboration between researchers at four institutions (Inria, University of Nice, University of Grenoble, and EPFL (groups of V. Kuncak and W. Zwaenepoel)). Baptiste Lepers (EPFL) is supported in 2017 as a postdoc as part of the Inria-EPFL joint laboratory.

Selfishness is one of the key problems that confronts developers of cooperative distributed systems (e.g., file-sharing networks, voluntary computing). It has the potential to severely degrade system performance and to lead to instability and failures. Current techniques for understanding the impact of selfish behaviours and designing effective countermeasures remain manual and time-consuming, requiring multi-domain expertise. To overcome these difficulties, we have proposed SEINE, a simulation framework for rapid modelling and evaluation of selfish behaviours in a cooperative system. SEINE relies on a domain-specific language (SEINE-L) for specifying selfishness scenarios, and provides semi-automatic support for their implementation and study in a state-of-the-art simulator. We show in a paper published at DSN 2017 [14] that (1) SEINE-L is expressive enough to specify fifteen selfishness scenarios taken from the literature, (2) SEINE is accurate in predicting the impact of selfishness compared to real experiments, and (3) SEINE substantially reduces the development effort compared to traditional manual approaches.

ALPINES Project-Team

7. New Results

7.1. Communication avoiding algorithms for preconditioned iterative methods

Our group continues to work on algorithms for dense and sparse linear algebra operations that minimize communication, introduced in [1], [4]. An overview of communication avoiding algorithms for dense linear algebra operations is presented in [18]. During this year we focused on communication avoiding iterative methods and designing algorithms for computing rank revealing and low rank approximations of dense and sparse matrices.

Iterative methods are widely used in industrial applications, and in the context of communication avoiding algorithms, our research is related to increasing the scalability of Krylov subspace iterative methods. Indeed the dot products related to the orthogonalization of the Krylov subspace and performed at each iteration of the Krylov method require collective communication among all processors. This collective communication does not scale to very large number of processors, and thus is a main bottleneck in the scalability of Krylov subspace methods. Our research focuses on enlarged Krylov subspace methods, a new approach that we have introduced in the recent years [5] that consists of enlarging the Krylov subspace by a maximum of t vectors per iteration, based on a domain decomposition of the graph of the input matrix. The solution of the linear system is searched in the enlarged subspace, which is a superset of the classic subspace. The enlarged Krylov projection subspace methods lead to faster convergence in terms of iterations and parallelizable algorithms with less communication, with respect to Krylov methods.

In [20] we propose an algebraic method in order to reduce dynamically the number of search directions during block Conjugate Gradient iterations. Indeed, by monitoring the rank of the optimal step α_k it is possible to detect inexact breakdowns and remove the corresponding search directions. We also propose an algebraic criterion that ensures in theory the equivalence between our method with dynamic reduction of the search directions and the classical block Conjugate Gradient. Numerical experiments show that the method is both stable, the number of iterations with or without reduction is of the same order, and effective, the search space is significantly reduced. We use this approach in the context of enlarged Krylov subspace methods which reduce communication when implemented on large scale machines. The reduction of the number of search directions further reduces the computation cost and the memory usage of those methods.

In [19] we propose a variant of the GMRES method for solving linear systems of equations with one or multiple right-hand sides. Our method is based on the idea of the enlarged Krylov subspace to reduce communication. It can be interpreted as a block GMRES method. Hence, we are interested in detecting inexact breakdowns. We introduce a strategy to perform the test of detection. Furthermore, we propose an eigenvalues deflation technique aiming to have two benefits. The first advantage is to avoid the plateau of convergence after the end of a cycle in the restarted version. The second is to have a very fast convergence when solving the same system with different right-hand sides, each given at a different time (useful in the context of CPR preconditioner). With the same memory cost, we obtain a saving of up to 50% in the number of iterations to reach convergence with respect to the original method.

7.2. Communication avoiding algorithms for low rank matrix approximation

Our work focuses on computing the low rank approximation of a sparse or dense matrix, while also minimizing communication, [3].

In [21] we introduce an URV Factorization with Random Orthogonal System Mixing. The unpivoted and pivoted Householder QR factorizations are ubiquitous in numerical linear algebra. A difficulty with pivoted Householder QR is the communication bottleneck introduced by pivoting. In this paper we propose using random orthogonal systems to quickly mix together the columns of a matrix before computing an unpivoted QR factorization. This method computes a URV factorization which forgoes expensive pivoted QR steps in exchange for mixing in advance, followed by a cheaper, unpivoted QR factorization. The mixing step typically reduces the variability of the column norms, and in certain experiments allows us to compute an accurate factorization where a plain, unpivoted QR performs poorly. We experiment with linear least-squares, rank-revealing factorizations, and the QLP approximation, and conclude that our randomized URV factorization behaves comparably to a similar randomized rank-revealing URV factorization, but at a fraction of the computational cost. Our experiments provide evidence that our proposed factorization might be rank-revealing with high probability.

7.3. Domain decomposition preconditioning for high frequency wave propagation problems

This work studies preconditioning the Helmholtz and Maxwell equations, where the preconditioner is constructed using two-level overlapping Additive Schwarz Domain Decomposition. The coarse space is based on the discretisation of the PDE on a coarse mesh. The PDE is discretised using finite-element methods of fixed, arbitrary order. The theoretical part of this work is the Maxwell analogue of a previous work for Helmholtz equation, and shows that for Maxwell problems with absorption, if the absorption is large enough and if the subdomain and coarse mesh diameters are chosen appropriately, then classical two-level overlapping Additive Schwarz Domain Decomposition preconditioning performs optimally – in the sense that GMRES converges in a wavenumber-independent number of iterations. An important feature of the theory is that it allows the coarse space to be built from low-order elements even if the PDE is discretised using high-order elements. This theory is presented in [24] and is illustrated by numerical experiments, which also (i) explore replacing the PEC boundary conditions on the subdomains by impedance boundary conditions, and (ii) show that the preconditioner for the problem with absorption is also an effective preconditioner for the problem with no absorption. The numerical results include two substantial examples arising from applications; the first (a problem arising in medical imaging from the Medimax ANR project) shows the robustness of the preconditioner against heterogeneity, and the second (scattering by a COBRA cavity) shows good scalability of the preconditioner with up to 3000 processors. The parallel implementation was done using FreeFem++ and HPDDM. We performed additional numerical studies of this two-level Domain Decomposition preconditioner for the Maxwell equations in [23], and for the Helmholtz equation (in 2D and 3D) in [25], where we also compare it to another two-level Domain Decomposition preconditioner where the coarse space is built by solving local eigenproblems on the interface between subdomains involving the Dirichlet-to-Neumann (DtN) operator.

7.4. First kind boundary integral formulation for the Hodge-Helmholtz equation

We adapt the variational approach to the analysis of first-kind boundary integral equations associated with strongly elliptic partial differential operators from [M. COSTABEL, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.] to the (scaled) Hodge-Helmholtz equation $\mathbf{curl}\,\mathbf{curl}\,\mathbf{u} - \eta\nabla\mathrm{div}\,\mathbf{u} - \kappa^2\mathbf{u} = 0, \eta > 0, \mathrm{Im}\,\kappa^2 \geq 0$, on Lipschitz domains in 3D Euclidean space, supplemented with natural complementary boundary conditions, which, however, fail to bring about strong ellipticity.

Nevertheless, a boundary integral representation formula can be found, from which we can derive boundary integral operators. They induce bounded and coercive sesqui-linear forms in the natural energy trace spaces for the Hodge-Helmholtz equation. We can establish precise conditions on η, κ that guarantee unique solvability of the two first-kind boundary integral equations associated with the natural boundary value problems for the Hodge-Helmholtz equations. Particular attention needs to be given to the case $\kappa = 0$.

7.5. Integral equation based optimized Schwarz method for electromagnetics

The optimized Schwarz method (OSM) is recognised as one of the most efficient domain decomposition strategies without overlap for the solution to wave propagation problems in harmonic regime. For the Helmholtz equation, this approach originated from the seminal work of Després, and led to the development of an abundent literature offering more elaborated but more efficient transmission conditions. Most contributions focus on transmission conditions based on local operators.

In recent years, F. Collino, P. Joly and M. Lecouvez introduced non-local transmission conditions that can drastically improve the convergence rate of OSM. The performance of this strategy seems to remain robust at high frequency. Such an approach was proposed only for the Helmholtz equation, and has still not been adapted to electromagnetics.

In this work we investigated such an approach for Maxwell's equations in a simple spherical geometry that allows explicit calculus by means of separation of variables. The transmission condition that we propose involves a non-local operator that is a dissipative counterpart of the so-called Electric Field integral operator (EFIE) which is a classical object in electromagnetic potential theory. We show that the iterative solver associated to our strategy converges at an exponential rate.

7.6. Quasi-local Multi-Trace formulations for electromagnetics

Multi-trace formulations (MTF) are a general methodology to derive first kind boundary integral formulations for harmonic wave scattering problems posed in multi-domain geometrical confgurations. There exists both a local and a global variant of MTF that only differ through the way transmission conditions are imposed across interfaces. Global MTF is easier to analyse but, from a computational viewpoint, local MTF appears more appealing because it looks computationally cheaper.

As regards local MTF, a decent stability theory has been developped for acoustic scalar wave propagation, but no such result as Garding inequality or uniform discrete inf-sup condition has been established so far for local MTF in the case of electromagnetics. Wether or not local MTF is stable for electromagnetics is actually an open question presently.

In this work, we have adopted a slightly modified version of local MTF where tranmission conditions are imposed by means of an operator that is non-local, but with a kernel whose support can be as small as desired. This so-called quasi-local MTF approach has previsouly been developped for acoustics and we adapted it to the case of electromagnetics. We could in particular prove a Garding inequality for quasi-local MTF applied to electromagnetics, and thus obtain uniform discrete inf-sup condition.

7.7. Domain decomposition preconditioning with approximate coarse solve

Convergence of domain decomposition methods relies heavily on the efficiency of the coarse space used in the second level. The GenEO coarse space has been shown to lead to a fully robust two-level Schwarz preconditioner which scales well over multiple cores [9], [2] as has been proved rigorously in [9]. The robustness is due to its good approximation properties for problems with highly heterogeneous material parameters. It is available in the finite element packages FreeFem++ [7], Feel++ [31] and recently in Dune [30] and is implemented as a standalone library in HPDDM [8]. But the coarse component of the preconditioner can ultimately become a bottleneck if the number of subdomains is very large and exact solves are used. It is therefore interesting to consider the effect of approximate coarse solves. In [28], robustness of GenEO methods is analyzed with respect to approximate coarse solves. Interestingly, the GenEO-2 method introduced in [6] has to be modified in order to be able to prove its robustness in this context.

AVALON Project-Team

6. New Results

6.1. Energy Efficiency in HPC and Large Scale Distributed Systems

Participants: Mathilde Boutigny, Radu Carpa, Marcos Dias de Assunção, Thierry Gautier, Olivier Glück, Laurent Lefèvre, Jean-Christophe Mignot, Issam Rais.

6.1.1. Combining Shutdown Policies with Multiple Constraints

Large scale distributed systems (high performance computing centers, networks, data centers) are expected to consume huge amounts of energy. In order to address this issue, shutdown policies constitute an appealing approach able to dynamically adapt the resource set to the actual workload. However, multiple constraints have to be taken into account for such policies to be applied on real infrastructures: the time and energy cost of switching on and off, the power and energy consumption bounds caused by the electricity grid or the cooling system, and the availability of renewable energy. We propose models translating these various constraints into different shutdown policies that can be combined for a multi-constraint purpose. Our models and their combinations are validated through simulations on a real workload trace [4], [13]. This work is done through the PhD of Issam Rais in the FSN ELCI Project with the collaboration of Anne Benoit (Roma team) and Anne-Cécile Orgerie (Myriads team).

6.1.2. Evaluating the Impact of SDN-Induced Frequent Route Changes on TCP Flows

Traffic engineering technologies such as MPLS have been proposed to adjust the paths of data flows according to network availability. Although the time interval between traffic optimisations is often on the scale of hours or minutes, modern SDN techniques enable reconfiguring the network more frequently. It is argued, however, that changing the paths of TCP flows too often could severely impact their performance by incurring packet loss and reordering. This work analyses and evaluates the impact of frequent route changes on the performance of TCP flows. Experiments carried out on a network testbed show that rerouting a flow can affect its throughput when reassigning it a path either longer or shorter than the original path. Packet reordering has a negligible impact when compared to the increase of RTT. Moreover, constant rerouting influences the performance of the congestion control algorithm. Designed to assess the limits on SDN-induced reconfiguration, a scenario where the traffic is rerouted every 0.1s demonstrates that the throughput can be as low as 35% of that achieved without rerouting.[7], [14].

6.1.3. Evaluating Energy Consumption of OpenMP Runtime

In a joint-work with J.V. Lima from UFSM, Santa Maria, Brazil [26], we analyse performance and energy consumption of four OpenMP runtime systems over a NUMA platform. We present an experimental study to characterize OpenMP runtime systems on the three main kernels in dense linear algebra algorithms (Cholesky, LU and QR) in terms of performance and energy consumption. Our experimental results suggest that OpenMP runtime systems can be considered as a new energy leverage. For instance, a LU factorization with concurrent write extension from libKOMP achieved up to 1.75 of performance gain and 1.56 of energy decrease.

6.2. Modeling and Simulation of Parallel Applications and Distributed Infrastructures

Participant: Frédéric Suter.

6.2.1. Simulating MPI Applications: the SMPI Approach

Predicting the behavior of distributed algorithms has always been a challenge, and the scale of next-generation High Performance Computing (HPC) systems will only make the situation more difficult. Performance modeling and software engineering for these systems increasingly require a simulation-based approach, and this need will only become more apparent with the arrival of Exascale computing by the end of the decade. In [6] we summarized our recent work and developments on SMPI, a flexible simulator of MPI applications. In this tool, we took a particular care to ensure our simulator could be used to produce fast and accurate predictions in a wide variety of situations. Although we did build SMPI on SimGrid whose speed and accuracy had already been assessed in other contexts, moving such techniques to a HPC workload required significant additional effort. Obviously, an accurate modeling of communications and network topology was one of the key to such achievements. Another less obvious key was the choice to combine in a single tool the possibility to do both offline and online simulation.

6.2.2. Modeling Distributed Platforms from Application Traces

Simulation is a fast, controlled, and reproducible way to evaluate new algorithms for distributed computing platforms in a variety of conditions. However, the realism of simulations is rarely assessed, which critically questions the applicability of a whole range of findings.

In [15], we present our efforts to build platform models from application traces, to allow for the accurate simulation of file transfers across a distributed infrastructure. File transfers are key to performance, as the variability of file transfer times has important consequences on the dataflow of the application. We present a methodology to build realistic platform models from application traces and provide a quantitative evaluation of the accuracy of the derived simulations. Results show that the proposed models are able to correctly capture real-life variability and significantly outperform the state-of-the-art model.

6.3. Data Stream Processing and Edge Computing

Participants: Eddy Caron, Marcos Dias de Assunção, Alexandre Da Silva Veith, Laurent Lefèvre, Felipe Rodrigo de Souza.

6.3.1. Resource Elasticity and Edge Computing for Data Stream Processing

We carried out an extensive survey on techniques for enabling resource elasticity for data stream processing applications. Moreover we have been investigating algorithms for placing stream processing tasks onto environments that comprise both cloud and edge computing resources [29].

We are currently working on modelling the placement scenario as a constraint programming problem as well as measuring the energy consumption of constrained devices, such as Raspberry Pi's. The power consumption information is being used for creating a model on power consumption model.

6.4. Large-Scale Cloud Resource Management

Participants: Yves Caniou, Eddy Caron, Marcos Dias de Assunção, Christian Perez, Pedro de Souza Bento Da Silva.

6.4.1. An Efficient Communication Aware Heuristic for Multiple Cloud Application Placement

To deploy a distributed application on the cloud, cost, resource and communication constraints have to be considered to select the most suitable Virtual Machines (VMs), from private and public cloud providers. This process becomes very complex in large scale scenarios and, as this problem is NP-Hard, its automation must take scalability into consideration. In this work [21], we propose a heuristic able to calculate initial placements for distributed component-based applications on possibly multiple clouds with the objective of minimizing VM renting costs while satisfying applications' resource and communication constraints. We evaluate the heuristic performance and determine its limitations by comparing it to other placement approaches, namely exact algorithms and meta-heuristics. We show that the proposed heuristic is able to compute a good solution much faster than them.

6.4.2. Production Deployment Tools for IaaSes: an Overall Model and Survey

Emerging applications for the Internet of Things (IoT) are complex programs which are composed of multiple modules (or services). For scalability, reliability and performance, modular applications are distributed on infrastructures that support utility computing (*e.g.,* Cloud, Fog). In order to simply operate such infrastructures, an Infrastructure-as-a-Service (IaaS) manager is required. OpenStack is the de-facto open-source solution to address the IaaS level of the Cloud paradigm. However, OpenStack is itself a large modular application composed of more than 150 modules that make it hard to deploy manually. To fully understand how IaaSes are deployed today, we propose in [16] an overall model of the application deployment process which describes each step with their interactions. This model then serves as the basis to analyse five different deployment tools used to deploy OpenStack in production: Kolla, Enos, Juju, Kubernetes, and TripleO. Finally, a comparison is provided and the results are discussed to extend this analysis.

6.4.3. Communication Aware Task Placement for Workflow Scheduling on DaaS-based Cloud

We proposed a framework for building an autonomous workflow manager and developped the different components that are required for this design to work. We believe that this design will help solve current issues with workflow deployment and scalling in the context of shared IaaS Cloud platforms. In that regard, our first contribution is the modelization of network topology [24], which is a key factor in predicting communication patterns and should therefore be considered by clustering algorithms. By designing a generic network model, we managed to improve the results of static scheduling in the context of DaaS-based Cloud platforms. In fact, the resulting clusters are both more efficient in terms of makespan (primary objective) and in terms of deployment cost compared to previous non-network-aware clustering algorithms.

6.4.4. Communication Aware Stochastic Tasks Scheduling Composing Scientific Workflows on a Cloud

In order to study the scheduling of workflows composed of stochastic tasks on a set of resources managed as a cloud, we firstly proposed a new execution model taking into account data transfers, heterogeneity, billing of used resources as close to reality based to a great extend on the offers of three big cloud providers: Google Cloud, Amazon EC2 and OVH [25]. We then studied new scheduling heuristics on a set of worflows taken from the Pegasus benchmark suite [23]. During the mapping process, the budget-aware algorithms make conservative assumptions to avoid exceeding the initial budget; we further improve our results with refined versions that aim at re-scheduling some tasks onto faster virtual machines, thereby spending any budget fraction leftover by the first allocation. These refined variants are much more time-consuming than the former algorithms, so there is a trade-off to find in terms of scalability. We report an extensive set of simulations. Most of the time our budget-aware algorithms succeed in achieving efficient makespans while enforcing the given budget, and despite the uncertainty in task weights.

6.5. HPC Component Models and Domain Specific Languages

Participants: Thierry Gautier, Christian Perez, Jérôme Richard.

6.5.1. Combining Both a Component Model and a Task-based Model for HPC Applications: a Feasibility Study on GYSELA

In [12], we studied the feasibility of efficiently combining both a software component model and a task-based model. Task based models are known to enable efficient executions on recent HPC computing nodes while component models ease the separation of concerns of application and thus improve their modularity and adaptability. This paper describes a prototype version of the COMET programming model combining concepts of task-based and component models, and a preliminary version of the COMET runtime built on top of StarPU and L2C. Evaluations of the approach have been conducted on a real-world use-case analysis of a subpart of the production application GYSELA. Results show that the approach is feasible and that it enables easy composition of independent software codes without introducing overheads. Performance results are equivalent to those obtained with a plain OpenMP based implementation.

6.5.2. Extensibility and Composability of a Multi-Stencil Domain Specific Framework

As the computation power of modern high performance architectures increases, their heterogeneity and complexity also become more important. One of the big challenges of exascale is to reach programming models that give access to high performance computing (HPC) to many scientists and not only to a few HPC specialists. One relevant solution to ease parallel programming for scientists is domain specific language (DSL). However, one problem to avoid with DSLs is to mutualize existing codes and libraries instead of implementing each solution from scratch. For example, this phenomenon occurs for stencil-based numerical simulations, for which a large number of languages has been proposed without code reuse between them. The Multi-Stencil Framework (MSF) presented in this paper [5] combines a new DSL to component-based programming models to enhance code reuse and separation of concerns in the specific case of stencils. MSF can easily choose one parallelization technique or another, one optimization or another, as well as one back-end implementation or another. It is shown that MSF can reach same performances than a non component-based MPI implementation over 16,384 cores. Finally, the performance model of the framework for hybrid parallelization is validated by evaluations.

DATAMOVE Project-Team

7. New Results

7.1. Integration of High Performance Computing and Data Analytics

New results on the topic *Integration of High Performance Computing and Data Analytics* are related to compression [15], automatic data extraction for in situ processing [14], in transit sensitivity analysis [16] and management of heterogeneous HPC and BigData workloads [21]. We detail the two last here.

- **Large Scale In Transit Sensitivity Analysis Avoiding Intermediate Files [16].** Global sensitivity analysis is an important step for analyzing and validating numerical simulations. One classical approach consists in computing statistics on the outputs from well-chosen multiple simulation runs. Simulation results are stored to disk and statistics are computed postmortem. Even if supercomputers enable to run large studies, scientists are constrained to run low resolution simulations with a limited number of probes to keep the amount of intermediate storage manageable. In this paper we propose a file avoiding, adaptive, fault tolerant and elastic framework that enables high resolution global sensitivity analysis at large scale. Our approach combines iterative statistics and in transit processing to compute Sobol' indices without any intermediate storage. Statistics are updated on-the-fly as soon as the in transit parallel server receives results from one of the running simulations. For one experiment, we computed the Sobol' indices on 10M hexahedra and 100 timesteps, running 8000 parallel simulations executed in 1h27 on up to 28672 cores, avoiding 48TB of file storage. Based on this work we open sourced the associated framework called Melissa (https://melissa-sa.github.io).
- **Big Data and HPC collocation: Using HPC idle resources for Big Data Analytics [21].** Executing Big Data workloads upon High Performance Computing (HPC) infrastructures has become an attractive way to improve their performances. However, the collocation of HPC and Big Data workloads is not an easy task, mainly because of their core concepts' differences. This paper focuses on the challenges related to the scheduling of both Big Data and HPC workloads on the same computing platform. In classic HPC workloads, the rigidity of jobs tends to create holes in the schedule: we can use those idle resources as a dynamic pool for Big Data workloads. We propose a new idea based on Resource and Job Management System's (RJMS) configuration, that makes HPC and Big Data systems to communicate through a simple prolog/epilog mechanism. It leverages the built-in resilience of Big Data frameworks, while minimizing the disturbance on HPC workloads. We present the first study of this approach, using the production RJMS middleware OAR and Hadoop YARN from the HPC and Big Data ecosystems respectively. Our new technique is evaluated with real experiments upon the Grid5000 platform. Our experiments validate our assumptions and show promising results. The system is capable of running an HPC workload with 70% cluster utilization, with a Big Data workload that fills the schedule holes to reach a full 100% utilization. We observe a penalty on the mean waiting time for HPC jobs of less than 17% and a Big Data effectiveness of more than 68% in average.

7.2. Data Aware Batch Scheduling

New results on the topic *Data Aware Batch Scheduling* are related to graph algorithm for dense k-subset detection [8], scheduling heuristic for multi-CPU multi-GPU computing platform with performance guarantee [9], machine learning for designing scheduling policies [11] and multi-objective scheduling heuristic [13]. We detail the two last here.

- **Obtaining Dynamic Scheduling Policies with Simulation and Machine Learning [11].** Dynamic scheduling of tasks in large-scale HPC platforms is normally accomplished using ad-hoc heuristics, based on task characteristics, combined with some backfilling strategy. Defining heuristics that work efficiently in different scenarios is a difficult task, specially when considering the large variety of

task types and platform architectures. In this work, we present a methodology based on simulation and machine learning to obtain dynamic scheduling policies. Using simulations and a workload generation model, we can determine the characteristics of tasks that lead to a reduction in the mean slowdown of tasks in an execution queue. Modeling these characteristics using a nonlinear function and applying this function to select the next task to execute in a queue dramatically improved the mean task slowdown in synthetic workloads. When applied to real workload traces from highly different machines, these functions still resulted in important performance improvements, attesting the generalization capability of the obtained heuristics.

- **A new on-line method for scheduling independent tasks [13].** We present a new method for scheduling independent tasks on a parallel machine composed of identical processors. This problem has been studied extensively for a long time with many variants. We are interested here in designing a generic algorithm in the on-line non-preemptive setting whose performance is good for various objectives. The basic idea of this algorithm is to detect some problematic tasks that are responsible for the delay of other shorter tasks. Then the former tasks are redirected to be executed in a dedicated part of the machine. We show through an extensive experimental campaign that this method is effective and in most cases is closer to some standard lower bounds than the base-line method for the problem.

HIEPACS Project-Team

7. New Results

7.1. High-performance computing on next generation architectures

7.1.1. Bridging the gap between OpenMP and task-based runtime systems

With the advent of complex modern architectures, the low-level paradigms long considered sufficient to build High Performance Computing (HPC) numerical codes have met their limits. Achieving efficiency, ensuring portability, while preserving programming tractability on such hardware prompted the HPC community to design new, higher level paradigms while relying on runtime systems to maintain performance. However, the common weakness of these projects is to deeply tie applications to specific expert-only runtime system APIs. The OpenMP specification, which aims at providing common parallel programming means for shared-memory platforms, appears as a good candidate to address this issue thanks to the latest task-based constructs introduced in its revision 4.0. The goal of this paper is to assess the effectiveness and limits of this support for designing a high-performance numerical library, ScalFMM, implementing the fast multipole method (FMM) that we have deeply redesigned with respect to the most advanced features provided by OpenMP 4. We show that OpenMP 4 allows for significant performance improvements over previous OpenMP revisions on recent multicore processors and that extensions to the 4.0 standard allow for strongly improving the performance, bridging the gap with the very high performance that was so far reserved to expert-only runtime system APIs. More details on this work can be found in [17].

7.1.2. Modeling Irregular Kernels of Task-based codes: Illustration with the Fast Multipole Method

The significant increase of the hardware complexity that occurred in the last few years led the high performance community to design many scientific libraries according to a task-based parallelization. The modeling of the performance of the individual tasks (or kernels) they are composed of is crucial for facing multiple challenges as diverse as performing accurate performance predictions, designing robust scheduling algorithms, tuning the applications, etc. Fine-grain modeling such as emulation and cycle-accurate simulation may lead to very accurate results. However, not only their high cost may be prohibitive but they furthermore require a high fidelity modeling of the processor, which makes them hard to deploy in practice. In this paper, we propose an alternative coarse-grain, empirical methodology oblivious to both the target code and the hardware architecture, which leads to robust and accurate timing predictions. We illustrate our approach with a task-based Fast Multipole Method (FMM) algorithm, whose kernels are highly irregular, implemented in the `ScalFMM` library on top of the `StarPU` task-based runtime system and the simgrid simulator. More details on this work can be found in [41].

7.1.3. Task-based fast multipole method for clusters of multicore processors

Most high-performance, scientific libraries have adopted hybrid parallelization schemes - such as the popular MPI+OpenMP hybridization - to benefit from the capacities of modern distributed-memory machines. While these approaches have shown to achieve high performance, they require a lot of effort to design and maintain sophisticated synchronization/communication strategies. On the other hand, task-based programming paradigms aim at delegating this burden to a runtime system for maximizing productivity. In this article, we assess the potential of task-based fast multipole methods (FMM) on clusters of multicore processors. We propose both a hybrid MPI+task FMM parallelization and a pure task-based parallelization where the MPI communications are implicitly handled by the runtime system. The latter approach yields a very compact code following a sequential task-based programming model. We show that task-based approaches can compete with a hybrid MPI+OpenMP highly optimized code and that furthermore the compact task-based scheme fully matches the performance of the sophisticated, hybrid MPI+task version, ensuring performance while maximizing productivity. We illustrate our discussion with the `ScalFMM` FMM library and the `StarPU` runtime system. More details on this work can be found in [40].

7.1.4. Achieving high-performance with a sparse direct solver on Intel KNL

The need for energy-efficient high-end systems has led hardware vendors to design new types of chips for general purpose computing. However, designing or porting a code tailored for these new types of processing units is often considered as a major hurdle for their broad adoption. In this paper, we consider a modern Intel Xeon Phi processor, namely the Intel Knights Landing (KNL) and a numerical code initially designed for a classical multi-core system. More precisely, we consider the qr_mumps scientific library implementing a sparse direct method on top of the StarPU runtime system. We show that with a portable programming model (task-based programming), a good software support (a robust runtime system coupled with an efficient scheduler) and some well defined hardware and software settings, we are able to transparently run the exact same numerical code. This code not only achieves very high performance (up to 1 TFlop/s) on the KNL but also significantly outperforms a modern Intel Xeon multi-core processor both in terms of time to solution and energy efficiency up to a factor of 2.0. More details on this work can be found in [42].

7.2. High performance solvers for large linear algebra problems

7.2.1. Blocking strategy optimizations for sparse direct linear solver on heterogeneous architectures

The preprocessing steps of sparse direct solvers, ordering and block-symbolic factorization, are two major steps that lead to a reduced amount of computation and memory and to a better task granularity to reach a good level of performance when using BLAS kernels. With the advent of GPUs, the granularity of the block computation became more important than ever. In this paper, we present a reordering strategy that increases this block granularity. This strategy relies on the block-symbolic factorization to refine the ordering produced by tools such as METIS or Scotch, but it does not impact the number of operations required to solve the problem. We integrate this algorithm in the PaStiX solver and show an important reduction of the number of off-diagonal blocks on a large spectrum of matrices. This improvement leads to an increase in efficiency of up to 20% on GPUs.

These contributions have been published in SIAM Journal on Matrix Analysis and Applications [22].

7.2.2. Sparse supernodal solver using block low-rank compression

In the context of FASTLA associate team, during the last 4 years, we are collaborating with Eric Darve, professor in the Institute for Computational and Mathematical Engineering and the Mechanical Engineering Department at Stanford, on the design of a new efficient sparse direct solvers. We have been working on applying fast direct solvers for dense matrices to the solution of sparse direct systems. We observed that the extend-add operation (during the sparse factorization) is the most time-consuming step. We have therefore developed a series of algorithms to reduce this computational cost.

We presented two approaches using a Block Low-Rank (BLR) compression technique to reduce the memory footprint and/or the time-to-solution of the sparse supernodal solver PaStiX. This flat, non-hierarchical, compression method allows to take advantage of the low-rank property of the blocks appearing during the factorization of sparse linear systems, which come from the discretization of partial differential equations. The first approach, called *Minimal Memory*, illustrates the maximum memory gain that can be obtained with the BLR compression method, while the second approach, called *Just-In-Time*, mainly focuses on reducing the computational complexity and thus the time-to-solution. Singular Value Decomposition (SVD) and Rank-Revealing QR (RRQR), as compression kernels, are both compared in terms of factorization time, memory consumption, as well as numerical properties. Experiments on a single node with 24 threads and 128 GB of memory are performed to evaluate the potential of both strategies. On a set of matrices from real-life problems, we demonstrate a memory footprint reduction of up to 4 times using the *Minimal Memory* strategy and a computational time speedup of up to 3.5 times with the *Just-In-Time* strategy. Then, we study the impact of configuration parameters of the BLR solver that allowed us to solve a 3D laplacian of 36 million unknowns a single node, while the full-rank solver stopped at 8 million due to memory limitation.

These contributions have been presented at the PDSEC workshop of IPDPS'17 conference [30] and an extended version has been submitted in Journal of Computational Science [48].

7.2.3. Towards a hierachical symbol factorization for data sparse direct solvers

Hierarchical algorithms based on low-rank compression techniques have led to fully re-design the methods of solving dense linear systems at the dawn of the twenty-first century, significantly reducing the computational costs. However, their application to the treatment of sparse linear systems remains today a major challenge to which both the community of hierarchical matrices and that of the sparse matrices are tackling. For this purpose, a first class of approach has been developed by the community of hierarchical matrices to exploit the sparse matrix structure. If the strong point of these methods is that the resulting algorithm remains hierarchical, these do not manage exploit some zeros as naturally do sparse solvers. In contrast, the fact that a sparse factorization can be seen as a sequence of smaller, dense operations, the community of hollow treasure has explored this property to introduce hierarchical techniques within these elementary operations. However, the resulting algorithm loses the fundamental property of hierarchical algorithms, since the compression hierarchy is only local. As part of this doctorate, we introduce a new algorithm, performing a sparse hierarchical symbolic factorization that allows to exploit precisely the sparse structure the matrix and its factors while preserving a global hierarchical structure for to ensure effective compression. We have shown experimentally that this new approach allows us to obtain at the same time a reduced number of operations (because of its hierarchical character) and a number of non-zero elements as small as a hollow method (through the use of a symbolic factorization).

This work is developped in the A. Falco PhD thesis, it led to a publication in a national conference [31] and will give rise to a submission in an international journal in 2018

7.3. High performance fast multipole method for N-body problems

7.3.1. Modeling Irregular Kernels of Task-based codes

The significant increase of the hardware complexity that occurred in the last few years led the high performance community to design many scientific libraries according to a task-based parallelization. The modeling of the performance of the individual tasks (or kernels) they are composed of is crucial for facing multiple challenges as diverse as performing accurate performance predictions, designing robust scheduling algorithms, tuning the applications, etc. Fine-grain modeling such as emulation and cycle-accurate simulation may lead to very accurate results. However, not only their high cost may be prohibitive but they furthermore require a high fidelity modeling of the processor, which makes them hard to deploy in practice. In this paper, we propose an alternative coarse-grain, empirical methodology oblivious to both the target code and the hardware architecture, which leads to robust and accurate timing predictions. We illustrate our approach with a task-based Fast Multipole Method (FMM) algorithm, whose kernels are highly irregular, implemented in the `ScalFMM` library on top of the starpu task-based runtime system and the simgrid simulator. More details on this work can be found in [41].

7.3.2. Task-based fast multipole method for clusters of multicore processors

Most high-performance, scientific libraries have adopted hybrid parallelization schemes - such as the popular MPI+OpenMP hybridization - to benefit from the capacities of modern distributed-memory machines. While these approaches have shown to achieve high performance, they require a lot of effort to design and maintain sophisticated synchronization/communication strategies. On the other hand, task-based programming paradigms aim at delegating this burden to a runtime system for maximizing productivity. In this article, we assess the potential of task-based fast multipole methods (FMM) on clusters of multicore processors. We propose both a hybrid MPI+task FMM parallelization and a pure task-based parallelization where the MPI communications are implicitly handled by the runtime system. The latter approach yields a very compact code following a sequential task-based programming model. We show that task-based approaches can compete with a hybrid MPI+OpenMP highly optimized code and that furthermore the compact task-based scheme fully matches the performance of the sophisticated, hybrid MPI+task version, ensuring performance while

maximizing productivity. We illustrate our discussion with the ScalFMM FMM library and the StarPU runtime system. More details on this work can be found in [40].

7.4. Efficient algorithmic for load balancing and code coupling in complex simulations

7.4.1. Comparison of initial partitioning methods for multilevel direct k-way graph partitioning with fixed vertices

In scientific computing, load balancing is a crucial step conditioning the performance of large-scale applications. In this case, an efficient decomposition of the workload to a number of processors is highly necessary. A common approach to solve this problem is to use graph representation and perform a graph partitioning in k parts using the multilevel framework and the recursive bisection (RB) paradigm. However, in graph instances where fixed vertices are used to model additional constraints, RB often produces partitions of poor quality.In this paper, we investigate the difficulties of RB to handle fixed vertices and we compare its results with two different alternatives. The first one, called KGGGP is a direct k-way greedy graph growing partitioning that properly handles fixed vertices while the second one, introduced in kPaToH, uses RB and a post-processing technique to correct the obtained partition. Finally, experimental results on graphs that represent real-life numerical simulations show that both alternative methods provide improved partitions compared to RB. More details on this work can be found in [23].

7.5. Application Domains

7.5.1. Material physics

7.5.1.1. EigenSolver

The adaptive vibrational configuration interaction algorithm has been introduced as a new method to efficiently reduce the dimension of the set of basis functions used in a vibrational configuration interaction process. It is based on the construction of nested bases for the discretization of the Hamiltonian operator according to a theoretical criterion that ensures the convergence of the method. In the present work, the Hamiltonian is written as a sum of products of operators. The purpose of this paper is to study the properties and outline the performance details of the main steps of the algorithm. New parameters have been incorporated to increase flexibility, and their influence has been thoroughly investigated. The robustness and reliability of the method are demonstrated for the computation of the vibrational spectrum up to 3000 cm-1 of a widely studied 6-atom molecule (acetonitrile). Our results are compared to the most accurate up to date computation; we also give a new reference calculation for future work on this system. The algorithm has also been applied to a more challenging 7-atom molecule (ethylene oxide). The computed spectrum up to 3200 cm-1 is the most accurate computation that exists today on such systems. More details on this work can be found in [43], [21].

7.5.1.2. Dislocation

We have focused on the improvements of the parallel collision detection and of the accuracy in the force field computation in the OPTIDIS code.

- a new collision detection algorithm to reliably handle junction formation for Dislocation Dynamics using hybrid OpenMP + MPI parallelism has been developed. The enhanced precision and reliability of this new algorithm allows the use of larger time-steps for faster simulations. Hierarchical methods for collision detection, as well as hybrid parallelism are also used to improve performance;
- we observed that the force field computation depends on how the traversal of the segments list or boxes in the octree was done. New accurate formulas to remove this issue have been developed and we are implementing them in the code. They will be used in the Fast Multipole Method that we have developed previously.

Finally, a new distributed data structure has been developed to enhance the reliability and modularity of OPTIDIS. The new data structure provides an interface to modify safely and reliably the distributed dislocation mesh in order to enforce data consistency across all computation nodes. This interface also improves code modularity allowing the study of data layout performance without modifying the algorithms.

7.5.2. Co-design for scalable numerical algorithms in scientific applications

7.5.2.1. High performance simulation for ITER tokamak

Concerning the GYSELA global non-linear electrostatic code, the efforts during the period have concentrated on the design of a more efficient parallel gyro-average operator for the deployment of very large (future) GYSELA runs. The main unknown of the computation is a distribution function that represents either the density of the guiding centers, either the density of the particles in a tokamak. The switch between these two representations is done thanks to the gyro-average operator. In the previous version of GYSELA, the computation of this operator was achieved thanks to a Padé approximation. In order to improve the precision of the gyro-averaging, a new parallel version based on an Hermite interpolation has been done (in collaboration with the Inria TONUS project-team and IPP Garching). The integration of this new implementation of the gyro-average operator has been done in GYSELA and the parallel benchmarks have been successful. This work is carried on in the framework of the PhD of Nicolas Bouzat (funded by IPL C2S@EXA) co-advised with Michel Mehrenberger from TONUS project-team and in collaboration with Guillaume Latu from CEA-IRFM. The scientific objectives of this work is first to consolidate the parallel version of the gyro-average operator, in particular by designing a scalable MPI+OpenMP parallel version and using a new communication scheme, and second to design new numerical methods for the gyro-average, source and collision operators to deal with new physics in GYSELA. The objective is to tackle kinetic electron configurations for more realistic complex large simulations.

In the context of the EoCoE project, we have collaborations with CEA-IRFM. First, with G. Latu, we have investigated the potential of using the last release of the PaStiX solver (version 6.0) on Intel KNL architecture, and more especially on the MARCONI machine (one of the PRACE supercomputers at Cineca, Italia). The results obtained on this architecture are really promising since we are able to reach more than 1 Tflops using a single node. Secondly, we also have a collaboration with P. Tamain and G. Giorgani on the TOKAM3X code to analyze the performance of using PaStiX as a preconditioner. Since a distributed memory is required during the simulation, the previous release of PaStiX is then used. Some difficulties regarding the Fortran wrapper and some memory issues should be fixed when we will have reimplemented the MPI interface in the current release.

7.5.2.2. High performance simulation for 3D frequency-domain Maxwell's equations

We also recently developed a collaboration with NACHOS on the HORSE (High Order solver for Radar cross Section Evaluation) simulation code. The aim was to integrate the PaStiX solver, with low-rank compression technique, in a domain decomposition framework to solve 3D frequency-domain Maxwell's equations. The results are promising since we were able to reduce by two the factorization and the solve time for each subdomain. And we were also able to reduce by two the memory requirements thanks to our compression techniques. This would allow us to consider larger subdomains with the same memory constraints that currently limit the simulations.

7.5.2.3. High performance simulation for atmospheric chemistry

We worked on the development and tests of the Adaptative Semi-Implicit Scheme (ASIS) solver for the simulation of atmospheric chemistry. To solve the Ordinary Differential Equation systems associated with the time evolution of the species concentrations, ASIS adopts a one step linearized implicit scheme with specific treatments of the Jacobian of the chemical fluxes. It conserves mass and has a time stepping module to control the accuracy of the numerical solution. In 5 idealized box model simulations ASIS gives results similar to the higher order implicit schemes derived from the Rosenbrock's and Gear's methods and requires less computation and run time at the moderate precision required for atmospheric applications. When implemented in the MOCAGE CTM and the LMD Mars GCM the ASIS solver performs well and reveals weaknesses and limitations of the original semi-implicit solvers used by these two models. ASIS can be easily adapted to

various chemical schemes and further developments are foreseen to increase its computational efficiency, and to include the computation of the 10 concentrations of the species in aqueous phase in addition to gas phase chemistry.

More details on this work can be found in [19].

KERDATA Project-Team

6. New Results

6.1. Convergence of HPC and Big Data

6.1.1. Týr: Blob-based storage convergence of HPC and Big Data

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

The increasingly growing data sets processed on HPC platforms raise major challenges for the underlying storage layer. A promising alternative to POSIX-I/O-compliant file systems are simpler blobs (binary large objects), or object storage systems. They offer lower overhead, better performance and horizontal scalability at the cost of largely unused features such as file hierarchies or permissions. Similarly, blobs are increasingly considered for replacing distributed file systems for big data analytics or as a base for storage abstractions like key-value stores or time-series databases.

This growing interest from both HPC and Big Data communities towards blob storage naturally fits with the current trend towards HPC and Big Data convergence. In this context, we seek to demonstrate that blob storage indeed constitutes a strong alternative to current storage infrastructures. Additionally, the data model of blob storage is close enough to that of distributed file systems so that this change is largely transparent for the applications running atop them.

In [22] we provide a preliminary evaluation of blob storage in HPC and Big Data contexts. We leverage a series of real-world HPC applications as well as an industry-standard HPC benchmark. We analyze for each of these applications the storage requests sent to the underlying storage system. We discover that over 98% of these storage calls can be directly mapped to the data model offered by blobs. Interestingly, we also note that the remaining calls are using file systems features for convinience rather than by necessity. These calls may consequently be performed as offline pre- or post-processing, or avoided altogether without altering the application.

6.1.2. Modeling elastic storage

Participants: Nathanaël Cheriere, Gabriel Antoniu.

For efficient Big Data processing, efficient resource utilization becomes a major concern as large-scale computing infrastructures such as supercomputers or clouds keep growing in size. Naturally, energy and cost savings can be obtained by reducing idle resources. Malleability, which is the possibility for resource managers to *dynamically* increase or reduce the resources of jobs, appears as a promising means to progress towards this goal.

However, state-of-the-art parallel and distributed file systems have not been designed with malleability in mind. This is mainly due to the supposedly high cost of storage decommission, which is considered to involve expensive data transfers. Nevertheless, as network and storage technologies evolve, old assumptions on potential bottlenecks can be revisited.

In [18], we evaluate the viability of malleability as a design principle for a distributed file system. We specifically model the duration of the decommission operation, for which we obtain a theoretical lower bound. Then we consider HDFS as a use case and we show that our model can explain the measured decommission times.

The existing decommission mechanism of HDFS is good when the network is the bottleneck, but could be accelerated by up to a factor 3 when the storage is the limiting factor. With the highlights provided by our model, we suggest improvements to speed up decommission in HDFS and we discuss open perspectives for the design of efficient malleable distributed file systems.

6.1.3. Eley: Leveraging burst-buffers for efficient Big Data processing on HPC systems

Participants: Orçun Yildiz, Chi Zhou, Shadi Ibrahim.

Burst Buffer is an effective solution for reducing the data transfer time and the I/O interference in HPC systems. Extending Burst Buffers (BBs) to handle Big Data applications is challenging because BBs must account for the large data inputs of Big Data applications and the performance guarantees of HPC applications – which are considered as first-class citizens in HPC systems. Existing BBs focus on only intermediate data of Big Data applications and incur a high performance degradation of both Big Data and HPC applications.In [26], we present *Eley*, a burst buffer solution that helps to accelerate the performance of Big Data applications while guaranteeing the performance of HPC applications. In order to improve the performance of Big Data applications, *Eley* employs a prefetching technique that fetches the input data of these applications to be stored close to computing nodes thus reducing the latency of reading data inputs. Moreover, Eley is equipped with a full delay operator to guarantee the performance of HPC applications – as they are running independently on a HPC system. The experimental results show the effectiveness of *Eley* in obtaining shorter execution time of Big Data applications (shorter map phase) while guaranteeing the performance of HPC applications.

6.2. Scalable data processing on clouds

6.2.1. Low-latency storage for stream processing

Participants: Ovidiu-Cristian Marcu, Alexandru Costan, Gabriel Antoniu, María Pérez, Radu Tudoran, Stefano Bortoli, Bogdan Nicolae.

We are now witnessing an unprecedented growth of data that needs to be processed at always increasing rates in order to extract valuable insights. Big Data applications are rapidly moving from a batch-oriented execution model to a streaming execution model in order to extract value from the data in real-time. Big Data streaming analytics tools have been developed to cope with the online dimension of data processing: they enable real-time handling of live data sources by means of stateful aggregations (window-based operators). In [21] we design a deduplication method specifically for window-based operators that rely on key-value stores to hold a shared state. Our key finding is that more fine-grained interactions between streaming engines and (key-value) stores (i.e., the data ingest, store, and process interfaces) need to be designed in order to better respond to scenarios that have to overcome memory scarcity.

Moreover, processing live data alone is often not enough: in many cases, such applications need to combine the live data with previously archived data to increase the quality of the extracted insights. Current streaming-oriented runtimes and middlewares are not flexible enough to deal with this trend, as they address ingestion (collection and pre-processing of data streams) and persistent storage (archival of intermediate results) using separate services. This separation often leads to I/O redundancy (e.g., write data twice to disk or transfer data twice over the network) and interference (e.g., I/O bottlenecks when collecting data streams and writing archival data simultaneously). In [20] and [27] we argue for a unified ingestion and storage architecture for streaming data that addresses the aforementioned challenge and we identify a set of constraints and benefits for such a unified model, while highlighting the important architectural aspects required to implement it in real life.

Based on these findings, we are currently developing a low-latency stream storage framework that addresses such critical real-time needs for efficient stream processing, exposing high-performance interfaces for stream ingestion, storage, and processing.

6.2.2. A Performance Evaluation of Apache Kafka in Support of Big Data Streaming Applications

Participants: Paul Le Noac'h, Alexandru Costan.

Stream computing is becoming a more and more popular paradigm as it enables the real-time promise of data analytics. Apache Kafka is currently the most popular framework used to ingest the data streams into the processing platforms. However, how to tune Kafka and how much resources to allocate for it remains a challenge for most users, who now rely mainly on empirical approaches to determine the best parameter settings for their deployments. Our goal in [28] is to make a through evaluation of several configurations and performance metrics of Kafka in order to allow users avoid bottlenecks, reach its full potential and avoid bottlenecks and eventually leverage some good practice for efficient stream processing.

6.2.3. Hot metadata management for geographically distributed workflows

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu, Ji Liu, Esther Pacitti, Patrick Valduriez, Marta Mattoso.

Large-scale scientific applications are often expressed as scientific workflows (SWfs) that help defining data processing jobs and dependencies between jobs' activities. Several SWfs have huge storage and computation requirements, and so they need to be processed in multiple (cloud-federated) datacenters. It has been shown that efficient metadata handling plays a key role in the performance of computing systems. However, most of this evidence concern only single-site, HPC systems to date. In addition, the efficient scheduling of tasks among different datacenters is critical to the SWf execution. In [19], we present a hybrid distributed model and architecture, using hot metadata (frequently accessed metadata) for efficient SWf scheduling in a multisite cloud. We couple our model with a scientific workflow management system (SWfMS) to validate its applicability to real-life scientific workflows with different scheduling algorithms. We show that the combination of efficient management of hot metadata and scheduling algorithms improves the performance of SWfMS, reducing the execution time of highly parallel jobs up to 64.1 % and that of the whole scientific workflows up to 37.5 %, by avoiding unnecessary cold metadata operations. We also further discuss how to dynamically handle such hot metadata.

6.3. Scalable I/O, storage and in-situ processing in Exascale environments

6.3.1. Extreme-scale logging through application-defined storage

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Applications generating data as logs and seeking to store it as such face hard challenges on HPC platforms. In distributed systems this storage model is key to ensuring fault-tolerance, developing transactional systems or publish-subscribe models. In scientific applications, distributed logs can play many roles such as in-situ visualization of large data streams, centralized collection of telemetry or monitoring events computational steering, data aggregation from array of physical sensors or live data indexing. Distributed shared logs are very difficult to implement on common HPC platforms due to the lack of efficient append operation in the current file-based storage infrastructures. While part of the POSIX standard, this operation has not been the main focus during the development of parallel file systems. While application-specific, custom-built solutions are possible, they require a significant development effort and often fail to meet the performance requirements of data-intensive applications running at large scale.

In this work we go through the basic requirements of storing telemetry data streams for computational steering and visualization. For simple use cases where the telemetry data is only temporary, we prove that distributed logging can be performed at scale by leveraging state-of-the-art blob storage systems such as Týr or RADOS. This approach is supported by the growing availability of node-local storage on a new generation of supercomputers, giving application developers the freedom to deploy transient storage systems alongside the application directly on the compute nodes.

When long-term storage of the generated data is needed for offline visualization or analytics, we prove that distributed logs require a significantly lower number of output logs to achieve peak performance compared to Lustre or GPFS. We also prove that this low number of output files obviates the need for an explicit post-processing merge step in most cases for iterating the whole output log in generation order. We finally prove on up to 100,000 cores of the Theta supercomputer that our findings are applicable to run distributed logging at large scale, while improving write throughput by several orders of magnitude compared to Lustre of GPFS.

6.3.2. Leveraging Damaris for in-situ visualization in support of GeoScience and CFD Simulations

Participants: Hadi Salimi, Matthieu Dorier, Luc Bougé.

Damaris is a middleware for in situ data analysis and visualization targeting extreme-scale, MPI-based simulations. The main goal of Damaris is to provide a simple method to instrument a simulation in order to benefit from in situ analysis and visualization. To this aim, the computing resources are partitioned such that a subset of cores in a SMP node or a subset of nodes of the underlying platform are dedicated to in situ processing. The data generated by the simulation are passed to these dedicated processes either through shared memory (in the case of dedicated cores) or through the MPI calls (in the case of dedicated nodes) and can be processed both in synchronous and asynchronous modes. Afterwards, the processed data can be analyzed or visualized. Damaris also supports a very simple API to instrument simulations developed in different domains. Moreover, using some XML configuration files for defining simulation data types (e.g. meshes) makes the instrumentation process easier with minimum code modifications. Active development is currently continuing within the KerData team, where it is at the center of several collaborations with industry (e.g Total) as well as with national and international academic partners.

In recent developments of Damaris, we have focused on two main targets that are: 1) Instrumenting new simulations codes from different scientific domains, i.e. geoscience and ocean modeling, 2) Implementing new storage backends, i.e. HDF5 for Damaris. In this regard, we report the results of some experiments we made to evaluate Damaris with respect to performance. These experiments were conducted on Grid'5000 test bed. In these experiments Damaris was employed to visualize the data generated by the Wave Propagation geoscience simulation and also the CROCO coastal and ocean simulation. During the experiments, the impact of Damaris was measured by comparing the simulations instrumented by Damaris (space partitioning approach) with a baseline where those simulations include data processing codes directly on their source code (time partitioning approach). The results of these simulations show that the incorporation of Damaris into a simulation decreases the total run time of the simulation due to its asynchronous data processing and visualization capabilities. In addition, using Damaris for data visualization has nearly no impact on the total run time of the mentioned simulation codes. We also have shown that the amount of code changes necessary for instrumenting the simulation codes is much less compared to the case that the simulation code is instrumented by native visualization or storage APIs. Moreover, we also have studied the impact of new HDF5 storage backend, on storing simulation results in HDF5 format in both file-per-dedicated-core and collective I/O scenarios.

6.3.3. Accelerating MPI collective operations on the Theta cupercomputer

Participants: Nathanaël Cheriere, Matthieu Dorier, Misbah Mubarak, Robert Ross, Gabriel Antoniu.

Recent network topologies in supercomputers have motivated new research on topology-aware collective communication algorithms for MPI. But such endeavor requires betting on the fact that topology-awareness is the primary factor to accelerate these collective operations. Besides, showing the benefit of a new, topology-aware algorithm requires not only access to a leadership-scale supercomputer with the desired topology, but also a large resource allocation on this supercomputer. Event-driven network simulations can alleviate such constraints and speed up the search for appropriate algorithms by providing early answers on their relative merit.

In our studies, we focus on the Scatter and AllGather operations in the context of the Theta supercomputer's dragonfly topology. We propose a set of topology-aware versions of these operations as well as optimizations of the old, non-topology-aware ones. We conduct an extensive simulation campaign using the CODES network simulator. Our results show that, contrary to our expectations, topology-awareness does not help improving significantly the speed of these operations. Rather, the high radix and low diameter of the dragonfly topology, along with already good routing protocols, enable simple algorithms based on non-blocking communications to perform better than state-of-the-art algorithms. A trivial implementation of Scatter using nonblocking point-to-point communications can be faster than state-of-the art algorithms by up to a factor of 6. Traditional AllGather algorithms can also be improved by the same principle and exhibit a 4x speedup in some

situations. These results highlight the need to rethink the collective operations under the light of nonblocking communications.

6.4. Energy-aware data storage and processing at large scale

6.4.1. Performance and energy-efficiency trade-offs in in-memory storage systems

Participants: Mohammed-Yacine Taleb, Shadi Ibrahim, Gabriel Antoniu, Toni Cortes.

Most large popular web applications, like Facebook and Twitter, have been relying on large amounts of in-memory storage to cache data and offer a low response time. As the main memory capacity of clusters and clouds increases, it becomes possible to keep most of the data in the main memory. This motivates the introduction of in-memory storage systems. While prior work has focused on how to exploit the low-latency of in-memory access at scale, there is very little visibility into the energy-efficiency of in-memory storage systems. Even though it is known that main memory is a fundamental energy bottleneck in computing systems (i.e., DRAM consumes up to 40% of a server's power). During this project, by the means of experimental evaluation, we have studied the performance and energy-efficiency of RAMCloud - a well-known in-memory storage system. We reveal that although RAMCloud is scalable for read-only applications, it exhibits non-proportional power consumption. We also find that the current replication scheme implemented in RAMCloud limits the performance and results in high energy consumption. Surprisingly, we show that replication can also play a negative role in crash-recovery.

6.4.2. Energy-aware straggler mitigation in Map-Reduce

Participants: Tien-Dat Phan, Chi Zhou, Shadi Ibrahim, Guillaume Aupy, Gabriel Antoniu.

Energy consumption is an important concern for large-scale data-centers, which results in huge monetary cost for data-center operators. Due to the hardware heterogeneity and contentions between concurrent workloads, straggler mitigation is important to many Big Data applications running in large-scale data-centers and the speculative execution technique is widely-used to handle stragglers. Although a large number of studies have been proposed to improve the performance of Big Data applications using speculative execution, few of them have studied the energy efficiency of their solutions.

In [23], we propose two techniques to improve the energy efficiency of speculative executions while ensuring comparable performance. Specifically, we propose a hierarchical straggler detection mechanism which can greatly reduce the number of killed speculative copies and hence save the energy consumption. We also propose an energy-aware speculative copy allocation method which considers the trade-off between performance and energy when allocating speculative copies. We implement both techniques into Hadoop and evaluate them using representative Map-Reduce benchmarks. Results show that our solution can reduce the energy waste on killed speculative copies by up to 100% and improve the energy efficiency by 20% compared to state-of-the-art mechanisms.

POLARIS Team

7. New Results

7.1. Simgrid for MPI (SMPI)

Several new results on the usage of Simgrid to assess MPI performance have been published in 2017. The general framework introducing the methodology for a proper use of SimGrid to simulate MPI applications was presented in [5]. One more specific line of work concerns the prediction of the performance and the energy consumption of MPI applications using SimGrid [39], [19]. Other applications have also been simulated using this approach. General capacity planning of supercomputers is analyzed in [35] using simulation. More specifically, we have shown that the HPL benchmark (high Performance Linpack), used to establish the top 500 ranking of the most powerful supercomputers in the world, can be emulated faithfully on a commodity server, at the scale of a supercomputer [36]. We have also shown that SimGrid is reliable and fast enough to evaluate and tune the performance of dynamic load balancing in seismic simulations [22].

7.2. Visualisation for Performance Analysis of Task-Based Applications

The performance of task-based application heavily depends on the runtime scheduling and on its ability to exploit computing and communication resources. Unfortunately, the traditional performance analysis strategies are unfit to fully understand task-based runtime systems and applications: they expect a regular behavior with communication and computation phases, while task-based applications demonstrate no clear phases. Moreover, the finer granularity of task-based applications typically induces a stochastic behavior that leads to irregular structures that are difficult to analyze. We have introduced a flexible framework combining visualization panels to understand and pinpoint performance problems incurred by bad scheduling decisions in task-based applications. Three case-studies using StarPU-MPI, a task-based multi-node runtime system, have been investigated in more details to show how our framework is used to study the performance of the well-known Cholesky factorization. Performance improvements include a better task partitioning among the multi-(GPU,core) to get closer to theoretical lower bounds, improved MPI pipelining in multi-(node,core,GPU) to reduce the slow start, and changes in the runtime system to increase MPI bandwidth, with gains of up to 13% in the total makespan [38].

7.3. Convergence of game dynamics

The study of game dynamics is crucial in understanding the long-run behavior of optimizing agents in an environment that changes dynamically over time, whether endogenously (i.e. via the agents' interactions) or exogenously (i.e. due to factors beyond the agents' influence). Starting with the observation that oblivious agents should seek to at least minimize their regret, we showed in [9] that players that "follow the regularized leader" in continuous time achieve no regret at an optimal rate. The multi-agent implications of this property where subsequently explored in [3], [24] (for games with finite and continuous action sets respectively), where we established a wide range of conditions guaranteeing convergence to Nash equilibrium, even when the players' payoff observations are subject to noise and/or other stochastic disturbances.

7.4. Multi-agent learning

In contrast to [9], [3], [24], the above works focus squarely on multi-agent interactions that occur in discrete time (as is typically the case in practical applications). In the case of games with finite action spaces, we showed in [16] that no-regret learning based on "following the regularized leader" converges to Nash equilibrium in potential games, thus complementing the analysis of [15] where it was shown that this family of learning methods eliminates dominated strategies and converges locally to strict Nash equilibria. The former result was extended to mixed-strategy learning in games with continuous action spaces in [11], while [42], [28] established the convergence of no-regret regularized learning to variationally stable equilibria in continuous games, even with imperfect and/or delayed/asynchronous feedback.

7.5. Selfishness vs efficiency in traffic networks

Empirical studies in real-world networks show that the efficiency ratio between selfishly and socially optimal states (the so-called price of anarchy) is close to 1 in both light and heavy traffic conditions, thus raising the question: can these observations be justified theoretically? In [17] we showed that this is not always the case: the price of anarchy may remain bounded away from 1 for all values of the traffic inflow, even in simple three-link networks with a single O/D pair and smooth, convex costs. On the other hand, for a large class of cost functions (including all polynomials), the price of anarchy does converge to 1 in both heavy and light traffic conditions, and irrespective of the network topology and the number of O/D pairs in the network.

7.6. Online Energy Optimization in Embedded Systems

We have used a Markov Decision Process (MDP) approach to compute the optimal on-line speed scaling policy to minimize the energy consumption of a single processor executing a finite or infinite set of jobs with real-time constraints. We provide several qualitative properties of the optimal policy: monotonicity with respect to the jobs parameters, comparison with on-line deterministic algorithms. Numerical experiments in several scenarios show that our proposition performs well when compared with off-line optimal solutions and out-performs on-line solutions oblivious to statistical information on the jobs [33]. Several extension to online learning (Q-learning) as well as hidden Markov chain theory for offline computation of the statistical parameters of the system are currently being investigated.

7.7. Asymptotic Models

- Mean field approximation is a popular means to approximate large and complex stochastic models that can be represented as N interacting objects. The idea of mean field approximation to study the limit of this system as N goes to infinity.

  In [18], we study how accurate is mean field approximation as N goes to infinity. We show that under very general conditions the expectation of any performance indicator converges at rate $O(1/N)$ to its mean field approximation. In [7] we continue this analysis and establish a result that expresses the constant associated with this $1/N$ term. This allows us to propose what we call a *refined mean field approximation*. By considering a variety of applications, we illustrate that the proposed refined mean field approximation is significantly more accurate that the classic mean field approximation for small and moderate values of N: the relative errors of this refined approximation is often below 1% for systems with $N = 10$.
- Computer system and network performance can be significantly improved by caching frequently used information. When the cache size is limited, the cache replacement algorithm has an important impact on the effectiveness of caching. In [8] we introduce time-to-live (TTL) approximations to determine the cache hit probability of two classes of cache replacement algorithms: h-LRU and LRU(m). Using a mean field approach, we provide both numerical and theoretical support for the claim that the proposed TTL approximations are asymptotically exact. We use this approximation and trace-based simulation to compare the performance of h-LRU and LRU(m). First, we show that they perform alike, while the latter requires less work when a hit/miss occurs. Second, we show that as opposed to LRU, h-LRU and LRU(m) are sensitive to the correlation between consecutive inter-request times. Last, we study cache partitioning. In all tested cases, the hit probability improved by partitioning the cache into different parts—each being dedicated to a particular content provider. However, the gain is limited and the optimal partition sizes are very sensitive to the problem's parameters.
- Mean field approximation is often use to characterize the transient or steady state performance of a stochastic system. In [6], we use this approach to compute absorbing times. We use mean field approximation to provide an asymptotic expansion of this absorbing time that uses the spectral decomposition of the kernel of the original chains. Our results rely on extreme values theory. We show the applicability of this approach with three different problems: the coupon collector, the erasure channel lifetime and the coupling times of random walks in high dimensional spaces.

7.8. Secret Key Generation

Secret key generation (SKG) from shared randomness at two remote locations has been shown to be vulnerable to denial of service attacks in the form of jamming. In [13], [2], we develop as a novel counter-jamming approach by using energy harvesting. The idea is that part of the jamming signal can potentially be harvested and converted into useful communication power. In [14], we investigate the use of frequency hopping/spreading in Rayleigh block fading additive white Gaussian noise (BF-AWGN) channels to counteract attacks from jamming. In both cases, we formulate the problems as a zero-sum game and characterize the unique Nash equilibrium of the game in closed form. Through numerical evaluations, we show that energy harvesting is an efficient counter-jamming approach that offers substantial gains in terms of relative SKG rates. In the case of BF-AWGN channels, we also use numerical results to show that frequency hopping/spreading is an effective technique for combating jamming attacks in SKG systems; a modest increase of the system bandwidth can substantially increase the SKG rates.

7.9. Power control in wireless systems

Channel state information (CSI) is essential for efficient power and spectrum allocation policies. In cognitive radio (CR) channels, although perfect CSI of the direct link (between the secondary transmitter and the secondary receiver) is a reasonable assumption at the secondary transmitter (ST), however, perfect knowledge of its interfering links to the primary receivers (PRs) is not. Power allocation and scheduling algorithms are often based on perfect global CSI at the secondary transmitter (ST). In [31], we analyze the impact of channel estimation errors on both the secondary and primary users. On the one hand, the robustness of water-filling type of algorithms allowing the secondary user (SU) to minimize its power consumption under QoS and CR interference power constraints to channel estimation errors in the SU interfering links is analyzed. On the other hand, the impact of these estimation errors on the PU interference constraints is also analyzed. To this aim, we consider the worst case with respect to these estimation errors. Our analysis shows that the water-filling algorithm provides robustness in terms of power consumption and scheduling of the SU given the realistic estimation error models especially when the SU is overestimating the interfering power gains. We also provide possible solutions to ensure that the created interference is below the tolerated thresholds.

7.10. Routing in SDN-based networks

A new adaptive multi-flow routing algorithm to select end-to-end paths in packet-switched networks has been proposed. This algorithm provides provable optimality guarantees in the following game theoretic sense: The network configuration converges to a configuration arbitrarily close to a pure Nash equilibrium. This algorithm has several robustness properties making it suitable for real-life usage: it is robust to measurement errors, outdated information, and clocks desynchronization. Furthermore, it is only based on local information and only takes local decisions, making it suitable for a distributed implementation. Our SDN-based proof-of-concept is built as an Openflow controller. We also set up an emulation platform based on Mininet to test the behavior of our proof-of- concept implementation in several scenarios. Although real-world conditions do not conform exactly to the theoretical model, all experiments exhibit satisfying behavior, in accordance with the theoretical predictions [20], [21].

7.11. Distributed Best Response

We have designed and analyzed distributed algorithms to compute a Nash equilibrium in random potential games. Our algorithms are based on best-response dynamics, with suitable revision sequences (orders of play). We compute the average complexity over all potential games of best response dynamics under a random i.i.d. revision sequence, since it can be implemented in a distributed way using Poisson clocks. We obtain a distributed algorithm whose execution time is within a constant factor of the optimal centralized one. We also showed how to take advantage of the structure of the interactions between players in a network game: non-interacting players can play simultaneously. This improves best response algorithm, both in the centralized and in the distributed case [32].

ROMA Project-Team

7. New Results

7.1. Acyclic partitioning of large directed acyclic graphs

Participant: Bora Uçar.

Finding a good partition of a computational directed acyclic graph associated with an algorithm can help find an execution pattern improving data locality, conduct an analysis of data movement, and expose parallel steps. The partition is required to be acyclic, i.e., the inter-part edges between the vertices from different parts should preserve an acyclic dependency structure among the parts. In this work [26], we adopt the multilevel approach with coarsening, initial partitioning, and refinement phases for acyclic partitioning of directed acyclic graphs and develop a direct k-way partitioning scheme. To the best of our knowledge, no such scheme exists in the literature. To ensure the acyclicity of the partition at all times, we propose novel and efficient coarsening and refinement heuristics. The quality of the computed acyclic partitions is assessed by computing the edge cut, the total volume of communication between the parts, and the critical path latencies. We use the solution returned by well-known undirected graph partitioners as a baseline to evaluate our acyclic partitioner, knowing that the space of solution is more restricted in our problem. The experiments are run on large graphs arising from linear algebra applications.

7.2. Further notes on Birkhoff-von Neumann decomposition of doubly stochastic matrices

Participants: Ioannis Panagiotas, Bora Uçar.

The well-known Birkhoff-von Neumann (BvN) decomposition expresses a doubly stochastic matrix as a convex combination of a number of permutation matrices. For a given doubly stochastic matrix, there are many BvN decompositions, and finding the one with the minimum number of permutation matrices is NP-hard. There are heuristics to obtain BvN decompositions for a given doubly stochastic matrix. A family of heuristics are based on the original proof of Birkhoff and proceed step by step by subtracting a scalar multiple of a permutation matrix at each step from the current matrix, starting from the given matrix. At every step, the subtracted matrix contains nonzeros at the positions of some nonzero entries of the current matrix and annihilates at least one entry, while keeping the current matrix nonnegative. Our first result shows that this family of heuristics can miss optimal decompositions. We also investigate the performance of two heuristics from this family theoretically [46].

7.3. Low-Cost Approximation Algorithms for Scheduling Independent Tasks on Hybrid Platforms

Participants: Louis-Claude Canon, Loris Marchal, Frédéric Vivien.

Hybrid platforms embedding accelerators such as GPUs or Xeon Phis are increasingly used in computing. When scheduling tasks on such platforms, one has to take into account that a task execution time depends on the type of core used to execute it. We focus on the problem of minimizing the total completion time (or makespan) when scheduling independent tasks on two processor types, also known as the $(Pm, Pk)||C_{\max}$ problem. We propose BalanceEstimate and BalanceMakespan, two novel 2-approximation algorithms with low complexity. Their approximation ratio is both on par with the best approximation algorithms using dual approximation techniques (which are, thus, of high complexity) and significantly smaller than the approximation ratio of existing low-cost approximation algorithms. We compared both algorithms by simulations to existing strategies in different scenarios. These simulations showed that their performance is among the best ones in all cases.

This work has been presented at the EuroPar 2017 conference [22].

7.4. Memory-aware tree partitioning on homogeneous platforms

Participants: Anne Benoit, Changjiang Gou, Loris Marchal.

Scientific applications are commonly modeled as the processing of directed acyclic graphs of tasks, and for some of them, the graph takes the special form of a rooted tree. This tree expresses both the computational dependencies between tasks and their storage requirements. The problem of scheduling/traversing such a tree on a single processor to minimize its memory footprint has already been widely studied. Hence, we move to parallel processing and study how to partition the tree for a homogeneous multiprocessor platform, where each processor is equipped with its own memory. We formally state the problem of partitioning the tree into subtrees such that each subtree can be processed on a single processor and the total resulting processing time is minimized. We prove that the problem is NP-complete, and we design polynomial-time heuristics to address it. An extensive set of simulations demonstrates the usefulness of these heuristics.

This work has been accepted as a short paper in the PDP 2018 conference [50].

7.5. Parallel scheduling of DAGs under memory constraints.

Participants: Loris Marchal, Bertrand Simon, Frédéric Vivien.

Scientific workflows are frequently modeled as Directed Acyclic Graphs (DAG) of tasks, which represent computational modules and their dependencies, in the form of data produced by a task and used by another one. This formulation allows the use of runtime sys- tems which dynamically allocate tasks onto the resources of increasingly complex and hetero- geneous computing platforms. However, for some workflows, such a dynamic schedule may run out of memory by exposing too much parallelism. This work focuses on the problem of transforming such a DAG to prevent memory shortage, and concentrates on shared memory platforms. We first propose a simple model of DAG which is expressive enough to emulate com- plex memory behaviors. We then exhibit a polynomial-time algorithm that computes the max- imum peak memory of a DAG, that is, the maximum memory needed by any parallel schedule. We consider the problem of reducing this maximum peak memory to make it smaller than a given bound by adding new fictitious edges, while trying to minimize the critical path of the graph. After proving this problem NP-complete, we provide an ILP solution as well as several heuristic strategies that are thoroughly compared by simulation on synthetic DAGs modeling actual computational workflows. We show that on most instances, we are able to decrease the maximum peak memory at the cost of a small increase in the critical path, thus with little im- pact on quality of the final parallel schedule.

This work has been accepted for presentation at the IPDPS 2018 conference [56].

7.6. On the Complexity of the Block Low-Rank Multifrontal Factorization

Participants: Patrick Amestoy [INP-IRIT], Alfredo Buttari [CNRS-IRIT], Jean-Yves L'Excellent, Théo Mary [UPS-IRIT].

Matrices coming from elliptic Partial Differential Equations have been shown to have a low-rank property: well defined off-diagonal blocks of their Schur complements can be approximated by low-rank products and this property can be efficiently exploited in multifrontal solvers to provide a substantial reduction of their complexity. Among the possible low-rank formats, the Block Low-Rank format (BLR) is easy to use in a general purpose multifrontal solver and has been shown to provide significant gains compared to full-rank on practical applications. However, unlike hierarchical formats, such as $\mathcal{H}$ and HSS, its theoretical complexity was unknown. We extended the theoretical work done on hierarchical matrices in order to compute the theoretical complexity of the BLR multifrontal factorization. We then studied several variants of the BLR multifrontal factorization, depending on the strategies used to perform the updates in the frontal matrices and on the constraints on how numerical pivoting is handled. We showed that these variants can further reduce the complexity of the factorization. In the best case (3D, constant ranks), we obtain a complexity of the order of $O(n^{4/3})$. We provide an experimental study with numerical results to support our complexity bounds.

This work has been published in the SIAM Journal on Scientific Computing [6].

7.7. Large-scale 3-D EM modelling with a Block Low-Rank multifrontal direct solver

Participants: Daniil Shantsev [EMGS-Univ. Oslo], Piyoosh Jaysaval [Univ. Oslo], Sébastien de La Kethulle de Ryhove [EMGS], Patrick Amestoy [INP-IRIT], Alfredo Buttari [CNRS-IRIT], Jean-Yves L'Excellent, Théo Mary [UPS-IRIT].

We put forward the idea of using a Block Low-Rank (BLR) multifrontal direct solver to efficiently solve the linear systems of equations arising from a finite-difference discretization of the frequency-domain Maxwell equations for 3-D electromagnetic (EM) problems. The solver uses a low-rank representation for the off-diagonal blocks of the intermediate dense matrices arising in the multifrontal method to reduce the computational load. A numerical threshold, the so-called BLR threshold, controlling the accuracy of low-rank representations was optimized by balancing errors in the computed EM fields against savings in floating point operations (flops). Simulations were carried out over large-scale 3-D resistivity models representing typical scenarios for marine controlled-source EM surveys, and in particular the SEG SEAM model which contains an irregular salt body. The flop count, size of factor matrices and elapsed run time for matrix factorization are reduced dramatically by using BLR representations and can go down to, respectively, 10, 30 and 40 per cent of their full-rank values for our largest system with N = 20.6 million unknowns. The reductions are almost independent of the number of MPI tasks and threads at least up to $90 \times 10 = 900$ cores. The BLR savings increase for larger systems, which reduces the factorization flop complexity from $O(N^2)$ for the full-rank solver to $O(N^m)$ with $m = 1.4$–1.6. The BLR savings are significantly larger for deep-water environments that exclude the highly resistive air layer from the computational domain. A study in a scenario where simulations are required at multiple source locations shows that the BLR solver can become competitive in comparison to iterative solvers as an engine for 3-D controlled-source electromagnetic Gauss–Newton inversion that requires forward modelling for a few thousand right-hand sides.

This work has been published in the Geophysical Journal International [16].

7.8. On exploiting sparsity of multiple right-hand sides in sparse direct solvers

Participants: Patrick Amestoy [INP-IRIT], Jean-Yves L'Excellent, Gilles Moreau.

The cost of the solution phase in sparse direct methods is sometimes critical. It can be larger than the one of the factorization in applications where systems of linear equations with thousands of right-hand sides (RHS) must be solved. In this work, we focus on the case of multiple *sparse* RHS with different nonzero structures in each column. Given a factorization $A = LU$ of a sparse matrix A and the system $AX = B$ (or $LY = B$ when focusing on the forward elimination), the sparsity of B can be exploited in two ways. First, *vertical* sparsity is exploited by pruning unnecessary nodes from the elimination tree, which represents the dependencies between computations in a direct method. Second, we explain how *horizontal* sparsity can be exploited by working on a subset of RHS columns at each node of the tree. A combinatorial problem must then be solved in order to permute the columns of B and minimize the number of operations. We propose a new algorithm to build such a permutation, based on the tree and on the sparsity structure of B. We then propose an original approach to split the columns of B into a minimal number of blocks (to preserve flexibility in the implementation or maintain high arithmetic intensity, for example), while reducing the number of operations down to a given threshold. Both algorithms are motivated by geometric intuitions and designed using an algebraic approach, and they can be applied to general systems of linear equations. We demonstrate the effectiveness of our algorithms on systems coming from real applications and compare them to other standard approaches. Finally, we give some perspectives and possible applications for this work.

This work is available as a research report [34] and has been submitted to a journal.

7.9. Revisiting temporal failure independence in large scale systems

Participants: Guillaume Aupy [Inria Tadaam], Leonardo Bautista Gomez [Barcelona Supercomputing Center, Spain], Yves Robert, Frédéric Vivien.

This work revisits the *failure temporal independence* hypothesis which is omnipresent in the analysis of resilience methods for HPC. We explain why a previous approach is incorrect, and we propose a new method to detect failure cascades, i.e., series of non-independent consecutive failures. We use this new method to assess whether public archive failure logs contain failure cascades. Then we design and compare several cascade-aware checkpointing algorithms to quantify the maximum gain that could be obtained, and we report extensive simulation results with archive and synthetic failure logs. Altogether, there are a few logs that contain cascades, but we show that the gain that can be achieved from this knowledge is not significant. The conclusion is that we can wrongly, but safely, assume failure independence!

This work is available as a research report and has been submitted to a journal. A preliminary version appears in the proceedings of the FTS'17 workshop.

7.10. Co-scheduling Amdahl applications on cache-partitioned systems

Participants: Guillaume Aupy [Inria Tadaam], Anne Benoit, Sicheng Dai [East China Normal University, China], Loïc Pottier, Padma Raghavan [Vanderbilt University, Nashville TN, USA], Yves Robert, Manu Shantharam [San Diego Supercomputer Center, San Diego CA, USA].

Cache-partitioned architectures allow subsections of the shared last-level cache (LLC) to be exclusively reserved for some applications. This technique dramatically limits interactions between applications that are concurrently executing on a multi-core machine. Consider n applications that execute concurrently, with the objective to minimize the makespan, defined as the maximum completion time of the n applications. Key scheduling questions are: (i) which proportion of cache and (ii) how many processors should be given to each application? In this work, we provide answers to (i) and (ii) for Amdahl applications. Even though the problem is shown to be NP-complete, we give key elements to determine the subset of applications that should share the LLC (while remaining ones only use their smaller private cache). Building upon these results, we design efficient heuristics for Amdahl applications. Extensive simulations demonstrate the usefulness of co-scheduling when our efficient cache partitioning strategies are deployed.

This work is available as a research report and has been accepted for publication in the IJHPCA journal.

7.11. Coping with silent and fail-stop errors at scale by combining replication and checkpointing

Participants: Anne Benoit, Franck Cappello [Argonne National Laboratory, USA], Aurélien Cavelan [University of Basel, Switzerland], Padma Raghavan [Vanderbilt University, Nashville TN, USA], Yves Robert, Hongyang Sun [Vanderbilt University, Nashville TN, USA].

This work provides a model and an analytical study of replication as a technique to detect and correct silent errors, as well as to cope with both silent and fail-stop errors on large-scale platforms. Fail-stop errors are immediately detected, unlike silent errors for which a detection mechanism is required. To detect silent errors, many application-specific techniques are available, either based on algorithms (ABFT), invariant preservation or data analytics, but replication remains the most transparent and least intrusive technique. We explore the right level (duplication, triplication or more) of replication for two frameworks: (i) when the platform is subject only to silent errors, and (ii) when the platform is subject to both silent and fail-stop errors. A higher level of replication is more expensive in terms of resource usage but enables to tolerate more errors and to correct some silent errors, hence there is a trade-off to be found. Replication is combined with checkpointing and comes with two flavors: *process replication* and *group replication*. Process replication applies to message-passing applications with communicating processes. Each process is replicated, and the platform is composed of process pairs, or triplets. Group replication applies to black-box applications, whose parallel execution is replicated several times. The platform is partitioned into two halves (or three thirds). In both scenarios, results are compared before each checkpoint, which is taken only when both results (duplication) or two out of three results (triplication) coincide. If not, one or more silent errors have been detected, and the application rolls back to the last checkpoint, as well as when fail-stop errors have struck. We provide a detailed analytical study for all of these scenarios, with formulas to decide, for each scenario, the optimal parameters as a function

of the error rate, checkpoint cost, and platform size. We also report a set of extensive simulation results that nicely corroborates the analytical model.

This work is available as a research report and has been submitted to a journal. A preliminary version appears in the proceedings of the FTXS'17 workshop.

7.12. Optimal checkpointing period with replicated execution on heterogeneous platforms

Participants: Anne Benoit, Aurélien Cavelan [University of Basel, Switzerland], Valentin Le Fèvre, Yves Robert.

In this work, we design and analyze strategies to replicate the execution of an application on two different platforms subject to failures, using checkpointing on a shared stable storage. We derive the optimal pattern size W for a periodic checkpointing strategy where both platforms concurrently try and execute W units of work before checkpointing. The first platform that completes its pattern takes a checkpoint, and the other platform interrupts its execution to synchronize from that checkpoint. We compare this strategy to a simpler on-failure checkpointing strategy, where a checkpoint is taken by one platform only whenever the other platform encounters a failure. We use first or second-order approximations to compute overheads and optimal pattern sizes, and show through extensive simulations that these models are very accurate. The simulations show the usefulness of a secondary platform to reduce execution time, even when the platforms have relatively different speeds: in average, over a wide range of scenarios, the overhead is reduced by 30%. The simulations also demonstrate that the periodic checkpointing strategy is globally more efficient, unless platform speeds are quite close.

This work is available as a research report. A preliminary version appears in the proceedings of the FTXS'17 workshop.

7.13. A Failure Detector for HPC Platforms

Participants: George Bosilca [ICL, University of Tennessee Knoxville, USA], Aurélien Bouteiller [ICL, University of Tennessee Knoxville, USA], Amina Guermouche [Telecom SudParis, France], Thomas Hérault [ICL, University of Tennessee Knoxville, USA], Yves Robert, Pierre Sens [LIP6, Université Paris 6, France].

Building an infrastructure for exascale applications requires, in addition to many other key components, a stable and efficient failure detector. This work describes the design and evaluation of a robust failure detector, that can maintain and distribute the correct list of alive resources within proven and scalable bounds. The detection and distribution of the fault information follow different overlay topologies that together guarantee minimal disturbance to the applications. A virtual observation ring minimizes the overhead by allowing each node to be observed by another single node, providing an unobtrusive behavior. The propagation stage is using a non uniform variant of a reliable broadcast over a circulant graph overlay network, and guarantees a logarithmic fault propagation. Extensive simulations, together with experiments on the Titan ORNL supercomputer, show that the algorithm performs extremely well and exhibits all the desired properties of an exascale-ready algorithm.

This work is available as a research report and has been accepted for publication in the IJHPCA journal. A preliminary version appears in the proceedings of the SC'16 conference.

7.14. Budget-aware scheduling algorithms for scientific workflows on IaaS cloud platforms

Participants: Yves Caniou [Inria Avalon], Eddy Caron [Inria Avalon], Aurélie Kong Win Chang, Yves Robert.

This work introduces several budget-aware algorithms to deploy scientific workflows on IaaS cloud platforms, where users can request Virtual Machines (VMs) of different types, each with specific cost and speed parameters. We use a realistic application/platform model with stochastic task weights, and VMs communicating through a datacenter. We extend two well-known algorithms, HEFT and MinMin, and make scheduling decisions based upon machine availability *and* available budget. During the mapping process, the budget-aware algorithms make conservative assumptions to avoid exceeding the initial budget; we further improve our results with refined versions that aim at re-scheduling some tasks onto faster VMs, thereby spending any budget fraction leftover by the first allocation. These refined variants are much more time-consuming than the former algorithms, so there is a trade-off to find in terms of scalability. We report an extensive set of simulations with workflows from the Pegasus benchmark suite. Budget-aware algorithms generally succeed in achieving efficient makespans while enforcing the given budget, and despite the uncertainty in task weights.

This work is available as a research report and has been submitted to a journal.

7.15. Resilience for stencil computations with latent errors

Participants: Aurélien Cavelan [University of Basel, Switzerland], Andrew Chien [University of Chicago, USA], Aiman Fang [University of Chicago, USA], Yves Robert.

Projections and measurements of error rates in near-exascale and exascale systems suggest a dramatic growth, due to extreme scale (10^9 cores), concurrency, software complexity, and deep submicron transistor scaling. Such a growth makes resilience a critical concern, and may increase the incidence of errors that "escape", silently corrupting application state. Such errors can often be revealed by application software tests but with long latencies, and thus are known as *latent errors*. We explore how to efficiently recover from latent errors, with an approach called application-based focused recovery (ABFR). Specifically we present a case study of stencil computations, a widely useful computational structure, showing how ABFR focuses recovery effort where needed, using intelligent testing and pruning to reduce recovery effort, and enables recovery effort to be overlapped with application computation. We analyze and characterize the ABFR approach on stencils, creating a performance model parameterized by error rate and detection interval (latency). We compare projections from the model to experimental results with the Chombo stencil application, validating the model and showing that ABFR on stencil can achieve a significant reductions in error recovery cost (up to 400x) and recovery latency (up to 4x). Such reductions enable efficient execution at scale with high latent error rates.

This work is available as a research report . A short version appears in the proceedings of the ICPP'17 conference.

7.16. Checkpointing workflows for fail-stop errors

Participants: Louis-Claude Canon, Henri Casanova [University of Hawai'i at Manoa, USA], Li Han, Yves Robert, Frédéric Vivien.

We consider the problem of orchestrating the execution of workflow applications structured as Directed Acyclic Graphs (DAGs) on parallel computing platforms that are subject to fail-stop failures. The objective is to minimize expected overall execution time, or makespan. A solution to this problem consists of a schedule of the workflow tasks on the available processors and of a decision of which application data to checkpoint to stable storage, so as to mitigate the impact of processor failures. For general DAGs this problem is hopelessly intractable. In fact, given a solution, computing its expected makespan is still a difficult problem. To address this challenge, we consider a restricted class of graphs, Minimal Series-Parallel Graphs (GSPGs). It turns out that many real-world workflow applications are naturally structured as GSPGs. For this class of graphs, we propose a recursive list-scheduling algorithm that exploits the GSPG structure to assign subgraphs to individual processors, and uses dynamic programming to decide which tasks in these sub-gaphs should be checkpointed. Furthermore, it is possible to efficiently compute the expected makespan for the solution produced by this algorithm, using a first-order approximation of task weights and existing evaluation algorithms for 2-state probabilistic DAGs. We assess the performance of our algorithm for production workflow configurations, comparing it to (i) an approach in which all application data is checkpointed, which

corresponds to the standard way in which most production workflows are executed today; and (ii) an approach in which no application data is checkpointed. Our results demonstrate that our algorithm strikes a good compromise between these two approaches, leading to lower checkpointing overhead than the former and to better resilience to failure than the latter. To the best of our knowledge, this is the first scheduling/checkpointing algorithm for workflow applications with fail-stop failures that considers workflow structures more general than mere linear chains of tasks.

This work is available as a research report and has been submitted to a journal. A short version appears in the proceedings of the IEEE Cluster'17 conference.

7.17. Optimal Cooperative Checkpointing for Shared High-Performance Computing Platforms

Participants: Dorian Arnold [Emory University, Atlanta, GA, USA], George Bosilca [ICL, University of Tennessee Knoxville, USA], Aurélien Bouteiller [ICL, University of Tennessee Knoxville, USA], Jack Dongarra [ICL, University of Tennessee Knoxville, USA], Kurt Ferreira [Center for Computing Research, Sandia National Laboratory, USA], Thomas Hérault [ICL, University of Tennessee Knoxville, USA], Yves Robert.

In high-performance computing environments, input/output (I/O) from various sources often contend for scarce available bandwidth. Adding to the I/O operations inherent to the failure-free execution of an application, I/O from checkpoint/restart (CR) operations (used to ensure progress in the presence of failures) places an additional burden as it increases I/O contention, leading to degraded performance. In this work, we consider a cooperative scheduling policy that optimizes the overall performance of concurrently executing CR-based applications which share valuable I/O resources. First, we provide a theoretical model and then derive a set of necessary constraints needed to minimize the global *waste* on the platform. Our results demonstrate that the optimal checkpoint interval as defined by Young/Daly, while providing a sensible metric for a single application, is not sufficient to optimally address resource contention at the platform scale. We therefore show that combining optimal checkpointing periods with I/O scheduling strategies can provide a significant improvement on the overall application performance, thereby maximizing platform throughput. Overall, these results provide critical analysis and direct guidance on checkpointing large-scale workloads in the presence of competing I/O while minimizing the impact on application performance.

This work is available as a research report and has been submitted to a conference.

7.18. Parallel Code Generation of Synchronous Programs for a Many-core Architecture

Participant: Matthieu Moy.

Embedded systems tend to require more and more computational power. Many-core architectures are good candidates since they offer power and are considered more time predictable than classical multi-cores.

Data-flow Synchronous languages such as Lustre or Scade are widely used for avionic critical software. Programs are described by networks of computational nodes. Implementation of such programs on a many-core architecture must ensure a bounded response time and preserve the functional behavior by taking interference into account.

We consider the top-level node of a Lustre application as a software architecture description where each sub-node corresponds to a potential parallel task. Given a mapping (tasks to cores), we automatically generate code suitable for the targeted many-core architecture. This code uses hardware synchronization mechanisms and time-triggered execution. This minimizes memory interferences and allows usage of a framework to compute the Worst-Case Response Time.

This work was accepted for publication at the DATE 2018 conference [82].

7.19. Optimizing Affine Control with Semantic Factorizations

Participants: Christophe Alias, Alexandru Plesco.

Hardware accelerators generated by polyhedral synthesis techniques make an extensive use of affine expressions (affine functions and convex polyhedra) in control and steering logic. Since the control is pipelined, these affine objects must be evaluated at the same time for different values, which forbids aggressive reuse of operators.

In this work, we propose a method to factorize a collection of affine expressions without preventing pipelining. Our key contributions are (i) to use semantic factorizations exploiting arithmetic properties of addition and multiplication and (ii) to rely on a cost function whose minimization ensures a correct usage of FPGA resources. Our algorithm is totally parametrized by the cost function, which can be customized to fit a target FPGA. Experimental results on a large pool of linear algebra kernels show a significant improvement compared to traditional low-level RTL optimizations. In particular, we show how our method reduces resource consumption by revealing hidden strength reductions.

This work has been published in ACM TACO [5]

7.20. Improving Communication Patterns in Polyhedral Process Networks

Participant: Christophe Alias.

Process networks are a natural intermediate representation for HLS and more generally automatic parallelization. Compiler optimizations for parallelism and data locality restructure deeply the execution order of the processes, hence the read/write patterns in communication channels. This breaks most FIFO channels, which have to be implemented with addressable buffers. Expensive hardware is required to enforce synchronizations, which often results in dramatic performance loss.

In this work, we present an algorithm to partition the communications so that most FIFO channels can be recovered after a loop tiling, a key optimization for parallelism and data locality. Experimental results show a drastic improvement of FIFO detection for regular kernels at the cost of (few) additional storage. As a bonus, the storage can even be reduced in some cases.

This work will be presented at the HiP3ES'2018 workshop [32].

7.21. Static Analyses of pointers

Participants: Laure Gonnord, Maroua Maalej.

The design and implementation of static analyses that disambiguate pointers has been a focus of research since the early days of compiler construction. One of the challenges that arise in this context is the analysis of languages that support pointer arithmetics, such as C, C++ and assembly dialects. In 2017, we contributed to this research area with a conference paper and a journal paper.

The CGO'17 paper[27] contributes to solve this challenge. We start from an obvious, yet unexplored, observation: if a pointer is strictly less than another, they cannot alias. Motivated by this remark, we use the abstract interpretation framework to build strict less-than relations between pointers. To this end, we construct a program representation that bestows the Static Single Information (SSI) property onto our dataflow analysis. SSI gives us an efficient sparse algorithm, which, once seen as a form of abstract interpretation, is correct by construction. We have implemented our static analysis in LLVM. It runs in time linear on the number of program variables, and, depending on the benchmark, it can be as much as six times more precise than the pointer disambiguation techniques already in place in that compiler.

Pentagons is an abstract domain invented by Logozzo and Fahndrich to validate array accesses in low-level programming languages. This algebraic structure provides a cheap "less-than check", which builds a partial order between the integer variables used in a program. In the Science of Computer Programming journal paper[15], we show how we have used the ideas available in Pentagons to design and implement a novel alias analysis. With this new algorithm, we are able to disambiguate pointers with off- sets, that commonly occur in C programs, in a precise and efficient way. Together with this new abstract domain we describe several implementation decisions that let us produce a practical pointer disambiguation algorithm on top of the LLVM compiler. Our alias analysis is able to handle programs as large as SPEC CPU2006's gcc in a few minutes. Furthermore, it improves on LLVM's industrial quality analyses. As an extreme example, we have observed a 4x improvement when analyzing SPEC's lbm.

7.22. Dataflow static analyses and optimisations

Participants: Laure Gonnord, Lionel Morel, Szabolcs-Marton Bagoly, Romain Fontaine.

Nowadays, parallel computers have become ubiquitous and current processors contain several execution cores, anywhere from a couple to hundreds. This multi-core tendency is due to constraints preventing the increase of clock frequencies, such as heat generation and power consumption. A variety of low-level tools exist to program these chips efficiently, but they are considered hard to program, to maintain, and to debug, because they may exhibit non-deterministic behaviors. We explore the potentiality of the data flow programming, which allows the programmer to specify only the operations to perform and their dependencies, without actually scheduling them. The work is published in two research reports: [48] and [49].

In [48], we explore the combination of a dataflow paradigm language, SigmaC, with the Polyhedral Model, which allows automatic parallelization and optimization of loop nests, in order to make the programming easier by delegating work to the compilers and static analyzers, in various case studies.

In [49], we explore the expressivity of the horn clause format for static analyses of Lustre programs with arrays. We propose a translation from a Lustre core language to horn clauses, with or without array variables.

STORM Project-Team

6. New Results

6.1. Distributed Sequential Task Flow with StarPU

The emergence of accelerators as standard computing resources on supercomputers and the subsequent architectural complexity increase revived the need for high-level parallel programming paradigms. Sequential task-based programming model has been shown to efficiently meet this challenge on a single multicore node possibly enhanced with accelerators, which motivated its support in the OpenMP 4.0 standard. We showed [5] that this paradigm can also be employed to achieve high performance on modern supercomputers composed of multiple such nodes, with extremely limited changes in the user code. To prove this claim, we have extended the StarPU runtime system with an advanced inter-node data management layer that supports this model by posting communications automatically. We illustrates our discussion with the task-based tile Cholesky algorithm that we implemented on top of this new runtime system layer. We showed that it enables very high productivity while achieving a performance competitive with both the pure Message Passing Interface (MPI)-based ScaLAPACK Cholesky reference implementation and the DPLASMA Cholesky code, which implements another (non-sequential) task-based programming paradigm.

6.2. Distributed StarPU on top of a High Performance Communication Library

A new implementation of the StarPU's distributed engine is being currently developed on top of the NewMadeleine library. The first version of this engine had been written directly on top of MPI. The performance were not as good as expected when dealing with applications exchanging huge number of messages, and we had to implement within StarPU mechanisms to control the memory subscription [14].

NewMadeleine is a high performance communication library for clusters developed in the Tadaam team. It applies optimization strategy on data flows through dynamic packet scheduling, and is usable on various high performance networks. The new implementation of the StarPU's distributed engine no longer has to deal with communication-related issues, and provides a better reactivity as the communications progress is dealt with by NewMadeleine itself. First experiments with the Chameleon solver show promising results.

6.3. Bridging the Gap between a Standard Parallel Language and a Task-based Runtime System

With the advent of complex modern architectures, the low-level paradigms long considered sufficient to build High Performance Computing (HPC) numerical codes have met their limits. Achieving efficiency, ensuring portability, while preserving programming tractability on such hardware prompted the HPC community to design new, higher level paradigms while relying on runtime systems to maintain performance. However, the common weakness of these projects is to deeply tie applications to specific expert-only runtime system APIs. The OpenMP specification, which aims at providing common parallel programming means for shared-memory platforms, appears as a good candidate to address this issue thanks to the latest task-based constructs introduced in its revision 4.0. We assessed [4] the effectiveness and limits of this support for designing a high-performance numerical library, ScalFMM, implementing the fast multipole method (FMM) that we have deeply re-designed with respect to the most advanced features provided by OpenMP 4. We showed that OpenMP 4 allows for significant performance improvements over previous OpenMP revisions on recent multicore processors and that extensions to the 4.0 standard allow for strongly improving the performance, bridging the gap with the very high performance that was so far reserved to expert-only runtime system APIs. Our proposal for an OpenMP extension to let the programmer express the property of commutativity between multiple tasks has been presented by Inria and successfully voted-on and integrated as the notion of mutually exclusive input/output sets (mutexinoutset keyword) in OpenMP ARB's Technical Report 6: OpenMP Version 5.0 Preview 2, the last pre-version of the upcoming OpenMP 5.0 specification.

6.4. Combining a Component Model and a Task Parallelism Model

We demonstrated the feasibility of efficiently combining both a software component model and a task-based model [6]. Task based models are known to enable efficient executions on recent HPC computing nodes while component models ease the separation of concerns of application and thus improve their modularity and adaptability.

This paper describes a prototype version of the COMET programming model combining concepts of task-based and component models, and a preliminary version of the COMET runtime built on top of StarPU and L2C. Evaluations of the approach have been conducted on a real-world use-case analysis of a sub-part of the production application GYSELA.

Results show that the approach is feasible and that it enables easy composition of independent software codes without introducing overheads. Performance results are equivalent to those obtained with a plain OpenMP based implementation.

6.5. Tackling the granularity problem

One of the main issues encountered when trying to exploit both CPUs and accelerators is that these devices have very different characteristics and requirements. Indeed, GPUs typically exhibit better performance when executing kernels applied to large data sets while regular CPU cores reach their peak performance with fine grain kernels working on a reduced memory footprint. To work around this granularity problem, task-based applications running on such heterogeneous platforms typically adopt a medium granularity, chosen as a trade-off between coarse-grain and fine-grain kernels. To tackle this granularity problem, we investigated different complementary technics. The first two technics are based on StarPU, performing both load-balancing and scheduling, the third one splits automatically kernels at compile-time and then performs load-balancing.

- The first technic is based on resource aggregation : we aggregate CPU cores to execute coarse grain tasks in a parallel manner. We have showed that this technic for a dense Cholesky factorization kernel outperforms state of the art implementations on a platform equipped with 24 CPU cores and 4 GPU devices (reaching a peak performance of 4.8 TFlop/s) and on the Intel KNL processor (reaching a peak performance 1.58 TFlop/s).
- The second technic splits dynamically coarse grain tasks when they are assigned to CPU cores. Tasks can be replaced by a subgraph of tasks of finer granularity, allowing for a finer handling of depencencies and a better pipelining of kernels. This mechanism allowing to deal with hierarchical task graphs has been designed within StarPU. Moreover, it allows to parallelize the task submission flow while preserving the simplicity of the sequential task flow submission paradigm. First experimental results for dense Cholesky factorization kernel show good performance improvements with respect to the native StarPU's implementation.
- The third technic extends our previous work that provides an automatic compiler and runtime technique to execute single OpenCL kernels on heterogeneous multi-device architectures. Our technique splits computation and data automatically across the multiple computing devices. OpenCL applications that consist in a chain of data-dependent kernels in an iterative computation are now considered.

  The technique proposed is completely transparent to the user, and does not require off-line training or a performance model. It manages sometimes conflicting needs between load balancing each kernel in the chain and minimizing data transfer between consecutive kernels, taking data locality into account. Load-balancing issues, resulting from hardware heterogeneity, load imbalance within a kernel itself, and load variations between repeated executions are also managed.

  Experiments on some benchmarks show the interest of our approach and we are currently implementing it in an OpenCL N-body computation with short-range interactions.

6.6. Interfacing MAQAO and BOAST Frameworks for Kernel Autotuning on ARM Platforms

In Project MontBlanc 2's deliverable D5.11 [13] we presented the integration of STORM's MAQAO software (a binary-level code analysis framework) with BOAST (an automatic performance tuning framework for meta-programming and optimizing computing kernels) developed at LIG's NANOSIM. From source meta-kernels written in the RUBY language, BOAST generates multiple versions, in various target languages, optionally applying optimization transformations and strategies, and exploring the space of compiler flags combinations, to discover the most effective kernel tuning parameters. MAQAO offers a scriptable framework to disassemble kernel binaries, explore binary instruction flows, register-level data dependencies, program control structures, to patch, re-assemble and instrument kernel binaries for tracing data access patterns, and to process them from custom analyzers written in the LUA language. This integration work built on the complementarity of these two environments by enabling MAQAO to process binary kernels generated by BOAST, and lead developers in a guided tuning cycle.

6.7. Using heterogeneous memories

Heterogeneous memories, such as the MCDRAM in the Xeon Phi architecture, with different latency and bandwidth characteristics, complexify the way the users allocate and use memory. In 2017, we have designed, in collaboration with CEA, an automatic tool to characterize the bandwidth needs of an application, in particular finding the functions and the arrays in these functions that would benefit the most of a high bandwidth. This tool is a plugin of gcc, and has been applied successfully to large CORAL benchmarks (Lulesh, MiniFE, AMG2013, Mcb and Snap). This characterization is essential in the common case where all data cannot fit into the MCDRAM but a more selective use of the MCDRAM is needed. The transformation of the memory allocations is automatic, based on these metrics. The development of better metrics, allowing to choose the most appropriate array is on going work.

6.8. Rewriting System for Profile-Guided Data Layout Transformations on Binaries

Careful data layout design is crucial for achieving high performance. However exploring data layouts is time-consuming and error-prone, and assessing the impact of a layout transformation on performance is difficult without performing it. We proposed [7] a method and implemented a prototype to guide application programmers through data layout restructuring for improving kernel performance and SIMDizability, by providing a comprehensive multidimensional description of the initial layout, built from trace analysis, and then by giving a performance evaluation of the transformations tested and an expression of each transformed layout. The programmer can limit the exploration to layouts matching some patterns. We apply this method to two multithreaded applications. The performance prediction of multiple transformations matches within 5% the performance of hand-transformed layout code.

6.9. Correctness of HPC Applications

The current supercomputer hardware trends lead to more complex HPC applications (heterogeneity in hardware and combinations of parallel programming models) that pose programmability challenges. Furthermore, progress to exascale stresses the requirement for convenient and scalable debugging methods to help developers fully exploit the future machines. Despite advances in the domain, this still remains a manual complex task. We aim to develop tools and methods to aid developers with problems of correctness in HPC applications for exascale systems. There are several requirements for such tools: 1) precision - report and handle only real problems, areas of interest; 2) scalability in LoCs and execution time; 3) heterogeneity - ability to handle multiple languages, runtime and execution models; and 4) soundness - ability to prove code properties. In order to improve developer productivity, we aim to develop a combination of static and dynamic analyses. Static analysis techniques will enable soundness and scalability in execution time. Dynamic analysis techniques will enable precision, scalability in LoCs and heterogeneity for hybrid parallelism.

The achieved results this year allow to perform an interprocedural static data- and control-flow analysis: its improves precision, by only detecting possible correctness issues related to MPI rank dependent variables. It improves scalability also by reducing the amount of dead-lock avoiding code added. This new method has been applied to CUDA, MPI, OpenMP and UPC parallel codes to detect collective deadlocks.

6.10. AMR-Based Dynamic Load Balancing for Molecular Dynamics Simulations

Modern parallel architectures require applications to generate enough parallelism to feed many cores, which require in turn regular data-parallel instructions to exploit large vector units. We revisit the extensively-studied Classical Molecular Dynamics N-body problem in the light of these hardware constraints. A new data layout is proposed with efficient force computation methods focusing on adaptive mesh refinement techniques, multi-threading, vectorization-friendly, using low memory footprint. Our design is guided by the need for load balancing and adaptivity raised by highly dynamic particle sets, as typically observed in simulations of strong shocks resulting in material micro-jetting. We analyze performance results on several simulation scenarios, over clusters equipped with Intel Xeon Phi knl processors. Performance obtained with our implementation using OpenMP is close to state-of-the-art implementations (LAMMPS) using MPI on steady particles simulations, and outperform them by 1.2 on micro-jetting simulations on Intel Xeon Phi (KNL).

6.11. Resource-Management Study in HPC Runtime-Stacking Context

With the advent of multicore and manycore processors as building blocks of HPC supercomputers, many applications shift from relying solely on a distributed programming model (e.g., MPI) to mixing distributed and shared-memory models (e.g., MPI+OpenMP), to better exploit shared-memory communications and reduce the overall memory footprint. One side effect of this programming approach is runtime stacking: mixing multiple models involve various runtime libraries to be alive at the same time and to share the underlying computing resources. This paper explores different configurations where this stacking may appear and introduces algorithms to detect the misuse of compute resources when running a hybrid parallel application. We have implemented our algorithms inside a dynamic tool that monitors applications and outputs resource usage to the user. We validated this tool on applications from CORAL benchmarks. This leads to relevant information which can be used to improve runtime placement, and to an average overhead lower than 1% of total execution time.

TADaaM Project-Team

7. New Results

7.1. Network Modeling

NETLOC (see Section 6.3) is a tool in HWLOC to discover the network topology. The information gathered and analysed are now saved in XML format. It brings more flexibiliy, readability and compatibility. Henceforth, in the display tool, we compute the positions of the nodes rather than use physics algorithm provided by vis.js library for node placement. Thus, it makes the visualization faster and we can display a fat-tree with around 41k nodes in less than 1 second.

Moreover, we can deal with other kinds of topologies. We handle topologies in a generic way and can have nested topologies. For the mapping, we build a deco graph in SCOTCH. Consequently, the mapping will be possible for any architecture. [17]

We have also optimized the mapping by giving a preconditioned matrix to SCOTCH, and by computing some metrics in order to evaluate mappings and keep the best one.

The part about discovering network have been improved and we support now, in addition to Infiniband, Omnipath fat-trees, Cray Torus.

7.2. Locality Aware Roofline Model

The trend of increasing the number of cores on-chip is enlarging the gap between compute power and memory performance. This issue leads to design systems with heterogeneous memories, creating new challenges for data locality. Before the release of those memory architectures, the Cache-Aware Roofline Model [47] (CARM) offered an insightful model and methodology to improve application performance with knowledge of the cache memory subsystem.

With the help of hwloc library, we are able to leverage the machine topology to extend the CARM for modeling NUMA and heterogeneous memory systems, by evaluating the memory bandwidths between all combinations of cores and NUMA nodes. The new Locality Aware Roofline Model [19] (LARM) scopes most contemporary types of large compute nodes and characterizes three bottlenecks typical of those systems, namely contention, congestion and remote access.

This work has been achieved in collaboration with the authors of the CARM and the source code of the associated tool is publicly available at https://github.com/NicolasDenoyelle/Locality-Aware-Roofline-Model.

In the future we plan to design and embed in the model an hybrid memory bandwidth model to provide an automatic roof matching feature.

7.3. Scalable Management of Platform Topologies

HWLOC (see Section 6.2) is used for gathering the topology of computing nodes. Those nodes are now growing to hundreds of cores, making the overall amount of topology information non-negligible. We studied the overhead of topology discovery on the overall execution time and showed that the Linux kernel is bottleneck on large nodes. It raised the need to use exported and/or abstracted topologies to factorize this overhead [22].

The memory footprint of locality information is also becoming an issue on large many-core. We designed a way to share this information between processes inside nodes so as to factorize this memory consumption [45].

7.4. New algorithm for I/O scheduling

We started working on I/O scheduling for HPC applications. HPC applications can be caracterized by I/O patterns that are repeated periodically. We showed in a simple context how this information can be taken into account to outperform state of the art I/O schedulers [15].

These preliminary results led to the obtention of the ANR DASH (see Section 9.1.2).

After which, we have performed a theoretical analysis to show how one should size the burst-buffers and the bandwidth to those buffers on a HPC system depending on the applications running. In our study we focused on one role of the buffers (namely the role of buffer to the PFS) [42]. This study is particularly important since those buffers are limited and can be used for many usage. Over or under booking the buffers for a specific use leads to an increase of congestion.

7.5. Topology-Aware Data Aggregation on Large-Scale Supercomputers

We have continue our work on on two-phase i/O and data aggregation. This strategy consists of selecting a subset of processes to aggregate contiguous pieces of data before performing reads/writes. In collaboration with Argonne National Lab, we have worked on TAPIOCA, an MPI-based library implementing an efficient topology-aware two-phase I/O algorithm. TAPIOCA can take advantage of double-buffering and one-sided communication to reduce as much as possible the idle time during data aggregation. We also introduce our cost model leading to a topology-aware aggregator placement optimizing the movements of data. We validate our approach at large scale on two leadership-class supercomputers: Mira (IBM BG/Q) and Theta (Cray XC40). On BG/Q+GPFS, for instance, our algorithm leads to a performance improvement by a factor of twelve while on the Cray XC40 system associated with a Lustre filesystem, we achieve an improvement of four [27]

7.6. Empirical Study of the Impact on Performance of Process Affinity and Metrics

Process placement, also called topology mapping, is a well-known strategy to improve parallel program execution by reducing the communication cost between processes. It requires two inputs: the topology of the target machine and a measure of the affinity between processes. In the literature, the dominant affinity measure is the communication matrix that describes the amount of communication between processes. We have studied the accuracy of the communication matrix as a measure of affinity. We have done an extensive set of tests with two fat-tree machines and a 3d-torus machine to evaluate several hypotheses that are often made in the literature and to discuss their validity. First, we check the correlation between algorithmic metrics and the performance of the application. Then, we check whether a good generic process placement algorithm never degrades performance. And finally, we see whether the structure of the communication matrix can be used to predict gain.

7.7. Automatic, Abstracted and Portable Topology-Aware Thread Placement

Efficiently programming shared-memory machines is a difficult challenge because mapping application threads onto the memory hierarchy has a strong impact on the performance. However, optimizing such thread placement is difficult: architectures become increasingly complex and application behavior changes with implementations and input parameters, e.g problem size and number of threads.We have worked on a fully automatic, abstracted and portable affinity module. It produces and implements an optimized affinity strategy that combines knowledge about application characteristics and the platform topology. Implemented in the back-end of our runtime system (ORWL), our approach was used to enhance the performance and the scalability of several unmodified ORWL-coded applications [23]

7.8. Process Placement with TreeMatch

We released TREEMATCH version 1.0 in June. The new feature are: a stabilize API, optional integration of SCOTCH, extensive testing of all the features.

7.9. Managing StarPU Communications with NewMadeleine

We have worked on the scalability with the number of communication requests in the NewMadeleine 6.4 communication library, so as to be able to manage communication patterns from the StarPU runtime. We have ported [44] StarPU on top of NewMadeleine so as to take benefit from NewMadeleine scalability in StarPU. Preliminary results are encouraging.

7.10. New abstraction to manage hardware topologies in MPI applications

Since the end of year 2016, we have been working on new abstractions and mechanisms that can allow the programmer to take advantage of the underlying hardware topology in their parallel applications developed in MPI. For instance, taking into account the intricate network/memory hierarchy can lead to substantial improvements in communication performance and reduce altogether the overall execution time of the application. However, it is important to find the relevant level of abstraction, as too much details are not usable practically because the programmer is not a a hardware specialist most of the time. Also, MPI being hardware-agnostic, it is important to find means to use the hardware specifics without being tied to a particular architecture or hardware design.

With these goals in mind, we proposed the HSPLIT (see Section 6.1) libray that implements a solution based on a well-known MPI concept, the *communicators* (that can be seen as groups of communicating processes). With HSPLIT, each level in the hardware hierarchy is accessible through a dedicated communicator. In this way, the programmer can leverage the underlying hierarchy in their application quite simply . The current implementation of HSPLIT is based on both HWLOC and NETLOC.

This work led to the creation of a new active working group within the MPI Forum, coordinated and lead by Inria.

7.11. Empirical Study of the Impact on Performance of Process Affinity and Metrics

Process placement, also called topology mapping, is a well-known strategy to improve parallel program execution by reducing the communication cost between processes. It requires two inputs: the topology of the target machine and a measure of the affinity between processes. In the literature, the dominant affinity measure is the communication matrix that describes the amount of communication between processes. We have studied the accuracy of the communication matrix as a measure of affinity. We have done an extensive set of tests with two fat-tree machines and a 3d-torus machine to evaluate several hypotheses that are often made in the literature and to discuss their validity. First, we check the correlation between algorithmic metrics and the performance of the application. Then, we check whether a good generic process placement algorithm never degrades performance. And finally, we see whether the structure of the communication matrix can be used to predict gain [35].

7.12. Gradient reconstruction in a legacy CFD application using task-based programming models

We investigated different runtime systems, namely StarPU and PaRSEC and their use in a legacy CFD code from EDF R&D. We assessed both runtimes in terms of performance, ease of implementation and various others criterion such as maintainability, documentation and team activity. By experimenting these solutions out of classical linear algebra problems, we push them out of their comfort zone into the common issues seen in Computational Fluid Dynamics codes with unstructured meshes [30].

7.13. Efficient multi-constraint graph partitioning algorithms

Although several tools provide multi-constraint graph partitioning features, this problem had not been thoroughly investigated. In the context of the PhD of Rémi Barat, several significant results were achieved regarding the multi-constraint graph partitioning problem.

Firstly, a theoretical analysis of the solution space of the mono-criterion, balanced graph bipartitioning problem showed that this space is strongly connected. Hence, local optimization algorithms may indeed succeed in finding paths to better solutions, from some existing solution. A conjecture on the multi-criteria case has been derived. These findings reversed our view on partitioning: while most tools try to find a possibly unbalanced partition of small cut, and then try to rebalance it, it is in fact possible to compute a balanced partition of arbitrary cut, and then to improve the cut.

Secondly, a thorough investigation of the multilevel framework, and of its implementations in several existing tools, allowed us to define the characteristics of an effective coarsening method, both in the mono-criterion and multi-criteria case. Also, new multi-criteria graph algorithms were designed for the initial partitioning and local optimization phases of the multilevel framework [43]. A new data structure has been devised, which speeds-up the computation of balanced partitions in the multi-criteria case.

Thirdly, all of the aforementioned algorithms were implemented in a prototype version of SCOTCH.

7.14. Progress threads placement for MPI Non-Blocking Collectives

MPI Non-Blocking Collectives (NBC) allow for communication overlap with computation. A good overlapping ratio is obtained when computation and communication are running in parallel. To achieve this, each MPI task generates a progress thread to manage communication tasks. The progression of these communications requires regular access to the processors. These threads compete with each other and with MPI tasks. In order to run threads with minimal disruption, we bound the progress threads on free cores when it is possible. Then, we showed that folding all progress threads on very few cores does not work for tree algorithms. The number of communication generated are too important. The solution that we propose is to perform a number of levels (S) of the dependency tree on MPI tasks. We get a reasonable execution time (less than compute time + communication time) while reserving fewer cores for progress threads. All these methods have been implemented in the MPC framework, which contributes to its development.

7.15. Use of PaMPA on large-scale simulations

Many improvements have been brought to PaMPA this year, to improve its robustness and scalability, and to extend its features. In the context of a joint work with CERFACS, PaMPA was subsequently used to remesh the mesh of a helicopter turbine combustion chamber, up to 1 billion elements. This allowed to run a Large-Eddy Simulation (LES) simulation that was out of reach of previous state-of-the-art remeshing software [38].

7.16. Co-scheduling applications on cache-partitioned systems

Cache-partitioned architectures allow subsections of the shared last-level cache (LLC) to be exclusively reserved for some applications. This technique dramatically limits interactions between applications that are concurrently executing on a multi- core machine. We have provided efficient algorithms to co-schedule multiple applications on cache-partitioned systems and evaluations showing that they performed well [13], [6]. We are currently in the process of evaluating them on real machines.

7.17. Dynamic memory-aware task-tree scheduling

We have provided new efficient algorithms that can be used for sparse matrices factorizations under memory constraints. We provide speedup of 15 to 45% over existing strategies and we are working on an actual implementation in QR-MUMPS [14].

ASCOLA Project-Team

7. New Results

7.1. Cloud programming and management

7.1.1. Cloud infrastructures

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

7.1.1.1. Geo-distributed Clouds

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

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

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

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

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

7.1.1.2. Cloud and HPC convergence

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

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

7.1.1.3. Virtualization simulation

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

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

7.1.2. Deployment and reconfiguration in the Cloud

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

7.1.2.1. Initial deployment and placement strategies

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

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

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

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

7.1.2.2. Capacity planning and scheduling

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

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

7.1.2.3. Reconfiguration and self-management

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

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

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

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

7.2. Energy-aware computing

7.2.1. Renewable energy

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

In [9], we present a scheme for green energy management in the presence of explicit and implicit integration of renewable energy in data center. More specifically we propose three contributions: i) we introduce the concept of *virtualization of green energy* to address the uncertainty of green energy availability, ii) we extend the Cloud Service Level Agreement (CSLA) language [0] to support Green SLA by introducing two new threshold parameters and iii) we introduce green SLA algorithm which leverages the concept of virtualization of green energy to provide per interval specific Green SLA. Experiments were conducted with real workload profile from PlanetLab and server power model from SPECpower to demonstrate that Green SLA can be successfully established and satisfied without incurring higher cost.

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

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

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

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

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

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

7.2.2. Energy-aware consolidation and reconfiguration

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

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

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

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

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

7.3. Software engineering

7.3.1. Security and privacy

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

7.3.1.1. System-level security for virtualized environments

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

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

7.3.1.2. Privacy-Aware Data Storage.

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

7.3.1.3. Accountability for Cloud applications.

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

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

7.3.2. Software development and programming languages

7.3.2.1. Industrial Internet

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

7.3.2.2. Cloud and HPC programming

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

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

DIVERSE Project-Team

7. New Results

7.1. Results on Variability modeling and management

7.1.1. Variability and testing.

Many approaches for testing configurable software systems start from the same assumption: it is impossible to test all configurations. This motivated the definition of variability-aware abstractions and sampling techniques to cope with large configuration spaces. Yet, there is no theoretical barrier that prevents the exhaustive testing of all configurations by simply enumerating them, if the effort required to do so remains acceptable. Not only this: we believe there is lots to be learned by systematically and exhaustively testing a configurable system. We report on the first ever endeavor to test all possible configurations of an industry-strength, open source configurable software system, JHipster, a popular code generator for web applications. We built a testing scaffold for the 26,000+ configurations of JHipster using a cluster of 80 machines during 4 nights for a total of 4,376 hours (182 days) CPU time. We find that 35.70% configurations fail and we identify the feature interactions that cause the errors. We show that sampling testing strategies (like dissimilarity and 2-wise) (1) are more effective to find faults than the 12 default configurations used in the JHipster continuous integration; (2) can be too costly and exceed the available testing budget. We cross this quantitative analysis with the qualitative assessment of JHipster's lead developers. Additional resources: preliminary effort on JHipster [32], https://arxiv.org/abs/1710.07980https://github.com/axel-halin/Thesis-JHipster/

7.1.2. Variability and teaching.

Software Product Line (SPL) engineering has emerged to provide the means to efficiently model, produce, and maintain multiple similar software variants, exploiting their common properties, and managing their variabilities (differences). With over two decades of existence, the community of SPL researchers and practitioners is thriving as can be attested by the extensive research output and the numerous successful industrial projects. Education has a key role to support the next generation of practitioners to build highly complex, variability-intensive systems. Yet, it is unclear how the concepts of variability and SPLs are taught, what are the possible missing gaps and difficulties faced, what are the benefits, or what is the material available. Also, it remains unclear whether scholars teach what is actually needed by industry. We report on three initiatives we have conducted with scholars, educators, industry practitioners, and students to further understand the connection between SPLs and education, i.e., an online survey on teaching SPLs we performed with 35 scholars, another survey on learning SPLs we conducted with 25 students, as well as two workshops held at the International Software Product Line Conference in 2014 and 2015 with both researchers and industry practitioners participating. We build upon the two surveys and the workshops to derive recommendations for educators to continue improving the state of practice of teaching SPLs, aimed at both individual educators as well as the wider community. Finally, we are developing and maintaining a repository for teaching SPLs and variability. Additional resources: https://teaching.variability.io

7.1.3. Variability and constraint solving.

Array constraints are essential for handling data structures in automated reasoning and software verification. Unfortunately, the use of a typical finite domain (FD) solver based on local consistency-based filtering has strong limitations when constraints on indexes are combined with constraints on array elements and size. This work proposes an efficient and complete FD-solving technique for extended constraints over (possibly unbounded) arrays. We describe a simple but particularly powerful transformation for building an equisatisfiable formula that can be efficiently solved using standard FD reasoning over arrays, even in the unbounded case. Experiments show that the proposed solver significantly outperforms FD solvers, and successfully competes with the best SMT-solvers [38]. This work is not directly related to variability and SPL. But it contributes to DiverSE's attempts to connect artificial intelligence techniques to software variability engineering, in which constraint solving or machine learning are typically applied.

7.1.4. Variability and machine learning (performance specialization of variability-intensive systems).

We propose the use of a machine learning approach to infer variability constraints from an oracle that is able to assess whether a given configuration is correct. We propose an automated procedure to randomly generate configurations, classify them according to the oracle, and synthesize cross-tree constraints. Specifically, based on an oracle (e.g. a runtime test) that tells us whether a given configuration meets the requirements (e.g. speed or memory footprint), we leverage machine learning to retrofit the acquired knowledge into a variability model of the system that can be used to automatically specialize the configurable system. We validate our approach on a set of well-known configurable software systems (Apache server, x264, etc.) Our results show that, for many different kinds of objectives and performance qualities, the approach has interesting accuracy, precision and recall after a learning stage based on a relative small number of random samples [43]. Additional resources: https://learningconstraints.github.io and VaryVary ANR project

7.1.5. Variability and machine learning (learning contextual variability models).

Modeling how contextual factors relate to a software system's configuration space is usually a manual, error-prone task that depends highly on expert knowledge. Machine-learning techniques can automatically predict the acceptable software configurations for a given context. Such an approach executes and observes a sample of software configurations within a sample of contexts. It then learns what factors of each context will likely discard or activate some of the software's features. This lets developers and product managers automatically extract the rules that specialize highly configurable systems for specific contexts [27] Additional resources: https://learningconstraints.github.io and VaryVary ANR project

We are currently exploring the use of machine learning for variability-intensive systems in the context of VaryVary ANR project (see also VaryLaTeX [28]).

7.2. Results on Software Language Engineering

7.2.1. On Language Interfaces

Complex systems are developed by teams of experts from multiple domains , who can be liberated from becoming programming experts through domain-specific languages (DSLs). The implementation of the different concerns of DSLs (including syntaxes and semantics) is now well-established and supported by various languages workbenches. However, the various services associated to a DSL (e.g., editors, model checker, debugger or composition operators) are still directly based on its implementation. Moreover, while most of the services crosscut the different DSL concerns, they only require specific information on each. Consequently, this prevents the reuse of services among related DSLs, and increases the complexity of service implementation. Leveraging the time-honored concept of interface in software engineering, we discuss in [40] the benefits of language interfaces in the context of software language engineering. In particular, we elaborate on particular usages that address current challenges in language development.

7.2.2. Revisiting Visitors for Modular Extension of Executable DSMLs

Executable Domain-Specific Modeling Languages (xDSMLs) are typically defined by metamodels that specify their abstract syntax, and model interpreters or compilers that define their execution semantics. To face the proliferation of xDSMLs in many domains, it is important to provide language engineering facilities for opportunistic reuse, extension, and customization of existing xDSMLs to ease the definition of new ones. Current approaches to language reuse either require to anticipate reuse, make use of advanced features that are not widely available in programming languages, or are not directly applicable to metamodel-based xDSMLs. In [35], we propose a new language implementation pattern, named REVISITOR, that enables independent extensibility of the syntax and semantics of metamodel-based xDSMLs with incremental compilation and without anticipation. We seamlessly implement our approach alongside the compilation chain of the Eclipse Modeling Framework, thereby demonstrating that it is directly and broadly applicable in various modeling environments. We show how it can be employed to incrementally extend both the syntax and semantics of the fUML language without requiring anticipation or re-compilation of existing code, and with acceptable performance penalty compared to classical handmade visitors.

7.2.3. Advanced and efficient execution trace management for executable domain-specific modeling languages

Executable Domain-Specific Modeling Languages (xDSMLs) enable the application of early dynamic verification and validation (V&V) techniques for behavioral models. At the core of such techniques, execution traces are used to represent the evolution of models during their execution. In order to construct execution traces for any xDSML, generic trace metamodels can be used. Yet, regarding trace manipulations, generic trace metamodels lack efficiency in time because of their sequential structure, efficiency in memory because they capture superfluous data, and usability because of their conceptual gap with the considered xDSML. We contributed in [22] a novel generative approach that defines a multidimensional and domain-specific trace metamodel enabling the construction and manipulation of execution traces for models conforming to a given xDSML. Efficiency in time is improved by providing a variety of navigation paths within traces, while usability and memory are improved by narrowing the scope of trace metamodels to fit the considered xDSML. We evaluated our approach by generating a trace metamodel for fUML and using it for semantic differencing, which is an important V&V technique in the realm of model evolution. Results show a significant performance improvement and simplification of the semantic differencing rules as compared to the usage of a generic trace metamodel.

7.2.4. Omniscient Debugging for Executable DSLs

Omniscient debugging is a promising technique that relies on execution traces to enable free traversal of the states reached by a model (or program) during an execution. While a few General-Purpose Languages (GPLs) already have support for omniscient debugging, developing such a complex tool for any executable Domain Specific Language (DSL) remains a challenging and error prone task. A generic solution must: support a wide range of executable DSLs independently of the metaprogramming approaches used for implementing their semantics; be efficient for good responsiveness. Our contribution in [21] relies on a generic omniscient debugger supported by efficient generic trace management facilities. To support a wide range of executable DSLs, the debugger provides a common set of debugging facilities, and is based on a pattern to define runtime services independently of metaprogramming approaches. Results show that our debugger can be used with various executable DSLs implemented with different metaprogramming approaches. As compared to a solution that copies the model at each step, it is on average six times more efficient in memory, and at least 2.2 faster when exploring past execution states, while only slowing down the execution 1.6 times on average.

7.2.5. Reverse Engineering Language Product Lines from Existing DSL Variants

The use of domain-specific languages (DSLs) has become a successful technique in the development of complex systems. Nevertheless, the construction of this type of languages is time-consuming and requires highly-specialized knowledge and skills. An emerging practice to facilitate this task is to enable reuse through the definition of language modules which can be later put together to build up new DSLs. In [26], we propose a reverse-engineering technique to ease-off such a development scenario. Our approach receives a set of DSL variants which are used to automatically recover a language modular design and to synthesize the corresponding variability models. The validation is performed in a project involving industrial partners that required three different variants of a DSL for finite state machines. This validation shows that our approach is able to correctly identify commonalities and variability.

7.2.6. Software Language Engineering for Virtual Reality Software Development

Due to the nature of Virtual Reality (VR) research, conducting experiments in order to validate the researcher's hypotheses is a must. However, the development of such experiments is a tedious and time-consuming task. In [48], we propose to make this task easier, more intuitive and faster with a method able to describe and generate the most tedious components of VR experiments. The main objective is to let experiment designers focus on their core tasks: designing , conducting, and reporting experiments. To that end, we applied well-established SLE concepts promoted in DIVERSE to the VR domain to ease the development of VR experiments. More precisely, we propose the use of DSLs to ease the description and generation of VR experiments. An analysis of published VR experiments is used to identify the main properties that characterize VR experiments.

This allowed us to design AGENT (Automatic Generation of ExperimeNtal proTocol runtime), a DSL for specifying and generating experimental protocol runtimes. We demonstrated the feasibility of our approach by using AGENT on two experiments published in the VRST'16 proceedings.

7.2.7. Create and Play your Pac-Man Game with the GEMOC Studio

Executable Domain-Specific Languages (DSLs) are used for defining the behaviors of systems. In particular, the operational semantics of such DSLs may define how conforming models react to stimuli from their environment. This commonly requires adapting the semantics to define both the possible domain-level stimuli, and their handling during the execution. However, manually adapting the semantics for such cross-cutting concern is a complex and error-prone task. In , we demonstrate a tool addressing this problem by allowing the augmentation of operational semantics for handling stimuli, and by automatically generating a complete behavioral language interface from this augmentation. At runtime, this interface can receive stimuli sent to models, and can safely handle them by automatically interrupting the execution flow. This tool is an extension to the GEMOC Studio, a language and modeling workbench for executable DSLs We demonstrate how it can be used to implement a Pac-Man DSL enabling to create and play Pac-Man games.

7.3. Results on Heterogeneous and dynamic software architectures

We have selected three main contributions for DIVERSE's research axis #4: one is in the field of runtime management, while the two others one are in the field of Privacy and Security.

7.3.1. Verifying the configuration of Virtualized Network Functions in Software Defined Networks

In Kevoree, one of the goal is to work on the shipping pases in which we aim at making deployment, and the reconfiguration simple and accessible to the whole team. This year we work to include the capacity to manage network configuration when reconfiguring application stack. In this context, the deployment of modular virtual network functions (VNFs) in software defined infrastructures (SDI) enables cloud and network providers to deploy integrated network services across different resource domains. It leads to a large interleaving between network configuration through software defined network controllers and VNF deployment within this network. Most of the configuration management tools and network orchestrator used to deploy VNF lack of an abstraction to express Assume-Guarantee contracts between the VNF and the SDN configuration. Consequently, VNF deployment can be inconsistent with network configurations.

Contribution. To tackle this challenge, in this work [41], we develop an approach to check the consistency between the VNF description described from a set of structural models and flow-chart models and a proposed deployment on a real SDN infrastructure with its own configuration manager. We illustrate our approach on virtualized Evolved Packet Core function.

Originality. The originality of this work is to propose a model to capture VNF.

Impact. Beyond the scientific originality of this work, the main impacts of this novel approach to check SDN configuration has been to (i) reinforce DIVERSE's visibility in the academic and industrial communities on software components and (ii) to create several research tracks that are currently explored in different projects of the team (B-com PhD thesis and Nokia common labs). This work is being integrated within the Kevoree platform.

7.3.2. Identity Negotiation at Runtime

Authentication delegation is a major function of the modern web. Identity Providers (IdP) acquired a central role by providing this function to other web services. By knowing which web services or web applications access its service, an IdP can violate the end-user privacy by discovering information that the user did not want to share with its IdP. For instance, WebRTC introduces a new field of usage as authentication delegation happens during the call session establishment, between two users. As a result, an IdP can easily discover that Bob has a meeting with Alice. A second issue that increases the privacy violation is the lack of choice for the end-user to select its own IdP. Indeed, on many web-applications, the end-user can only select between a subset of IdPs, in most cases Facebook or Google.

Contribution. This year, we analyze this phenomena [23], in particular why the end-user cannot easily select its preferred IdP, though there exists standards in this field such as OpenID Connect and OAuth 2? To lead this analysis, we conduct three investigations. The first one is a field survey on OAuth 2 and OpenID Connect scope usage by web sites to understand if scopes requested by web-sites could allow for user defined IdPs. The second one tries to understand whether the problem comes from the OAuth 2 protocol or its implementations by IdP. The last one tries to understand if trust relations between websites and IdP could prevent the end user to select its own IdP. Finally, we sketch possible architecture for web browser based identity management, and report on the implementation of a prototype. We also describe our implementation of the WebRTC identity architecture [24]. We adapt OpenID Connect servers to support WebRTC peer to peer authentication and detail the issues and solutions found in the process.

Originality. We observe that although WebRTC allows for the exchange of identity assertion between peers, users lack feedback and control over the other party authentication. To allow identity negotiation during a WebRTC communication setup, we propose an extension to the Session Description Protocol. Our implementation demonstrates current limitations with respect to the current WebRTC specification.

Impact. This work is done with Orange.

7.3.3. Raising Time Awareness in Model-Driven Engineering

The conviction that big data analytics is a key for the success of modern businesses is growing deeper, and the mobilisation of companies into adopting it becomes increasingly important. Big data integration projects enable companies to capture their relevant data, to efficiently store it, turn it into domain knowledge, and finally monetize it. In this context, historical data, also called temporal data, is becoming increasingly available and delivers means to analyse the history of applications, discover temporal patterns, and predict future trends. Despite the fact that most data that today's applications are dealing with is inherently temporal current approaches, methodologies, and environments for developing these applications don't provide sufficient support for handling time. We envision that Model-Driven Engineering (MDE) would be an appropriate ecosystem for a seamless and orthogonal integration of time into domain modeling and processing.

Contribution. This year, we investigate the state-of-the-art in MDE techniques and tools in order to identify the missing bricks for raising time-awareness in MDE and outline research directions in this emerging domain [30].

Originality. We propose an extended context representation for self-adaptive software that integrates the history of planned actions as well as their expected effects over time into the context representations. We demonstrate on a cloud elasticity manager case study that such *temporal action-aware context* leads to improved reasoners while still be highly scalable. This work is original with respect to the state of the art since it provides a way to represent and take into account the impact of reconfiguration actions on a system.

Impact. This work is done through a collaboration with the SnT in Luxembourg and a startup called DataThings, working on domain model representation for various industrial domains.

7.3.4. Collaborations

This year, we had a close and fruitful collaboration with the industrial partners that are involved in the HEADS and Occiware projects, in particular an active interaction with the Tellu company in Norway in the Heads context. Tellu relies on Kevoree and KevoreeJS to build their health management systems. They will be also an active member the new Stamp project led by DIVERSE. We can cite also an active collaboration with Orange Labs through Kevin Corre's joint PhD thesis. Another joint industrial (CIFRE) PhD started in September 2016, and we are also partner in a new starting FUI project. Finally, DIVERSE collaborates with the B-COM IRT (https://b-com.com/en), as one permanent member has a researcher position of one day per week at B-COM and a new joint PhD started in September.

At the academic level we collaborate actively with the Spiral team at Inria Lille (several joint projects), the Tacoma team (with two co-advised PhD students), the Myriad team (1 co-advised PhD student) and we have started two collaborations with the ASAP team.

7.4. Results on Diverse Implementations for Resilience

Diversity is acknowledged as a crucial element for resilience, sustainability and increased wealth in many domains such as sociology, economy and ecology. Yet, despite the large body of theoretical and experimental science that emphasizes the need to conserve high levels of diversity in complex systems, the limited amount of diversity in software-intensive systems is a major issue. This is particularly critical as these systems integrate multiple concerns, are connected to the physical world, run eternally and are open to other services and to users. Here we present our latest observational and technical results about (i) new approaches to increase diversity in software systems, and (ii) software testing to assess the validity of software.

7.4.1. Software diversification

Our work on software diversification explores various ways of adding randomness in program executions: state perturbations that preserve functional correctness [25]; randomizing of web APIs to mitigate browser fingeprinting [33].

Can the execution of software be perturbed without breaking the correctness of the output? In this work [25], we devise a protocol to answer this question from a novel perspective. In an experimental study, we observe that many perturbations do not break the correctness in ten subject programs. We call this phenomenon "correctness attraction". The uniqueness of this protocol is that it considers a systematic exploration of the perturbation space as well as perfect oracles to determine the correctness of the output. To this extent, our findings on the stability of software under execution perturbations have a level of validity that has never been reported before in the scarce related work. A qualitative manual analysis enables us to set up the first taxonomy ever of the reasons behind correctness attraction.

The rich programming interfaces (APIs) provided by web browsers can be diverted to collect a browser fingerprint. A small number of queries on these interfaces are sufficient to build a fingerprint that is statistically unique and very stable over time. Consequently, the fingerprint can be used to track users. Our work [33] aims at mitigating the risk of browser fingerprinting for users privacy by 'breaking' the stability of a fingerprint over time. We add randomness in the computation of selected browser functions, in order to have them deliver slightly different answers for each browsing session. Randomization is possible thanks to the following properties of browsers implementations: (i) some functions have a nondeterministic specification, but a deterministic implementation ; (ii) multimedia functions can be slightly altered without deteriorating user's perception. We present FPRandom, a modified version of Firefox that adds randomness to mitigate the most recent fingerprinting algorithms, namely canvas fingerprinting, AudioContext fingerprinting and the unmasking of browsers through the order of JavaScript properties. We evaluate the effectiveness of FPRandom by testing it against known fingerprinting tests. We also conduct a user study and evaluate the performance overhead of randomization to determine the impact on the user experience.

The other aspect in the area of software diversity is about the statistical analysis of browser fingerprinting on a large industrial dataset [17], [31].

7.4.2. Software testing

Generative software development has paved the way for the creation of multiple code generators and compilers that serve as a basis for automatically generating code to a broad range of software and hardware platforms. With full automatic code generation, the user is able to easily and rapidly synthetize software artifacts for various software platforms. In addition, modern generators (i.e., C compilers) become highly configurable, offering numerous configuration options that the user can use to easily customize the generated code for the target hardware platform. In this context, it is crucial to verify the correct behaviour of code generators. Numerous approaches have been proposed to verify the functional outcome of generated code but few of them evaluate the non-functional properties of automatically generated code, namely the performance and resource usage properties. The thesis of Mohamed Boussaa [16] has addressed this limitation.

FOCUS Project-Team

7. New Results

7.1. Service-Oriented Computing

Participants: Mario Bravetti, Maurizio Gabbrielli, Saverio Giallorenzo, Claudio Guidi, Ivan Lanese, Cosimo Laneve, Fabrizio Montesi, Davide Sangiorgi, Gianluigi Zavattaro.

7.1.1. Microservices

Microservices represent an architectural style inspired by service-oriented computing that has recently started gaining popularity. As we have discussed in [37], one of the main advantages of the microservices approach is that it improves scalability of the developed applications. In [43] we have analyzed the impact of microservices on the overall line of research on software architectures, by pointing out specific open problems and future challenges. One of the challenges is concerned with programming languages because microservice systems are currently developed using general-purpose programming languages that do not provide dedicated abstractions for service composition. In [46] we have discussed the limitations of the current practices and we have proposed a novel language-based approach to the engineering of microservices based on the Jolie programming language.

7.1.2. Orchestrations and choreographies

The practice of programming distributed systems is extremely error-prone, due to the complexity in correctly implementing separate components that, put together, enact an agreed protocol. Theoretical and applied research is, therefore, fundamental, to explore new tools to assist the development of such systems. In particular, usage of so-called session types in orchestration languages guarantees correct communication by means of corresponding type system theories. In this context, we carried out studies about: foundations of the classical theory of session types, by providing an encoding into pi-calculus typing [17]; and subtyping in the context of asynchronous communication, showing it to be an undecidable problem [15]. Choreographies are also an important specification tool in that they can be compiled to obtain projected orchestrations that enjoy deadlock freedom by construction. Moreover they allow one to express dynamic behaviours at the level of the whole system, which then reflect on each involved orchestration. In this context, in [16] we studied the theory and implementation of dynamic choreographies and in [45] we showed how to use them for programming microservice-based applications. Finally, we considered applications in the context of Mobility-as-a-Service (MaaS) scenarios, where solutions of different transportation providers are dynamically composed into a single, consistent interface. We devised the prototype of an enabling software platform for MaaS [27] and we studied MaaS security issues [28].

7.2. Models for Reliability

Participant: Ivan Lanese.

7.2.1. Reversibility

We have continued the study of reversibility started in the past years. In particular, in [19], we thoroughly studied causal-consistent reversibility in the coordination language μKlaim [50], a distributed version of Linda [49]. More specifically, we gave an abstract specification of a causal-consistent rollback operator and showed that our semantics satisfies it. The main novelty of μKlaim w.r.t. process calculi studied in past work is that it includes a primitive to read a datum without consuming it, that, from the causality point of view, creates asymmetric dependencies. The same technique could be used to reverse languages with shared memory.

In [24] we studied how to exploit reversibility to improve client-server interactions. In particular, we defined retractable contracts, namely contracts including the possibility of undoing past agreements, which are more expressive than standard session contracts for binary interactions [48], yet preserve their nice properties: compliance and the subcontract relation are both decidable in polynomial time, the dual of a contract always exists and has a simple syntactic characterization. Furthermore we showed that the same contracts can also describe speculative interactions.

7.3. Probabilistic Systems and Resource Control

Participants: Martin Avanzini, Raphaelle Crubillé, Ugo Dal Lago, Francesco Gavazzo, Charles Grellois, Davide Sangiorgi, Valeria Vignudelli.

7.3.1. Probabilistic termination

In Focus, we are interested in studying probabilistic higher-order programming languages and, more generally, the fundamental properties of probabilistic computation when placed in an interactive scenario. One of the most basic (but nevertheless desirable) properties of programs is certainly termination. When probabilistic choice comes into play, termination can be defined in more than one way. As an example, one can stipulate that a probabilistic program terminates if and only if its probability of convergence is 1, this way being *almost surely* terminating. Alternatively, a probabilistic program can be said to be *positively* almost surely terminating if its average runtime is finite. The latter condition easily implies the former. Termination, already undecidable for deterministic (universal) programming languages, remains so in presence of probabilistic choice. Actually, it becomes provably harder, being strictly higher in the arithmetical hierarchy. Probabilistic termination has received quite some attention in recent years, but most contributions are concerned either with its abstract nature, or with verification methodologies for imperative programs. Along 2017, we have initiated the study of probabilistic termination in probabilistic higher-order functional languages. Our contribution in this direction is twofold. On the one hand, we have analysed the impact of endowing a strongly normalising typed lambda calculus, namely Godel's **T**, with various forms of probabilistic choice operators [26]. Unsurprisingly, the obtained systems are all almost surely terminating, but interestingly, only *some* of them are positively so. In particular, binary probabilistic choice and the geometric distribution can have dramatically different effects. Another line of work has to do with types, and in particular with sized types, which we have generalised to a higher-order functional language with higher order recursion and binary probabilistic programs. We showed how the obtained system is sound for almost sure termination [35], but also that it captures interesting examples like various forms of random walks.

7.3.2. Automating complexity analysis of higher-order functional programs

Complexity analysis of higher-order functional programs has been one of the core research directions within Focus since its inception. Progressively, however, our interest has shifted from foundations to automation. The latter is indeed the main research direction we have pursued in 2017. More specifically, we have been trying to overcome the main shortcoming of our software tool HoCA, namely the fact that most analysis techniques it implements are not modular, and are thus bound not to scale. We have looked at sized type systems as a way to do complexity analysis of functional programs by performing type inference on a so-called ticking-transformed version of them [13]. The obtained design methodology has been proved to allow the analysis of programs which could not be handled by HoCA.

7.3.3. Relational reasoning about effectful and concurrent programs

Building on our knowledge on semantic and coinductive techniques for reasoning about higher-order programs, we have studied how to reason relationally when the programs at hand exhibit some form of effect including probabilistic choice, but also algebraic effects. We have first of all concluded our investigation about metric reasoning about terms in a probabilistic lambda calculus. We discovered that in the general case of a fully-fledged probabilistic lambda calculus, any reasonable metric is bound to trivialise to an equivalence [31]. This negative result convinced us that a richer and more refined notion of comparison is needed, on which we are currently investigating. We also looked at how Abramsky's applicative bisimilarity can be generalised to a

language with algebraic effects. Since the notion of algebraic effect is abstract, this is best done by injecting concepts from category theory, and in particular those of a monad and of a relator, into the playground. Mild conditions on the latter allow one to generalise the classic proof of congruence for applicative bisimilarity, due to Howe [33], [34]. This way, conductive proof techniques for equivalence can be shown sound with respect to context equivalence for various forms of algebraic effects including probabilistic choice, global state, exceptions, and combinations. One last line of work we have pursued in 2017 has to do with geometry of interaction, a dynamic semantic framework which is known to faithfully model higher-order computation. We have this year managed to show that multitoken machines, a generalisation of geometry of interaction we introduced three years ago, can faithfully model quantum lambda calculi [32], but also process algebras like the π-calculus, through multiport interaction combinators [14].

7.4. Verification Techniques

Participants: Mario Bravetti, Adrien Durier, Daniel Hirschkoff, Ivan Lanese, Cosimo Laneve, Davide Sangiorgi.

We analyze sensible properties of concurrent systems, including deadlock freedom and resource usages, and proof techniques for deriving behavioural equalities and preorders on processes.

7.4.1. Deadlock detection and cloud elasticity

In order to verify sensible properties of concurrent programs we use a technique consisting of (1) extracting information by means of behavioural type systems and (2) analyzing types by means of ad-hoc tools.

In [20] we study deadlock detection for value-passing CCS (and for π calculus). In this paper we analyze complex programs that create networks with arbitrary numbers of nodes. To enable the analysis of such programs, (1) we define an algorithm for detecting deadlocks of a basic model featuring recursion and fresh name generation, and (2) we design a type system that returns behavioural types. We show the soundness of the type system, and develop a type inference algorithm for it.

In [39] we apply the above technique to a language for stateful active objects. This is challenging because active objects use futures to refer to results of pending asynchronous invocations and because these futures can be stored in object fields, passed as method parameters, or returned by invocations. The type system traces the access to object fields by means of effects. For this reason, it is possible to compute behavioural types that express synchronisation patterns in a precise way. The behavioural types are thereafter analysed by a solver that discovers potential deadlocks. The PhD thesis of Vincenzo Mastandrea [11] addresses deadlock detection of stateful active objects.

In [44] we apply the same technique to Java byte-code. In particular [44] gives a practical presentation of JaDA, a static deadlock analyzer for Java that extracts behavioral types and analyzes these types by means of a fixpoint algorithm that reports potential deadlocks in the original Java code. We also present some of the features for customising the analysis: while the main strength of JaDA is to run in a fully automatic way, user interaction is possible and may enhance the accuracy of the results. The whole theory behind JaDa is fully developed in the PhD thesis of Abel Garcia Celestrin [10].

In [18] we address a concurrent language with explicit acquire and release operations on virtual machines. In our language it is possible to delegate other (ad-hoc or third party) concurrent code to release virtual machines (by passing them as arguments of invocations). In this case, we define (i) a type system associating programs with behavioural types that record relevant information for resource usage (creations, releases, and concurrent operations), (ii) a translation function that takes behavioural types and returns cost equations, and (iii) an automatic off-the-shelf solver for the cost equations. A soundness proof of the type system establishes the correctness of our technique with respect to the cost equations. We have experimentally evaluated our technique using a cost analysis solver. The experiments show that our analysis allows us to derive bounds for programs that are better than other techniques, such as those based on amortized analysis.

7.4.2. Most general property-preserving updates

Systems need to be updated to last for a long time in a dynamic environment, and to cope with changing requirements. It is important for updates to preserve the desirable properties of the system under update, while possibly enforcing new ones. We consider a simple yet general update mechanism [25] that replaces a component of the system with a new one. The context, i.e., the rest of the system, remains unchanged. We define contexts and components as Constraint Automata interacting via either asynchronous or synchronous communication, and we express properties using Constraint Automata too. Then we build most general updates which preserve specific properties, considering both a single property and all the properties satisfied by the original system, in a given context or in all possible contexts.

7.4.3. Proof techniques based on unique solutions

In [22], we study bisimilarity, a behavioural equivalence whose success is much due to the associated bisimulation proof method. In particular, we discuss a different proof method, based on unique solution of special forms of inequations called contractions, and inspired by Milner's theorem on unique solution of equations. The method is as powerful as the bisimulation proof method and its up-to context enhancements. The definition of contraction can be transferred onto other behavioural equivalences, possibly contextual and non-coinductive. This enables a coinductive reasoning style on such equivalences, either by applying the method based on unique solution of contractions, or by injecting appropriate contraction preorders into the bisimulation game.

In [38] we develop the above proof method in a different direction: rather than introducing contractions, we remain within equations, and we investigate conditions that guarantee unique solutions, for bisimilarity as well as for other behavioural equivalences such as trace equivalence. We also consider preorders such as trace inclusion. We finally develop abstract formulations of the theorems, on generic Labeled Transition Systems.

7.4.4. Fuzzy logics

In [14] we introduce a framework for detecting anomalies in the clocks of the different components of a network of sensor stations connected with a central server for measuring air quality. We propose a novel approach, supported by a formal representation of the network using fuzzy-timed automata, to precisely represent the expected behaviour of each component of the network. Using fuzzy logic concepts, we can specify admissible mismatches between the clocks.

7.5. Computer Science Education

Participants: Michael Lodi, Simone Martini.

We study why and how to teach computer science principles (nowadays often referred to as "computational thinking", CT), in particular in the context of K-12 education (students aged approximately from 5 to 18). We study philosophical, sociological and historical motivations to teach computer science at all school levels. Furthermore, we study what concepts and skills related to computer science are not barely technical abilities, but have a general value for all students. Finally we try to find/produce/evaluate suitable materials (tools, languages, lesson plans...) to teach these concepts, taking into account: difficulties in learning CS concepts (particularly programming); stereotypes about computer science (particularly gender related issues); teacher training (particularly non specialist teachers).

7.5.1. Computational thinking definition

There is no accepted definition of computational thinking. From one hand we tried to find out the main common elements in the most important proposed definitions, and investigate, in a large sample of K-12 teachers, if they have a correct idea [30] about CT. We found the vast majority of them held misconceptions or partial views about it. We argued these may be consequences of a massive use of the term in school context [23]; we made clear "computational thinking" is not a new subject, but just a name to indicate computer science principles that should be taught to all students [21].

7.5.2. Evaluation of popularization initiatives

We analyzed [29] the sentiment of a large sample of teachers participating in the national project "Programma il Futuro" (Program the Future) - an italian version of Code.org with support materials. The sentiment was largely positive. Among other results, we note reported interest is equally distributed between male and female students in primary school, and shifts towards a higher male interest only from secondary school, suggesting a social influence.

7.5.3. Growth mindset and teacher training

Every person holds an idea (mindset) about intelligence: someone thinks it is a fixed trait, like eye color (fixed mindset), while others think it can grow like muscles (growth mindset). The latter is beneficial for students to have better results, particularly in STEM disciplines, and to not being influenced by stereotypes. Computer science is a subject that can be affected by fixed ideas ("geek gene") and some (small) studies showed it can induce fixed ideas. Teachers' mindset directly affects students' one. We propose [42] a line of research to investigate mindset of pre-service primary school teachers before and after a "creative computing course", to analyze and, in perspective, to change their specific "computer science mindset".

7.6. Constraint Programming

Participants: Maurizio Gabbrielli, Liu Tong.

In Focus, we sometimes make use of constraint solvers (e.g., cloud computing, service-oriented computing). Since a few years we have thus began to develop tools based on constraints. This year, besides refining the work on SUNNY (described elsewhere, see also [40]) we have developed a new tool, NightSplitter, a scheduling tool to optimize (sub)group activities [41]. Humans are social animals and usually organize activities in groups. However, they are often willing to split temporarily a bigger group in subgroups to enhance their preferences. NightSplitter is an on-line tool that is able to plan movie and dinner activities for a group of users, possibly splitting them in subgroups to optimally satisfy their preferences. We first have modeled and proved that this problem is NP-complete. We have then used Constraint Programming (CP) or alternatively Simulated Annealing (SA) to solve it. Empirical results show the feasibility of the approach even for big cities where hundreds of users can select among hundreds of movies and thousands of restaurants. (More information on NightSplitter is found in the section on tools.)

INDES Project-Team

5. New Results

5.1. Type Abstraction for Relaxed Noninterference

Information-flow security typing statically prevents confidential information to leak to public channels. The fundamental information flow property, known as *noninterference*, states that a public observer cannot learn anything from private data. As attractive as it is from a theoretical viewpoint, noninterference is impractical: real systems need to intentionally declassify some information, selectively. Among the different information flow approaches to declassification, a particularly expressive approach was proposed by Li and Zdancewic, enforcing a notion of *relaxed noninterference* by allowing programmers to specify *declassification policies* that capture the intended manner in which public information can be computed from private data. The paper [15] shows how we can exploit the familiar notion of type abstraction to support expressive declassification policies in a simpler, yet more expressive manner. In particular, the type-based approach to declassification—which we develop in an object-oriented setting—addresses several issues and challenges with respect to prior work, including a simple notion of label ordering based on subtyping, support for recursive declassification policies, and a local, modular reasoning principle for relaxed noninterference. This work paves the way for integrating declassification policies in practical security-typed languages.

5.2. Multiparty Reactive Sessions

Synchronous reactive programming (SRP) is a well-established programming paradigm whose essential features are logical instants, broadcast events and event-based preemption. This makes it an ideal vehicle for the specification and analysis of reactive systems, and indeed several programming languages and frameworks based on SRP have been put forward. On the other hand, *session-based concurrency* is the model of concurrent computation induced by *session types*, a rich typing discipline designed to specify the structure of interactions. In a nutshell, session types describe communication protocols between two or more participants by specifying the sequencing of messages along communication channels, as well as their functionality (sender, receiver and type of carried data). Originally conceived as a static analysis technique for an enhanced version of the π-calculus, session types have been subsequently transferred to functional, concurrent, and object-oriented programming languages, and adapted to support run-time verification.

A combination of session-based concurrency and SRP features appears to be appropriate to specify and analyse communication-centric systems in which some components may have a reactive and/or timed behaviour. In joint work with colleagues from I3S and the University of Groningen, currently submitted, we study the integration of SRP and session-based concurrency. To this end, we propose a calculus for multiparty sessions enriched with features from SRP. In this calculus, protocol participants communicate by broadcast messages, have the ability to suspend themselves while waiting for an absent message, and may react to the presence of particular events by triggering alternative behaviours. We equip the calculus with a session type system which enforces expected session properties such as communication safety, protocol fidelity, and input lock freedom. This session type system departs significantly from existing ones: the interplay of classical, well-established assumptions of SRP with session-based constructs requires revisiting central notions of multiparty session types, such as those of global type, local type and projection.

5.3. Multiparty Reversible Sessions

Reversibility has been an active trend of research for the last fifteen years. A reversible computation is a computation that has the ability to roll back to a past state. Allowing computations to reverse is a means to improve system flexibility and reliability. In the setting of concurrent process calculi, reversible computations have been first studied for CCS, then for the π-calculus, and only recently for session calculi. In [14] we present a multiparty session calculus with reversible computations. Our proposal improves on existing reversible

session calculi in several respects: it allows for concurrent and sequential composition within processes and types, it gives a compact representation of the *past* of processes and types, which facilitates the definition of rollback, and it implements a fine-tuned strategy for backward computation. We propose a refined session type system for this calculus and show that it enforces the expected properties of session fidelity, forward and backward progress, as well as causal consistency. In conclusion, our calculus is a conservative extension of previous proposals, offering enhanced expressive power and refined analysis techniques.

5.4. JavaScript ahead-of-time compilation

Nowadays, JavaScript is no longer confined to the programming of web pages. It is also used for programming server-side parts of web applications, compilers, and there is a growing trend for using it for programming internet-of-things (IoT) applications. All major industrial actors of the field are looking for, or are already providing, JavaScript based development kits (IoT.js, Espruino, JerryScript, Kinoma.js, ...). In this application domain, JavaScript programs execute on tiny devices that have limited hardware capacities, for instance only a few kilobytes of memory. Just-in-time (JIT) compilation, which has proved to be so effective for improving JavaScript performances, is unthinkable in these constrained environments. There would be just not enough memory nor CPU capacity to execute them at runtime. Pure JavaScript interpreters are then used but this comes with a strong performance penalty, especially when compared to assembly or C programs, that limits the possible uses.

When JIT compilation is not an option and when interpretation is too slow, the alternative is static compilation, also known as ahead-of-time (AOT) compilation. It has the promise of combining small memory footprints and good performances. However, this implementation technique seems not to fit the JavaScript design whose unique combination of antagonistic features such as functional programming support, high mutation rates of applications, introspection, and dynamicity, makes most known classical AOT compilation techniques ineffective.

Indeed, JavaScript is hard to compile, much harder than languages such as C, Java, and even harder than other functional languages like Scheme and ML. This is because a JavaScript source code accepts many more possible interpretations than other languages do. It forces JavaScript compilers to adopt a defensive position by generating target codes that can cope with all the possible, even unlikely, interpretations. This difficulty probably explains why JavaScript AOT compilation has received so little attention from the scientific community. All these difficulties cannot be solved with traditional compilation techniques. They demand new strategies. This is what we explore. We are developing a prototype of a new compiler that distinguishes from classical compilers by relying on static program analyses that are not governed by approximating all possible program executions but by inferring properties that suit the compiler back-end. For instance, instead of inferring types that describe a super set of all possible executions, this compiler infers types for which the compiler is able to deliver good code.

The whole year has been devoted to implementing an operational prototype of the compiler. The preliminary results we have obtained are very promising but we still have to improve the code generation quality before writing and publishing complete reports describing it. This will be one of our main objectives for 2018.

5.5. Orchestration of Web applications

Modern Web applications are composed of numerous heterogeneous actors (users, distant servers and services, IoT devices, etc.) interacting together by means of asynchronous events. The harmonious interaction between these actors is called *orchestration*. JavaScript, the mainstream language for writing Web applications, enables programmers to orchestrate events with an asynchronous event-loop. However, event-loop based orchestration is known to be a difficult problem leading to programs which are difficult to write, read and maintain. To address this problem, Hiphop.js, a domain-specific language (DSL), has been developed during the last two years. It extends JavaScript by means of *temporal constructors* allowing explicit synchronization, parallelism and preemption. These constructors are inspired from the Esterel synchronous language.

During this year Hiphop.js has gained in maturity. First, a development environment has been developed. It is now possible to debug Hiphop.js programs by visualizing the code source during the execution, inspecting instructions state and signals value. Moreover, it is possible to queue reactions in order to analyze step-by-step the global state of the program between each reaction. The debugger can also be used and controlled remotely, using a simple Web browser. It is an important feature since Hiphop.js applications can run on different types of devices, especially smartphones or headless devices, on which debugging is impossible. Besides, in order to have a deeper integration with JavaScript and to make the adoption of Hiphop.js easier for new users, a new syntax has been designed and implemented.

A short paper describing HipHop has been accepted for publication at the SAC'18 symposium.

Finally, Hiphop.js is used in the context of a music show during the *MANCA* (http://www.cirm-manca.org/manca2017/) festival in Nice. It is used to orchestrate the composition of lights and songs during the show. Moreover, the public can interact with musicians by the means of smartphones, playing specific songs during delimited periods of the performance. Those interactions are implemented using Hiphop.js.

5.6. On the Content Security Policy Violations due to the Same-Origin Policy

Modern browsers implement different security policies such as the Content Security Policy (CSP), a mechanism designed to mitigate popular web vulnerabilities, and the Same Origin Policy (SOP), a mechanism that governs interactions between resources of web pages.

In the work [17], we describe how CSP may be violated due to the SOP when a page contains an embedded iframe from the same origin. We analyse 1 million pages from 10,000 top Alexa sites and report that at least 31.1% of current CSP-enabled pages are potentially vulnerable to CSP violations. Further considering real-world situations where those pages are involved in same-origin nested browsing contexts, we found that in at least 23.5% of the cases, CSP violations are possible.

During our study, we also identified a divergence among browsers implementations in the enforcement of CSP in srcdoc sandboxed iframes, which actually reveals a problem in Gecko-based browsers CSP implementation. To ameliorate the problematic conflicts of the security mechanisms, we discuss measures to avoid CSP violations.

5.7. Control What You Include! Server-Side Protection Against Third Party Web Tracking

Third party tracking is the practice by which third parties recognize users accross different websites as they browse the web. Recent studies show that 90% of websites contain third party content that is tracking its users across the web. Website developers often need to include third party content in order to provide basic functionality. However, when a developer includes a third party content, she cannot know whether the third party contains tracking mechanisms. If a website developer wants to protect her users from being tracked, the only solution is to exclude any third-party content, thus trading functionality for privacy.

We describe and implement a privacy-preserving web architecture [16] that gives website developers a control over third party tracking: developers are able to include functionally useful third party content, and at the same time ensuring that the end users are not tracked by the third parties.

5.8. A Better Facet of Dynamic Information Flow Control

Multiple Facets (MF) is a dynamic enforcement mechanism which has proved to be a good fit for implementing information flow security for JavaScript. It relies on multi executing the program, once per each security level or view, to achieve soundness. By looking inside programs, MF encodes the views to reduce the number of needed multi-executions.

We extend Multiple Facets in three directions. First, we propose a new version of MF for arbitrary lattices, called Generalised Multiple Facets, or GMF. GMF strictly generalizes MF, which was originally proposed for a specific lattice of principals. Second, we propose a new optimization on top of GMF that further reduces the number of executions. Third, we strengthen the security guarantees provided by Multiple Facets by proposing a termination sensitive version that eliminates covert channels due to termination.

5.9. Impossibility of Precise and Sound Termination Sensitive Security Enforcements

An information flow policy is termination sensitive if it imposes that the termination behaviour of programs is not influenced by confidential input. Termination sensitivity can be statically or dynamically enforced. On one hand, existing static enforcement mechanisms for termination sensitive policies are typically quite conservative and impose strong constraints on programs like absence of while loops whose guard depends on confidential information. On the other hand, dynamic mechanisms can enforce termination sensitive policies in a less conservative way. SME, one of such mechanisms, was even claimed to be sound and precise in the sense that the enforcement mechanism will not modify the observable behaviour of programs that comply with the termination sensitive policy. However, termination sensitivity is a subtle policy, that has been formalized in different ways. A key aspect is whether the policy talks about actual termination, or observable termination.

We prove that termination sensitive policies that talk about actual termination are not enforceable in a sound and precise way. For static enforcements, the result follows directly from a reduction of the decidability of the problem to the halting problem. However, for dynamic mechanisms the insight is more involved and requires a diagonalization argument.

In particular, our result contradicts the claim made about SME. We correct this claim by showing that SME enforces a subtly different policy that we call indirect termination sensitive noninterference and that talks about observable termination instead of actual termination. We construct a variant of SME that is sound and precise for indirect termination sensitive noninterference. Finally, we also show that static methods can be adapted to enforce indirect termination sensitive information flow policies (but obviously not precisely) by constructing a sound type system for an indirect termination sensitive policy.

5.10. BELL: Browser fingerprinting via Extensions and Login-Leaks

Recent work showed that websites can detect browser extensions that users install and websites they are logged into. This poses significant privacy risks, since extensions and Web logins can leak sensitive information and be used to track users via fingerprinting.

In joint work with Gabor Gulyas and Claude Castelluccia (Privatics team, Inria Grenoble), we report on the first large-scale study of this new form of fingerprinting, based on more than 16,000 users who visited our website [0]. Our website identifies installed Google Chrome extensions via Web Accessible Resources, and detects logged in websites by methods that rely on URL redirection and CSP violation report. Our website is able to test and detect the presence of 16,743 Chrome extensions, covering 28% of all free Chrome extensions. We also test whether the user is connected to 60 different websites.

We compute uniqueness of collected fingerprints, and find out that 54.86% of users that have installed at least one detectable extension are unique; 19.53% are unique because they logged in one or more detectable websites; and 89.23% of users are unique because they have at least one extension and one login detected.

We optimize the fingerprinting algorithm and show that it is possible to fingerprint a user in less than 625 milliseconds by selecting the most identifying combinations of extensions. Moreover, we discover that 22.98% of users can be uniquely identified and tracked by Web logins, even if they disable JavaScript. We conclude with possible countermeasures.

[0] https://extensions.inrialpes.fr/

5.11. Large-scale measurement of invisible images for Web tracking

In joint work with Arnaud Legout (DIANA team, Inria Sophia Antipolis), we perform large scale Web measurements to evaluate Web tracking and privacy leaks in every-day Web browsing. Unlike the related work, our study focuses on the third-party HTTP requests for invisible images. We have identified two types of images that are invisible to the end user and most likely used for Web tracking: one-pixel images and empty images.

We have visited 4,351,318 pages from 38,000 web sites and identified that almost half of the third-party images are invisible to the end user. This finding raises a lot of concerns regarding Web tracking and user privacy on the Web. We made the first evaluations on how much of this invisible tracking is prevented by the popular browser extensions used for the privacy protection. We also find the top invisible trackers and the invisible trackers not blocked by the browser extensions.

We continue this work by analysing all the HTTP requests and responses that lead to invisible images in order to (1) provide a fine-grained classification of third-party cookie tracking; (2) analyse new techniques of cookie-synching used by the companies; (3) evaluate redirection chains that lead to user's information exchange between various companies; (4) identify companies that use invisible images for none of the known tracking techniques, and analyse such requests and responses further to reveal new Web tracking technologies.

PHOENIX Project-Team

7. New Results

7.1. Everyday Functioning Benefits from an Assisted Living Platform amongst Frail Older Adults and Their Caregivers

Ambient assisted living technologies (AAL) are regarded as a promising solution to support aging in place. Yet, their efficacy has to be demonstrated in terms of benefits for independent living and for work conditions of caregivers. Hence, the purpose of this study was to assess the benefits of a multi-task AAL platform for both Frail older Individuals (FIs) and professional caregivers with respect to everyday functioning and caregiver burden. In this context, a 6-month field study involved 32 FIs living at home (half of them were equipped by the platform and the remaining half were not, as a control condition) and their caregivers. Everyday functioning measures were reported by frail participants and caregivers. Self-reported burden measures of caregiver were also collected. The main results showed that the caregiver's estimates of everyday functioning of equipped participants were unchanged across time, while they decreased for the control participants. Also, a reduction of self-reported objective burden was obtained after 6 months of AAL intervention for the equipped group, compared to the control group. Overall, these results highlighted the potential of AAL as a relevant environmental support for preventing both functional losses in FIs and objective burden professional caregiver.

7.2. Designing Parallel Data Processing for Enabling Large-Scale Sensor Applications

Masses of sensors are being deployed at the scale of cities to manage parking spaces, transportation infrastructures to monitor traffic, and campuses of buildings to reduce energy consumption. These large-scale infrastructures become a reality for citizens via applications that orchestrate sensors to deliver high-value, innovative services. These applications critically rely on the processing of large amounts of data to analyze situations, inform users, and control devices. This work proposes a design-driven approach to developing orchestrating applications for masses of sensors that integrates parallel processing of large amounts of data. Specifically, an application design exposes declarations that are used to generate a programming framework based on the MapReduce programming model. We have developed a prototype of our approach, using Apache Hadoop. We applied it to a case study and obtained significant speedups by parallelizing computations over twelve nodes. In doing so, we demonstrate that our design-driven approach allows to abstract over implementation details, while exposing architectural properties used to generate high-performance code for processing large datasets. Furthermore, we show that this high-performance support enables new, personalized services in a smart city. Finally, we discuss the expressiveness of our design language, identify some limitations, and present language extensions.

7.3. Internet of Things: From Small-to Large-Scale Orchestration

The domain of Internet of Things (IoT) is rapidly expanding beyond research, and becoming a major industrial market with such stakeholders as major manufacturers of chips and connected entities (i.e., things), and fast-growing operators of wide-area networks. Importantly, this emerging domain is driven by applications that leverage an IoT infrastructure to provide users with innovative, high-value services. IoT infrastructures range from small scale (e.g., homes and personal health) to large scale (e.g., cities and transportation systems). In this work, we argue that there is a continuum between orchestrating connected entities in the small and in the large. We propose a unified approach to application development, which covers this spectrum. To do so, we examine the requirements for orchestrating connected entities and address them with domain-specific design concepts. We then show how to map these design concepts into dedicated programming patterns and runtime mechanisms. Our work revolves around domain-specific concepts and notations, integrated into a tool-based design methodology and dedicated to develop IoT applications. We have applied our work across a spectrum of infrastructure sizes, ranging from an automated pilot in avionics, to an assisted living platform for the home of seniors, to a parking management system in a smart city.

7.4. Designing an Accessible and Engaging Email Application for Aging in Place

Supporting independent everyday functioning of older adults is a major challenge for aging in place. In particular, communication and social activities need support to prevent social isolation, cognitive and psychosocial well-being decline, and a risk of depression. This paper focuses on how technology can bring social support to isolated older-old adults (over 75 years old) and allow them to communicate with members of their social network. We present the design of an accessible and engaging email application dedicated to this population. We propose design principles based on the older adults' specificities and then use these principles to develop a tablet-based email application. We conducted a field study to evaluate our email application during 9 months. We equipped 13 community-dwelling old-older adults with a touchscreen tablet and our application at their home (compared to 13 control counterparts). This field study validates our design principles as shown by the effectiveness and efficiency gained by the participants in using our application. Moreover, we reveal the influence of health indicators in the usage behaviors and the long-term use of our application.

7.5. HomeAssist: An Assisted Living Platform for Aging in Place Based on an Interdisciplinary Approach

HomeAssist is an assisted living platform aims to support aging in place. This platform was designed using a human-centered approach. It offers assistive services, addressing the main aspects of daily life: activities of daily living, home and user safety, and social participation. HomeAssist introduces key novel features: (1) it covers multiple aspects of daily life, addressing a variety of needs of older adults; (2) it provides customization mechanisms, adapting assistance to the user's abilities while preventing autonomy losses; (3) it relies on context awareness, delivering timely assistance; and, (4) it revolves around a unified user interface to achieve usability. All these features play a key role towards achieving high acceptance of HomeAssist and supporting autonomy effectively, as shown by our field study.

RMOD Project-Team

7. New Results

7.1. Software Quality: Testing and Tools

Testing Habits. What are the Testing Habits of Developers? We conducted a case study in a large IT company. Tests are considered important to ensure the good behavior of applications and improve their quality. But development in companies also involves tight schedules, old habits, less-trained developers, or practical difficulties such as creating a test database. As a result, good testing practices are not always used as often as one might wish. With a major IT company, we are engaged in a project to understand developers testing behavior, and whether it can be improved. Some ideas are to promote testing by reducing test session length, or by running automatically tests behind the scene and send warnings to developers about the failing ones. Reports on developers testing habits in the literature focus on highly distributed open-source projects, or involve students programmers. As such they might not apply to our industrial, closed source, context. We take inspiration from experiments of two papers of the literature to enhance our comprehension of the industrial environment. We report the results of a field study on how often the developers use tests in their daily practice, whether they make use of tests selection and why they do. Results are reinforced by interviews with developers involved in the study. The main findings are that test practice is in better shape than we expected; developers select tests ruthlessly (instead of launching an entire test suite); although they are not accurate in their selection, and; contrary to expectation, test selection is not influenced by the size of the test suite nor the duration of the tests. [23]

Tests in Open-Source. During the development, it is known that tests ensure the good behavior of applications and improve their quality. We studied developers testing behavior inside the Pharo community in the purpose to improve it. We report results of a field study on how often the developers use tests in their daily practice, whether they make use of tests selection and why they do. Results are strengthened by interviews with developers involved in the study. The main findings are that developers run tests every modifications of their code they did; most of the time they practice test selection (instead of launching an entire test suite); however they are not accurate in their selection; they change their selection depending on the duration of the tests and; contrary to expectation, test selection is not influenced by the size of the test suite. [35]

CodeCritics Applied to Database Schema: Challenges and First Results. Relational databases (DB) play a critical role in many information systems. For different reasons, their schemas gather not only tables and columns but also views, triggers or stored functions (i.e., fragments of code describing treatments). As for any other code-related artefact, software quality in a DB schema helps avoiding future bugs. However, few tools exist to analyze DB quality and prevent the introduction of technical debt. We present research issues related to assessing the software quality of a DB schema by adapting existing source code analysis research to database schemas. We present preliminary results that have been validated through the implementation of DBCritics, a prototype tool to perform static analysis on the SQL source code of a database schema. DBCritics addresses the limitations of existing DB quality tools based on an internal representation considering all entities of the database and their relationships. [26]

Recommending Source Code Locations for System Specific Transformations. From time to time, developers perform sequences of code transformations in a systematic and repetitive way. This may happen, for example, when introducing a design pattern in a legacy system: similar classes have to be introduced, containing similar methods that are called in a similar way. Automation of these sequences of transformations has been proposed in the literature to avoid errors due to their repetitive nature. However, developers still need support to identify all the relevant code locations that are candidate for transformation. Past research showed that these kinds of transformation can lag for years with forgotten instances popping out from time to time as other evolutions bring them into light. We evaluate three distinct code search approaches (structural, based on Information Retrieval, and AST based algorithm) to find code locations that would require similar transformations. We validate the resulting candidate locations from these approaches on real cases identified previously

in literature. The results show that looking for code with similar roles, e.g., classes in the same hierarchy, provides interesting results with an average recall of 87% and in some cases the precision up to 70%. [33]

Quality-oriented Move Method Refactoring. Restructuring is an important activity to improve software internal structure. Even though there are many restructuring approaches, very few consider the refactoring impact on the software quality. In this paper, we propose an semi-automatic software restructuring approach based on quality attributes. We rely on the measurements of the Quality Model for Object Oriented Design (QMOOD) to recommend Move Method refactorings that improve software quality. In our preliminary evaluation on three open-source systems, our approach achieved an average recall of 57%. [34]

7.2. Software Reengineering

The Case for Non-Cohesive Packages. While the lack of cohesiveness of modules in procedural languages is a good way to identify modules with potential quality problems, we doubt that it is an adequate measure for packages in object-oriented systems. Indeed, mapping procedural metrics to object-oriented systems should take into account the building principles of object-oriented programming: inheritance and late binding. Inheritance offers the possibility to create packages by just extending classes with the necessary increment of behavior. Late binding coupled to the "Hollywood Principle" are a key to build frameworks and let the users branch their extensions in the framework. Therefore, a package extending a framework does not have to be cohesive, since it inherits the framework logic, which is encapsulated in framework packages. In such a case, the correct modularization of an extender application may imply low cohesion for some of the packages. We confirm these conjectures on various real systems (JHotdraw, Eclipse, JEdit, JFace) using or extending OO frameworks. We carry out a dependency analysis of packages to measure their relation with their framework. The results show that framework dependencies form a considerable portion of the overall package dependencies. This means that non-cohesive packages should not be considered systematically as packages of low quality. [22]

Identifying Classes in Legacy JavaScript Code. JavaScript is the most popular programming language for the Web. Although the language is prototype-based, developers can emulate class-based abstractions in JavaScript to master the increasing complexity of their applications. Identifying classes in legacy JavaScript code can support these developers at least in the following activities: (i) program comprehension; (ii) migration to the new JavaScript syntax that supports classes; and (iii) implementation of supporting tools, including IDEs with class-based views and reverse engineering tools. We propose a strategy to detect class-based abstractions in the source code of legacy JavaScript systems. We report on a large and in-depth study to understand how class emulation is employed, using a dataset of 918 JavaScript applications available on GitHub. We found that almost 70% of the JavaScript systems we study make some usage of classes. We also performed a field study with the main developers of 60 popular JavaScript systems in order to validate our findings. The overall results range from 97% to 100% for precision, from 70% to 89% for recall, and from 82% to 94% for F-score. [20]

A Critical Analysis of String APIs: The Case of Pharo. Most programming languages, besides C, provide a native abstraction for character strings, but string APIs vary widely in size, expressiveness, and subjective convenience across languages. In Pharo, while at first glance the API of the String class seems rich, it often feels cumbersome in practice; to improve its usability, we faced the challenge of assessing its design. However, we found hardly any guideline about design forces and how they structure the design space, and no comprehensive analysis of the expected string operations and their different variations. We first analyze the Pharo 4 String library, then contrast it with its Haskell, Java, Python, Ruby, and Rust counterparts. We harvest criteria to describe a string API, and reflect on features and design tensions. This analysis should help language designers in understanding the design space of strings, and will serve as a basis for a future redesign of the string library in Pharo. [19]

7.3. Dynamic Languages: Language Constructs for Modular Design

An Experiment with Lexically-bound Extension Methods for a Dynamic Language. An extension method is a method declared in a package other than the package of its host class. Thanks to extension methods,

developers can adapt classes they do not own to their needs: adding methods to core classes is a typical use case. This is particularly useful for adapting software and therefore increasing reusability. In most dynamically-typed languages, extension methods are globally visible. Because any developer can define extension methods for any class, naming conflicts occur: if two developers define an extension method with the same signature in the same class, only one extension method is visible and overwrites the other. Similarly, if two developers each define an extension method with the same name in a class hierarchy, one overrides the other. Existing solutions typically rely on a dedicated and slow method lookup algorithm to resolve conflicts at runtime. We present a model of scoped extension methods that minimizes accidental overrides and we present an implementation in Pharo that incurs little performance overhead. This implementation is based on lexical scope and hierarchy-first strategy for extension scoping. [44]

Scoped Extension Methods in Dynamically-Typed Languages. An extension method is a method declared in a package other than the package of its host class. Thanks to extension methods, developers can adapt to their needs classes they do not own: adding methods to core classes is a typical use case. This is particularly useful for adapting software and therefore to increase reusability. Inquiry. In most dynamically-typed languages, extension methods are globally visible. Because any developer can define extension methods for any class, naming conflicts occur: if two developers define an extension method with the same signature in the same class, only one extension method is visible and overwrites the other. Similarly, if two developers each define an extension method with the same name in a class hierarchy, one overrides the other. To avoid such *accidental overrides*, some dynamically-typed languages limit the visibility of an extension method to a particular scope. However, their semantics have not been fully described and compared. In addition, these solutions typically rely on a dedicated and slow method lookup algorithm to resolve conflicts at runtime. Approach. In this article, we present a formalization of the underlying models of Ruby refinements, Groovy categories, Classboxes, and Method Shelters that are scoping extension method solutions in dynamically-typed languages. Knowledge. Our formal framework allows us to objectively compare and analyze the shortcomings of the studied solutions and other different approaches such as MultiJava. In addition, language designers can use our formal framework to determine which mechanism has less risk of *accidental overrides*. Grounding. Our comparison and analysis of existing solutions is grounded because it is based on denotational semantics formalizations. Importance. Extension methods are widely used in programming languages that support them, especially dynamically-typed languages such as Pharo, Ruby or Python. However, without a carefully designed mechanism, this feature can cause insidious hidden bugs or can be voluntarily used to gain access to protected operations, violate encapsulation or break fundamental invariants. [17]

First-Class Undefined Classes for Pharo: From Alternative Designs to a Unified Practical Solution. Loading code inside a Pharo image is a daily concern for a Pharo developer. Nevertheless, several problems may arise at loading time that can prevent the code to load or even worse let the system in an inconsistent state. We focus on the problem of loading code that references a class that does not exist in the system. We discuss the different flavors of this problem, the limitations of the existing Undeclared mechanism and the heterogeneity of Pharo tools to solve it. Then, we propose an unified solution for Pharo that reifies Undefined Classes. Our model of Undefined Classes is the result of an objective selection among different alternatives. We then validate our solution through two cases studies: migrating old code and loading code with circular dependencies. We also present the integration of this solution into Pharo regarding the needed Meta-Object Protocol for Undefined Classes and the required modifications of existing tools. [30]

Run-Fail-Grow: Creating Tailored Object-Oriented Runtimes. Producing a small deployment version of an application is a challenge because static abstractions such as packages cannot anticipate the use of their parts at runtime. Thus, an application often occupies more memory than actually needed. Tailoring is one of the main solutions to this problem i.e., extracting used code units such as classes and methods of an application. However, existing tailoring techniques are mostly based on static type annotations. These techniques cannot efficiently tailor applications in all their extent (e.g., runtime object graphs and metadata) nor be used in the context of dynamically-typed languages. We propose a run-fail-grow technique to tailor applications using their runtime execution. Run-fail-grow launches (a) a reference application containing the original application to tailor and (b) a nurtured application containing only a seed with a minimal set of code units the user wants to ensure in the final application. The nurtured application is executed, failing when it founds missing objects,

classes or methods. On failure, the necessary elements are installed into the nurtured application from the reference one, and the execution resumes. The nurtured application is executed until it finishes, or until the developer explicitly finishes it, for example in the case of a web application. resulting in an object memory (i.e., a heap) with only objects, classes and methods required to execute the application. To validate our approach we implemented a tool based on Virtual Machine modifications, namely Tornado. Tornado succeeds to create very small memory footprint versions of applications e.g., a simple object-oriented heap of 11kb. We show how tailoring works on application code, base and third-party libraries even supporting human interaction with user G. interfaces. These experiments show memory savings ranging from 95% to 99%. [18]

7.4. Dynamic Languages: Debugging

Unanticipated Debugging with Dynamic Layers. To debug running software we need unanticipated adaptation capabilities, especially when systems cannot be stopped, updated and restarted. Adapting such programs at runtime is an extreme solution given the delicate live contexts the debugging activity takes place. We introduce the Dynamic Layer, a construct in which behavioral variations are gathered and activated as a whole set of adaptations. Dimensions of Dynamic Layers activation are reified to allow very fine definitions of layer scopes and a fine grained selection of adapted entities. An experimental implementation with the Pharo language is evaluated through a runtime adaptation example. [25]

New Generation Debuggers. Locating and fixing bugs is a well-known time consuming task. Advanced approaches such as object-centric or back-in-time debuggers have been proposed in the literature, still in many scenarios developers are left alone with primitive tools such as manual breakpoints and execution stepping. We explore several advanced on-line debugging techniques such as advanced breakpoints and on-line execution comparison, that could help developers solve complex debugging scenarios. We analyze the challenges and underlying mechanisms required by these techniques. We present some early but promising prototypes we built on the Pharo programming language. We finally identify future research paths by analyzing existing research and connecting it to the techniques we presented before. [27]

Debugging Cyber-Physical Systems. Cyber-Physical Systems (CPS) integrate sensors and actuators to collect data and control entities in the physical world. Debugging CPS systems is hard due to the time-sensitive nature of a distributed applications combined with the lack of control on the surrounding physical environment. This makes bugs in CPS systems hard to reproduce and thus to fix. In this context, on-line debugging techniques are helpful because the debugger is connected to the device when an exception or crash occurs. We report on our experiences on applying two different on-line debugging techniques for a CPS system: remote debugging using the Pharo remote debugger and our IDRA debugger. In contrast to traditional remote debugging, IDRA allows to on-line debug an application locally in another client machine by reproducing the runtime context where the bug manifested. Our qualitative evaluation shows that IDRA provides almost the same interaction capabilities than Pharo's remote debugger and is less intrusive when performing hot-modifications. Our benchmarks also show that IDRA is significantly faster than the Pharo remote debugger, although it increases the amount of data transferred over the network. [29]

Reflectogram. Reflective facilities in OO languages are used both for implementing language extensions (such as AOP frameworks) and for supporting new programming tools and methodologies (such as object-centric debugging and message-based profiling). Yet controlling the runtime behavior of these reflective facilities introduces several challenges, such as computational overhead, the possibility of meta-recursion and an unclean separation of concerns between base and meta-level. We present five dimensions of meta-level control from related literature that try to remedy these problems. These dimensions are namely: temporal and spatial control, placement control, level control and identity control. We then discuss how these dimensions interact with language semantics in class-based OO languages in terms of: scoping, inheritance and first-class entities. We argue that the reification of the descriptive notion of reflectogram can unify the control of meta-level execution in all these five dimensions while expressing properly the underlying language semantics. We present an extended model for the reification of the reflectogram based on our additional analysis and validate our approach through a new prototype implementation that relies on byte-code instrumentation. Finally, we illustrate our approach through a case study on runtime tracing. [16]

7.5. Dynamic Languages: Virtual Machines

VM Profiler. Code profiling enables a user to know where in an application or function the execution time is spent. The Pharo ecosystem offers several code profilers. However, most of the publicly available profilers (MessageTally, Spy, GadgetProfiler) largely ignore the activity carried out by the virtual machine, thus incurring inaccuracy in the gathered information and missing important information, such as the Just-in-time compiler activity. We describe the motivations and the latest improvements carried out in VMProfiler, a code execution pro-filer hooked into the virtual machine, that performs its analysis by monitoring the virtual machine execution. These improvements address some limitations related to assessing the activity of native functions (resulting from a Just-in-time compiler operation): as of now, VMProfiler provides more detailed profiling reports, showing for native code functions in which bytecode range the execution time is spent. [28]

Sista: Saving Optimized Code in Snapshots for Fast Start-Up. Modern virtual machines for object-oriented languages such as Java HotSpot, Javascript V8 or Python PyPy reach high performance through just-in-time compilation techniques, involving on-the-fly optimization and deoptimization of the executed code. These techniques require a warm-up time for the virtual machine to collect information about the code it executes to be able to generate highly optimized code. This warm-up time required before reaching peak performance can be considerable and problematic. We propose an approach, Sista (Speculative Inlining SmallTalk Architecture) to persist optimized code in a platform-independent representation as part of a snapshot. After explaining the overall approach, we show on a large set of benchmarks that the Sista virtual machine can reach peak performance almost immediately after start-up when using a snapshot where optimized code was persisted. [24]

7.6. Interaction

This work is done in collaboration with team Mjolnir.

Turning Function Calls Into Animations. Animated transitions are an integral part of modern interaction frameworks. With the increasing number of animation scenarios, they have grown in range of animatable features. Yet not all transitions can be smoothed: programming systems limit the flexibility of frameworks for animating new things, and force them to expose low-level details to programmers. We present an ongoing work to provide system-wide animation of objects, by introducing a delay operator. This operator turns setter function calls into animations. It offers a coherent way to express animations across frameworks, and facilitates the animation of new properties. [31]

7.7. Software Engineering for BlockChain and Smart Contracts

Solidity Parsing Using SmaCC: Challenges and Irregularities. Solidity is a language used to implement smart contracts on a blockchain platform. Since its initial conception in 2014, Solidity has evolved into one of the major languages for the Ethereum platform as well as other blockchain technologies. Due to its popularity, there are many tools specifically designed to handle smart contracts written in Solidity. However , there is a lack of tools for Pharo to handle Solidity contracts. Therefore, we implemented a parser using SmaCC to serve as a base for further developing Solidity support in Pharo. We describe the parser creation, the irregularities we found in the Solidity grammar specification, and common practices on how to adapt the grammar to an LR type parser. Our experiences with parsing the Solidity language using SmaCC may help other developers trying to convert similar grammars. [32]

SmartInspect: Smart Contract Inspection. Smart contracts are embedded procedures stored with the data they act upon. Debugging deployed Smart Contracts is a difficult task since once deployed, the code cannot be reexecuted and inspecting a simple attribute is not easily possible because data is encoded. In this technical report, we present SmartInspect to address the lack of inspectability of a deployed contract. Our solution analyses the contract state by using decompilation techniques and a mirror-based architecture to represent the object responsible for interpreting the contract state. SmartInspect allows developers and also end-users of a contract to better visualize and understand the contract stored state without needing to redeploy, nor develop any ad-hoc code. [43]

TACOMA Team

6. New Results

6.1. Smart City and ITS

Participants: Indra Ngurah, Djibrilla Amadou Kountche, Xavier Gilles, Christophe Couturier, Rodrigo Silva, Frédéric Weis, Jean-Marie Bonnin [contact].

The domain of Smart Cities is still young but it is already a huge market which attract number of companies and researchers. It is also multi-fold as the words "smart city" gather multiple meanings. Among them one of the main responsibilities of a city, is to organisation the transportation of goods and people. In intelligent transportation systems (ITS), ICT technologies have been involved to improve planification and more generally efficiency of journeys within the city. We are interested in the next step where efficiency would be improved locally relying on local interactions between vehicles, infrastructure and people (smartphones).

For the future "autonomous" vehicle are now in the spotlight, since a lot of works has been done in recent years in automotive industry as well as in academic research centers. Such unmanned vehicle could strongly impact the organisation of the transportation in our cities. However, due to the lack of a definition of what is an "autonomous" vehicle it remains still difficult to see how these vehicles will interact with their environment (eg. road, smart city, houses, grid, etc"). From augmented perception to fully cooperative automated vehicle, the autonomie cover various realities in terms of interaction the vehicle relies on. The extended perception relies on communication between the vehicle and surrounding roadside equipments. This help the driving system to build and maintain an accurate view of the environment. But at this first stage the vehicle only uses its own perception to make its decisions. At a second stage, it will take benefits of local interaction with other vehicles through car-to-car communications to elaborate a better view of its environment. Such "cooperative autonomy" does not try to reproduce the human behavior anymore, it strongly rely on communication between vehicles and/or with the infrastructure to make decision and to acquire information on the environment. Part of the decision could be centralized (almost everything for an automatic metro) or coordinated by a roadside component. The decision making could even be fully distributed but this put high constraints on the communications. Automated vehicles are just an exemple of smart city automated processes that will have to share information within the surrounding to make their decisions.

We participated in the definition of the distributed architecture that has been adopted by all partners of the SEAS project. The main principles of this architecture have been published and we developed several profs of concept that have been demonstrated in the project consortium. Our partner developed the components of the architecture that has been demonstrated in the final review of the project (in January). The principles of the architecture and data representation has been used to design an open reusable Data Manager in the context of the EkoHub projet. This modular software will be extended to fit the needs of Indra Ngurah and Rodrigo Silva works.

6.2. Convergence middleware for pervasive data

Participants: Yoann Maurel, Jules Desjardin, Paul Couderc [contact].

We are currently working on data driven middleware approaches dedicated to physical objects and smart spaces. We had previous contributions on the topic, where opportunistic collaborations between mobile devices were supported by Linda-like tuple space and IEEE 802.11 radios. However, these were adapted to relatively complex devices and the technological limitation at the time did not allow the full potential of the model. More recently, we investigated distributed storage spread over physical objects or fragments using RFID, enabling complex data to be directly reflected by passive objects (without energy). Yet other radio technologies, such as BLE, are emerging to support close range interactions with very low (or even zero) energy requirements.

Applications such as pervasive games (for ex. Pokemon Go), on the go data sharing, collaborative mobile app are often good candidates for opportunistic or dynamic interaction models. But they are not well supported by existing communication stacks, especially in context involving multiple technologies. Technological heterogeneity is not hidden, and high level properties associated with the interactions, such as proximity/range, or mobility-related parameters (speed, discovery latency) have to be addressed in an ad hoc manner. We think that a good way to solve these issues is to offer an abstract interaction model that could be mapped over the common proximity communication technologies, in a similar way as MOM (Message Oriented Middleware) such as MQTT abstract communications in many IoT and pervasive computing scenarios. However, they typically requires IP level communication, which far beyond the capabilities of ultra low energy proximity communication such as RFID and BLE. Moreover, they often rely on a coordinator node that is not adapted in highly dynamic context involving ephemeral communications and mobile nodes.

We started the implementation of an associative memory mechanism over BLE, as it is a common ground that can be shared with passive or semi passive communications (RFID, NFC). Such mechanism, although relatively low level, is still a very useful building block for opportunistic applications: it enables opportunistic data storage/sharing and signaling/synchronization (in space in particular). This approach is fully in line with more general trend developed in the team to build "smart" systems leveraging local resources and data oriented mediation. We have started validation work with a few applications, in particular regarding energy aspects and scalability with respect to the communication load. This should lead to publishing on both infrastructure and application level aspects of the approach.

6.3. Modeling activities and forecasting energy consumption and production to promote the use of self-produced electricity from renewable sources

Participants: Alexandre Rio, Yoann Maurel [contact].

This work began in 2017 and is carried out as part of a broader collaboration between Tacoma, the Diverse Team and OKWind, a company specialized in the production of renewable sources of energy. OKWind proposes to deploy self-production units directly where the consumption is done. It has developed expertise in vertical-axis wind turbines, photovoltaic trackers, and heat pump. This project aims at building a system that optimizes the use of different sources of renewable energy, choosing the most suitable source for the current demand and anticipating future needs. The goal is to favor the consumption of locally produced electricity and to maximize the autonomy of the equipped sites so as to reduce the infrastructure needed to distribute electricity, to set energy cost, and to reduce the ecological impact of energy consumption.

Modeling and forecasting production and consumption of a site is hard and raises several issues: how to precisely assess the consumption and production of energy on a given site with changing conditions ? How to adequately size energy sources and energy storage (wind turbine, solar panel and batteries) ? And what methods to use to optimize consumption and, whenever possible, act on installations and activities to reduce energy costs ? We aim to propose tools to predict the consumption of a site based on estimation and previous observation, monitor the site in real time and forecast evolution. We propose to build a DSL describing consumption and production processes, and a system providing recommendations based on the derived model at runtime.

The problem of forecasting is known from both a production and consumption point of view. OKWind has developed tools to predict the production of their renewable sources - the same goes for batteries - and a lot of theoretical work has been done on consumption in the literature. In our view little has been done to precisely model activities, their energy consumption and the associated variability. Indeed most of the current approaches are concerned with either large-scale forecasting for the Smart Grid, are based on coarse grain data (total energy consumption of the site), or focuses on modeling specific appliance without describing how and when they are used.

This is paradoxical considering that companies have spent a lot of time modeling their activities from a logistic point of view. Intuitively, the predictable and seasonal nature of a company's activities would allow building activity schedulers that favor the consumption of certain energy sources (the cleanest or least expensive one for

instance). The development of a DSL to describe the relationships between activities, their planning, and the production and environmental factors would make possible to simulate a given site at a given location, to make assumptions on sizing, and would be a basis to forecast energy consumption so as to provide recommendations for the organization of activities.

We already have developed part of this DSL that simulates activities and production. In particular, it is capable of simulating consumption and production over a given period based on available environmental data. This tool is in the experimentation phase. In particular, we are collecting information on several sites to measure the consumption of various activities.

6.4. Sharing knowledge and access-control

Participants: Adrien Capaine, Yasmina Andaloussi, Frédéric Weis, Yoann Maurel [contact].

Smart spaces (Smart-city, home, building, etc.) are complex environments made up of resources (cars, smartphones, electronic equipment, applications, servers, flows, etc.) that cooperate to provide a wide range of services to a wide range of users. They are by nature extremely fluctuating, heterogeneous, and unpredictable. In addition, applications are often mobile and have to migrate or are offered by mobile platforms such as smartphones or vehicles.

To be relevant, applications must be able to adapt to users by understanding their environment and anticipating its evolutions. They are therefore based, explicitly or implicitly, on a representation of their surrounding environment based on available data provided by sensors, humans, objects and applications when available. The accuracy of the adaptations made by the applications depends on the precision of this representation. Building and maintaining such knowledge is resource-intensive in terms of network exchanges, computing time and incidentally energy consumption. It is, therefore, crucial to find ways to improve this process. In practice, many applications build their own models without sharing them or delegating calculations to remote services, which is not optimal because many processes are redundant. A huge improvement would be to find mechanisms that allows sharing the information so as to reduce as much as possible the treatments necessary to obtain it.

However, it seems extremely complex to provide a global, complete and unified view of the environment that reflects the applications' concerns. If it were possible, such a single representation would by nature be incomplete or subjective. Our solution should be applicable to nowadays devices and applications with little adjustments to the underlying architectures. It should then be flexible enough to deal with the lack of standards in the domain without imposing architectural choices. Such lack of standard is very common in IT and mainly due to well-known factors: (1) for technical reasons, developers tend to think that their "standard" is better suited for their current use-case, or/and (2) for commercial reasons companies want to keep a closed siloed system to capture their users, or/and (3) because the domain is still new and evolving and no standard as emerged yet, or/and finally (4) because the problem is too complex to be standardized and most proposed standards tend to be bloated and hard to use. The IoT domain suffers from all of these impediments and solution targeting mid-term application have to take these factors into accounts. Many IoT applications are still organized in silos of information. This leads to the deployment of sensors with similar functions and redundant pieces of software providing exactly the same service. Many frameworks or ontologies have been developed in the field to provide a solution to this problem but their implementation depends on the goodwill of the companies who do not always see their interest in losing part of the control of their application and data. To be largely accepted, solutions should let companies decide what information to share and when with little impact on their current infrastructure.

We want to be able to develop collaborative mechanisms that allow applications to share some of their information about the immediate surrounding environment with their counterparts. The idea is to allow the construction of shared representations between groups of applications that manipulate the same concepts so that each group can construct a subjective and complete representation of the environment that corresponds to its concerns. In this context, we want to offer applications mechanisms allowing them to leave information about their environment by associating them directly with the flows, data, services and objects handled. This

information will be stored by the environment so that it will be possible for the application to retrieve it and for its peers to access it. From a logical point of view, applications will have the illusion of annotating objects directly; we make no assumptions about where this information will be stored, which will depend on the characteristics of the environment or the sharing solution chosen. Data should be stored as close as possible to the environments they qualify for reasons of performance, confidentiality and autonomy. To experience that idea, we have developed:

- Matriona, a globally distributed framework developed on top of OSGi. This project has been described in more details in the previous activity report. It is meant to be a global framework for exposing devices as REST-like resources. Resources functionalities can be extended through the mean of decorators. The system also provides access mechanisms. The main interest of Matriona with regards to the information enrichment is its ability to support the dynamic extension of resource meta-information by application and to provide means to share this meta-information with others. It implements the concept of groups of interest with access control on meta-information. The concept described in Matriona are in the process to be published.
- Little Thumb Base (LithBase) is an independent knowledge base that provides the same enrichment capabilities than Matriona but imposes fewer constraints on the architecture of applications. It is a shared database implemented on simple low power nodes (esp32) that are cheap to deploy, flash and use. The idea behind LithBase is to decouple the storage from the framework and to provide a standard mechanism to share information. Ultimately we want to use its capabilities to implement a registry in the manner of Consul with meta-information enrichment and sharing mechanisms. By focussing only on the discovery mechanism and information sharing, LithBase imposes fewer constraints on applications and comply more with the goal of being ready to use in existing applications. This is still a work in progress. This solution also raises the issue of trust and control over access to this information. It is indeed necessary for applications to be able to determine the source of the additional information and to determine who will have access to the information they add. We have also been experimenting with access control mechanism that is implemented by LithBase. We are currently using elliptic cryptography to allow private information sharing between groups. Ultimately the goal of this project is to produce a coordinating object that implements generic mechanisms favouring opportunistic behaviours of surrounding applications.

AGORA Team

7. New Results

7.1. Wireless network deployment

Participants: Ahmed Boubrima, Rodrigue Domga Komguem, Leo Le Taro, Jad Oueis, Walid Bechkit, Khaled Boussetta, Hervé Rivano, Razvan Stanica, Fabrice Valois.

7.1.1. Deployment of Wireless Sensor Networks for Pollution Monitoring

Air pollution has become a major issue of modern megalopolis because of industrial emissions and increasing urbanization along with traffic jams and heating/cooling of buildings. Monitoring urban air quality is therefore required by municipalities and by the civil society. Current monitoring systems rely on reference sensing stations that are precise but massive, costly and therefore seldom. In our work, we focus on an alternative or complementary approach, with a network of low cost and autonomic wireless sensors, aiming at a finer spatiotemporal granularity of sensing. Generic deployment models of the literature are not adapted to the stochastic nature of pollution sensing.

In this sense, in [2], our main contribution is to design integer linear programming models that compute sensor deployments capturing both the coverage of pollution under time-varying weather conditions and the connectivity of the infrastructure. We evaluate our deployment models on a real data set of Greater London. We analyze the performance of the proposed models and show that our joint coverage and connectivity formulation is tight and compact, with a reasonable enough execution time. We also conduct extensive simulations to derive engineering insights for effective deployments of air pollution sensors in an urban environment.

Unlike most of the existing methods, which rely on simple and generic detection models, our approach is based on the spatial analysis of pollution data, allowing to take into account the nature of the pollution phenomenon. As proof of concept, we apply our approach on real world data, namely the Paris pollution data, which was recorded in March 2014 [7]. In this paper, we consider citywide wireless sensor networks and tackle the minimum-cost node positioning issue for air pollution monitoring. We propose an efficient approach that aims to find optimal sensors and sinks locations while ensuring air pollution coverage and network connectivity.

Mobile wireless sensor networks can also be used for monitoring air pollution, where the aim is usually to generate accurate pollution maps in real time. The generation of pollution maps can be performed using either sensor measurements or physical models which simulate the phenomenon of pollution dispersion. The combination of these two information sources, known as data assimilation, makes it possible to better monitor air pollution by correcting the simulations of physical models while relying on sensor measurements. The quality of data assimilation mainly depends on the number of measurements and their locations. A careful deployment of nodes is therefore necessary in order to get better pollution maps. In an ongoing work [30], we tackle the placement problem of pollution sensors and design a mixed integer programming model allowing to maximize the assimilation quality while ensuring the connectivity of the network. We perform some simulations on a dataset of the city of Lyon in order to show the effectiveness of our model regarding the quality of pollution coverage.

For an air pollution monitoring system deployment to be relevant relative to urban air quality aspects, we are concerned with maintaining the system properties over time. Indeed, one of the major drawbacks of cheap sensors is their drift: chemical properties degrade over time and alter the measurement accuracy. We challenge this issue by designing distributed, online recalibration procedures. In [16], we present a simulation framework modelling a mobile wireless sensor network (WSN) and we assess the system's measurement confidence using trust propagation paradigms. As WSN calibrations translate to information exchange between sensors, we also study means of limiting the number of such transmissions by skipping the calibrations deemed least profitable to the system.

7.1.2. Wireless Sensor Networks with Linear Topology

In wireless sensor networks with linear topology, knowing the physical order in which nodes are deployed is useful not only for the target application, but also to some network services, like routing or data aggregation. Considering the limited resources of sensor nodes, the design of autonomous protocols to find this order is a challenging topic.

In [9], we propose a distributed and iterative centroid-based algorithm to address this problem. At each iteration, the algorithm selects two virtual anchors and finds the order of a subset of nodes, placed between these two anchors. The proposed algorithm requires local node connectivity knowledge and the identifier of the first sensor node of the network, which is the only manually configured parameter. This solution, scalable and lightweight from the deployment and maintenance point of view, is shown to be robust to connectivity degradation, correctly ordering more than 95% of the nodes, even under very low connectivity conditions.

7.1.3. Function Placement in Public Safety Networks

In response to the growing demand in the public safety community for broadband communication systems, LTE is currently being adopted as the base technology for next generation public safety networks. In parallel, notable efforts are being made by the 3GPP to enhance the LTE standard in order to offer public safety oriented services. In the recent Release 13, the Isolated E-UTRAN Operation for Public Safety (IOPS) concept was introduced. IOPS aims at maintaining a level of communication between public safety users, offering them local mission-critical services even when the backhaul connectivity to the core network is not fully functional. Isolated operation is usually needed in mission-critical situations, when the infrastructure is damaged or completely destroyed, and in out of coverage areas. In [6], we present a detailed technical overview on the IOPS specifications, and then identify several research prospects and development perspectives opened up by IOPS.

An isolated base station is a base station having no connection to a traditional core network. To provide services to users, an isolated base station is co-located with an entity providing the same functionalities as the traditional core network, referred to as Local EPC. In order to cover wider areas, several base stations are interconnected, forming a network that should be served by a single Local EPC. In [20], [24], we tackle the Local EPC placement problem in the network, to determine with which of the base stations the Local EPC must be co-located. We propose a novel centrality metric, flow centrality, which measures the capacity of a node to receive the total amount of flows in the network. We show that co-locating the Local EPC with the base station having the maximum flow centrality maximizes the total amount of traffic the Local EPC can receive from all base stations, under certain capacity and load distribution constraints. We compare the flow centrality to other state of the art centrality metrics, and emphasize its advantages.

7.1.4. User Association in Public Safety Oriented Mobile Networks

In many disaster scenarios, communication infrastructure fails to provide network services for both civilians and first responders. One solution is to have rapidly deployable mobile networks formed by interconnected base stations, that are easy to move, deploy, and configure. Such public safety-oriented networks are different from classical mobile networks in terms of scale, deployment, and architecture.

In this context, we revisit the user association problem [21], for two main reasons. First, the backhaul, formed by the links interconnecting the base stations, must be accounted for when deciding on the association, since it may present a bottleneck with its limited bandwidth. Second, the mission-critical nature of the traffic imposes strict guaranteed bit rate constraints, that must be respected when associating users. Therefore, we propose a network-aware optimal association that minimizes the bandwidth consumption on the backhaul, while still respecting the stringent performance requirements.

7.2. Wireless data collection

Participants: Yosra Bahri Zguira, Alexis Duque, Junaid Ahmed Khan, Abdoul-Aziz Mbacké, Romain Pujol, Hervé Rivano, Razvan Stanica, Fabrice Valois.

7.2.1. Smart Parking Systems

Considering the increase of urban population and traffic congestion, smart parking is always a strategic issue to work on, not only in the research field but also from economic interests. Thanks to information and communication technology evolution, drivers can more efficiently find satisfying parking spaces with smart parking services. The existing and ongoing works on smart parking are complicated and transdisciplinary. While deploying a smart parking system, cities, as well as urban engineers, need to spend a very long time to survey and inspect all the possibilities. Moreover, many varied works involve multiple disciplines, which are closely linked and inseparable.

To give a clear overview, we introduce a smart parking ecosystem and propose a comprehensive and thoughtful classification by identifying their functionalities and problematic focuses [5]. We go through the literature over the period of 2000-2016 on parking solutions as they were applied to smart parking development and evolution, and propose three macro-themes: information collection, system deployment, and service dissemination. In each macro-theme, we explain and synthesize the main methodologies used in the existing works and summarize their common goals and visions to solve current parking difficulties. Lastly, we give our engineering insights and show some challenges and open issues.

7.2.2. Data Offloading

Mobile users in an urban environment access content on the Internet from different locations. It is challenging for the current service providers to cope with the increasing content demand from a large number of collocated mobile users. In-network caching to offload content at nodes closer to users alleviates the issue, though efficient cache management is required to find out who should cache what, when and where in an urban environment, given nodes limited computing, communication and caching resources. To address this [14], we first define a novel relation between content popularity and availability in the network and investigate a node eligibility to cache content based on its urban reachability. We then allow nodes to self-organize into mobile fogs to increase the distributed cache and maximize content availability in a cost-effective manner. However, to cater rational nodes, we propose a coalition game for the nodes to offer a maximum virtual cache assuming a monetary reward is paid to them by the service/content provider. Nodes are allowed to merge into different spatio-temporal coalitions in order to increase the distributed cache size at the network edge. Results obtained through simulations using realistic urban mobility trace validate the performance of our caching system showing a ratio of 60 - 85% of cache hits compared to the 30 - 40% obtained by the existing schemes and 10% in case of no coalition.

Another option for data offloading is represented by vehicular traffic. With over 300 billion vehicle trips made in the USA and 64 billion in France per year, network operators have the opportunity to utilize the existing road and highway network as an alternative data network to offload large amounts of delay-tolerant traffic. To enable the road network as a large-capacity transmission system, we exploit the existing mobility of vehicles equipped with wireless and storage capacities together with a collection of offloading spots [1]. An offloading spot is a data storage equipment located where vehicles usually park. Data is transloaded from a conventional data network to the closest offloading spot and then shipped by vehicles along their line of travel. The subsequent offloading spots act as data relay boxes where vehicles can drop off data for later pickups by other vehicles, depending on their direction of travel. The main challenges of this offloading system are how to compute the road path matching the performance requirements of a data transfer and how to configure the sequence of offloading spots involved in the transfer. We propose a scalable and adaptive centralized architecture built on SDN that maximizes the utilization of the flow of vehicles connecting consecutive offloading spots. We simulate the performance of our system using real roads traffic counts for France. Results show that the centralized controlled offloading architecture can achieve an efficient and fair allocation of concurrent data transfers between major cities in France.

7.2.3. Hybrid Short/Long Range Networks

Despite the success of dedicated IoT networks, such as Sigfox or LoRa, several use cases can not be accommodated by these new technologies, mainly because of capacity constraints. For example, mobile sensing and proximity-based applications require smart devices to find other nodes in vicinity, though it is

challenging for a device to find neighbors in an energy efficient manner, while also running on low duty cycles.

Neighbor discovery schemes allow nodes to follow a schedule to become active and send beacons or listen for other active nodes in order to discover each other with a bounded latency. However, a trade-off exists between the energy consumption and the time a node takes to discover neighbors using a given activity schedule. Moreover, energy consumption is not the only bottleneck, as theoretically perfect schedules can result in discovery failures in a real environment. In [12], we provide an in-depth study on neighbor discovery, by first defining the relation between energy efficiency, discovery latency and the fraction of discovered neighbors. We evaluate existing mechanisms using extensive simulations for up to 100 nodes and testbed implementations for up to 15 nodes, with no synchronization between nodes and using duty cycles as low as 1% and 5%. Moreover, the literature assumes that multiple nodes active simultaneously always result in neighbor discovery, which is not true in practice as this can lead to collisions between the transmitted messages. Our findings reveal such scalability issues in existing schemes, where discovery fails because of collisions between beacons from multiple nodes active at the same time. Therefore, we show that energy efficient discovery schemes do not necessarily result in successful discovery of all neighbors, even when the activity schedules are computed in a deterministic manner.

A second use-case requiring a combination of long range and short range communications is related to intelligent transportation systems. As a matter of fact, communication is essential to the coordination of public transport systems. Nowadays, cities are facing an increasing number of bikes used by citizens therefore the need of monitoring and managing their traffic becomes crucial. Public bike sharing system has been introduced as an urban transportation system that can collect data from mobile devices. In this context, we introduce IoB-DTN [29], a protocol based on the Delay/Disruption Tolerant Network (DTN) paradigm adapted for an IoT-like applications running on bike sharing system based sensor network. We present simulation results obtained by evaluating the Binary Spray and Wait inspired variant of IoB-DTN with four buffer management policies and by comparing three variants of IoB-DTN by varying the number of packet copies sprayed in the network.

7.2.4. Visible Light Communications in IoT Networks

With the increasing consumer demand for smart objects, Visible Light Communications (VLC), and especially LED-to-Camera communication, appears as a low-cost alternative to radio to make any conventional device smart. Since LEDs are already on most electronics devices, that is achieved at the cost of negligible hardware modifications. However, as these LEDs are very different from the widely studied ceiling ones, several challenges need to be addressed to make this happen. In our work [31], we propose a line of sight bi-directional communication system between an ordinary LED and an off-the-shelf smartphone. We designed a cheap multi sensors device as a proof of concept of a near communication module for the IoT.

Among the issues we observed experimenting with this platform, we note the constrained physical layer data unit (PHY-SDU) length that complicates the use of coding strategies to cope with bits or packets erasure. To break this limitation, we present SeedLight [8], a coding scheme designed to face the inherent packet losses and enhance line-of-sight LED-to-Camera communication goodput. SeedLight leverages random linear coding to provide an efficient redundancy mechanism that works even on PHY-SDU of tens of bits. The key idea of SeedLight is to reduce the code overhead by replacing the usual coding coefficients by a seed. Since this work addresses IoT devices with low computational resources, SeedLight encoding algorithm complexity remains low. We develop an implementation of SeedLight on a low-cost MCU and a smartphone to evaluate both the communication and algorithmic performances. Experimental results show that SeedLight introduces a negligible overhead and can be implemented even on the cheapest MCU, such as the ones used in many IoT devices. The achievable goodput can be up to 2.5kbps, while the gain compared to a trivial retransmissions scheme is up to 100%.

To ease the evaluation of VLC systems, we present CamComSim [28], the first simulator for development and rapid prototyping of LED-to-Camera communication systems. Our event driven simulator relies on a standalone Java application that is easily extensible through a set of interfaces. A range of low and high-level parameters, such as the camera characteristics, the PHY-SDU size, or the redundancy mechanism can

be chosen. CamComSim uses empirically validated models for the LED-to-Camera channel and the broadcast protocols, configurable with a finely grained precision. To validate CamComSim implementation and accuracy, we use the previously discussed testbed, based on a color LED and a smartphone, and compare the performance reached by the testbed with the results given by our simulator. We illustrate with a real use case the full usage of CamComSim, tuning a broadcast protocol that implements the transmission of 1 kbyte of information. The results highlight that our simulator is very precise and predicts the performance of a real LED-to-Camera system with less than 10% of error in most cases.

7.2.5. Data Collection with RFID Devices

The popularization of Radio Frequency Identification (RFID) systems has conducted to large deployments of RFID solutions in various areas under different criteria. However, such deployments, specially in dense environments, can be subject to RFID collisions which in turn affect the quality of readings. In [17], [18], we propose two distributed and efficient solutions for dense mobile deployments of RFID systems. mDEFAR is an adaptation of a previous work highly performing in terms of collisions reduction, efficiency and fairness in dense static deployments. CORA is more of a locally mutual solution where each reader relies on its neighborhood to enable itself or not. Using a beaconing mechanism, each reader is able to identify potential (non-)colliding neighbors in a running frame and as such chooses to read or not. Performance evaluation shows high performance in terms of coverage delay for both proposals quickly achieving 100% coverage depending on the considered use case while always maintaining consistent efficiency levels above 70%. Compared to GDRA, our solutions proved to be better suited for highly dense and mobile environments, offering both higher throughput and efficiency. The results reveal that depending on the application considered, choosing either mDEFAR or CORA helps improve efficiency and coverage delay.

RFID solutions encounter two main issues: the first one is inherent to the technology itself which is readers collisions, the second one being the gathering of read data up to a base station, potentially in a multihop fashion. While the first one has been a main research subject in the late years, the second one has not been investigated for the sole purpose of RFID, but rather for wireless adhoc networks. This multihop tag information collection must be done in regards of the application requirements but it should also care for the deployment strategy of readers to take advantage of their relative positions, coverage, reading activity and deployment density to avoid interfering between tag reading and data forwarding. To the best of our knowledge, the issue for a joint scheduling between tag reading and forwarding has never been investigated so far in the literature, although important. In [17], we propose two new distributed, cross-layer solutions meant for the reduction of collisions and better efficiency of the RFID system, but also serving as a routing solution towards a base station. Simulations show high levels of throughput while not lowering on the fairness on medium access staying above 85% in the highest deployment density with up to 500 readers, also providing a 90% increase in data rate.

7.3. Network data exploitation

Participants: Panagiota Katsikouli, Elli Zavou, Stéphane D'Alu, Hervé Rivano, Razvan Stanica.

7.3.1. Spatio-temporal Characterization of Mobile Data Traffic

Mobile traffic data collected by network operators is a rich source of information about human habits, and its analysis provides insights relevant to many fields, including urbanism, transportation, sociology and networking. Urban landscapes present a variety of socio-topological environments that are associated to diverse human activities. As the latter affect the way individuals connect with each other, a bound exists between the urban tissue and the mobile communication demand. In [3], we investigate the heterogeneous patterns emerging in the mobile communication activity recorded within metropolitan regions. To that end, we introduce an original technique to identify classes of mobile traffic signatures that are distinctive of different urban fabrics. Our proposed technique outperforms previous approaches when confronted to ground-truth information, and allows characterizing the mobile demand in greater detail than that attained in the literature to date. We apply our technique to extensive real-world data collected by major mobile operators in ten cities. Results unveil the diversity of baseline communication activities across countries, but also evidence

the existence of a number of mobile traffic signatures that are common to all studied areas and specific to particular land uses.

Similarly to mobile phone data, GPS traces of vehicles convey information on transportation demand and human activities that can be related to the land use of the neighborhood where they take place. In [10], we investigate the land use patterns that emerge when studying simultaneously GPS traces of probe vehicles and mobile phone data collected by network providers. To this end, we extend previous definitions of mobile phone traffic signatures for land use detection, so as to incorporate additional information on human presence and mobility conveyed by GPS traces of vehicles. Leveraging these extended signatures, we exploit an unsupervised learning technique to identify classes of signatures that are distinctive of different land use. We apply our technique to real-world data collected in French and Italian cities. Results unveil the existence of signatures that are common to all studied areas and specific to particular land uses. The combined use of mobile phone data and GPS traces outperforms previous approaches when confronted to ground-truth information, and allows characterizing land use in greater detail than in the literature to date.

The spatial and temporal profiles of mobile phone traffic can be studied simultaneously. In [11], we present an original approach to infer both spatial and temporal structures hidden in the mobile demand, via a first-time tailoring of Exploratory Factor Analysis (EFA) techniques to the context of mobile traffic datasets. Casting our approach to the time or space dimensions of such datasets allows solving different problems in mobile traffic analysis, i.e., network activity profiling and land use detection, respectively. Tests with real-world mobile traffic datasets show that, in both its variants above, the proposed approach (i) yields results whose quality matches or exceeds that of state-of-the-art solutions, and (ii) provides additional joint spatiotemporal knowledge that is critical to result interpretation.

7.3.2. Using Mobile Phone Data in Cognitive Networking

In the next few years, mobile networks will undergo significant evolutions in order to accommodate the ever-growing load generated by increasingly pervasive smartphones and connected objects. Among those evolutions, cognitive networking upholds a more dynamic management of network resources that adapts to the significant spatiotemporal fluctuations of the mobile demand. Cognitive networking techniques root in the capability of mining large amounts of mobile traffic data collected in the network, so as to understand the current resource utilization in an automated manner. In [4], we take a first step towards cellular cognitive networks by proposing a framework that analyzes mobile operator data, builds profiles of the typical demand, and identifies unusual situations in network-wide usages. We evaluate our framework on two real-world mobile traffic datasets, and show how it extracts from these a limited number of meaningful mobile demand profiles. In addition, the proposed framework singles out a large number of outlying behaviors in both case studies, which are mapped to social events or technical issues in the network.

7.3.3. Study of Wi-Fi Localization from Crowdsourced Datasets

The wide adoption of mobile devices has created unprecedented opportunities to collect mobility traces and make them available for the research community to conduct interdisciplinary research. However, mobility traces available in the public domain are usually restricted to traces resulting from a single sensor (e.g., either GPS, GSM or WiFi). In [26], we present the PRIVA'MOV dataset, a novel dataset collected in the city of Lyon, France on which user mobility has been collected using multiple sensors. More precisely, this dataset contains mobility traces of about 100 persons including university students, staff and their family members over 15 months collected through the GPS, WiFi, GSM, and accelerometer sensors. We provide both a quantitative and a preliminary qualitative analysis of this dataset. Specifically, we report the number of visited points of interests, GSM antennas and WiFi hotspots and their distribution across the various users. We finally analyse the uniqueness of human mobility by considering the various sensors.

Thanks to this collected data, it is possible to combine information from several probes. A very common use case is the collection of network scans with location to help the localisation feature of these devices. Nevertheless, most users are not aware of this spying. The collected data might represent infringements of privacy. One possible solution to keep gathering these data while maintaining privacy would consist in device-to-device communications in order to break the links between data and users. In [22], we propose an approach

to test the feasibility of such a system. We collected data from mobile users to combine location and network scans data. With this data, we test the accuracy level we can reach while using Wi-Fi localisation. We analyse how a new measure should be pushed and how many scans should be realised to provide location-based Wi-Fi. We analyse the minimal dataset to cover the set of locations covered by users and prove that a multiuser gathering system can benefit the users.

COATI Project-Team

7. New Results

7.1. Network Design and Management

Participants: Christelle Caillouet, David Coudert, Frédéric Giroire, Frédéric Havet, Nicolas Huin, Joanna Moulierac, Nicolas Nisse, Stéphane Pérennes, Andrea Tomassilli.

Network design is a very wide subject which concerns all kinds of networks. In telecommunications, networks can be either physical (backbone, access, wireless, ...) or virtual (logical). The objective is to design a network able to route a (given, estimated, dynamic, ...) traffic under some constraints (e.g. capacity) and with some quality-of-service (QoS) requirements. Usually the traffic is expressed as a family of requests with parameters attached to them. In order to satisfy these requests, we need to find one (or many) paths between their end nodes. The set of paths is chosen according to the technology, the protocol or the QoS constraints.

We mainly focus on four topics: Firstly, we study the new network paradigms, Software-Defined Networks (SDN) and Network Function Virtualization (NFV). On the contrary to legacy networks, in SDN, a centralized controller is in charge of the control plane and takes the routing decisions for the switches and routers based on the network conditions. This new technology brings new constraints and therefore new algorithmic problems such as the problem of limited space in the switches to store the forwarding rules. We then tackle the problem of placement of virtualized resources. We validated our algorithms on a real SDN platform [0]. Secondly, we consider different scenarios regarding wireless networks, in particular, wireless backhaul networks, linear access networks for transportation systems, and connected Unmanned Aerial Vehicules (UAVs). Third, we tackle routing in the Internet. Last, we study live streaming in distributed systems.

7.1.1. Software Defined Networks (SDN)

Software-defined Networks (SDN), in particular OpenFlow, is a new networking paradigm enabling innovation through network programmability. SDN is gaining momentum with the support of major manufacturers. Over past few years, many applications have been built using SDN such as server load balancing, virtual-machine migration, traffic engineering and access control.

7.1.1.1. Minnie: an SDN World with Few Compressed Forwarding Rules

While SDN brings flexibility to the management of flows within the data center fabric, this flexibility comes at the cost of smaller routing table capacities. Indeed, the Ternary Content-Addressable Memory (TCAM) needed by SDN devices has smaller capacities than CAMs used in legacy hardware. Also, we investigate in [37] compression techniques to maximize the utility of SDN switches forwarding tables. We validate our algorithm, called MINNIE, with intensive simulations for well-known data center topologies, to study its efficiency and compression ratio for a large number of forwarding rules. Our results indicate that MINNIE scales well, being able to deal with around a million of different flows with less than 1000 forwarding entries per SDN switch, requiring negligible computation time.

To assess the operational viability of MINNIE in real networks, we deployed a testbed able to emulate a $k = 4$ Fat-Tree data center topology. We demonstrate on the one hand, that even with a small number of clients, the limit in terms of number of rules is reached if no compression is performed, increasing the delay of new incoming flows. MINNIE, on the other hand, reduces drastically the number of rules that need to be stored, with no packet losses, nor detectable extra delays if routing lookups are done in the Application-Specific Integrated Circuits (ASICs).

Hence, both simulations and experimental results suggest that MINNIE can be safely deployed in real networks, providing compression ratios between 70% and 99%.

[0] Testbed with SDN hardware, in particular a switch HP 5412 with 96 ports, hosted at I3S laboratory. A complete fat-tree architecture with 16 servers can be built on the testbed.

7.1.1.2. Bringing Energy Aware Routing closer to Reality with SDN Hybrid Networks

Energy aware routing aims at reducing the energy consumption of ISP networks. The idea is to adapt routing to the traffic load in order to turn off some hardware. However, it implies to make dynamic changes to routing configurations which is almost impossible with legacy protocols. The SDN paradigm bears the promise of allowing a dynamic optimization with its centralized controller.

In [49], [59], we propose SENAtoR, an algorithm to enable energy aware routing in a scenario of progressive migration from legacy to SDN hardware. Since in real life, turning off network equipments is a delicate task as it can lead to packet losses, SENAtoR provides also several features to safely enable energy saving services: tunneling for fast rerouting, smooth node disabling and detection of both traffic spikes and link failures.

We validate our solution by extensive simulations and by experimentation. We show that MINNIE can be progressively deployed in a network using the SDN paradigm. It allows to reduce the energy consumption of ISP networks by 5 to 35% depending on the penetration of SDN hardware, while diminishing the packet loss rate compared to legacy protocols.

7.1.1.3. Network Function Virtualization (NFV) and Service Function Chains

Network Function Virtualization (NFV) is a promising network architecture concept to reduce operational costs. In legacy networks, network functions, such as firewall or TCP optimization, are performed by specific hardware. In networks enabling NFV coupled with the Software Defined Network (SDN) paradigm, network functions can be implemented dynamically on generic hardware. The challenge is then to efficiently provision the service chain requests, while finding the best compromise between the bandwidth requirements, the number of locations for hosting Virtual Network Functions (VNFs), and the number of chain occurrences.

In [48], we propose two ILP (Integer Linear Programming) models for routing service chain requests, one of them with a decomposition modeling. We conduct extensive numerical experiments, and show we can solve exactly the routing of service chain requests in a few minutes for networks with up to 50 nodes, and traffic requests between all pairs of nodes. We investigate the best compromise between the bandwidth requirements and the number of VNF nodes.

In [50], we study how to use NFV coupled with SDN to improve the energy efficiency of networks. We consider a setting in which a flow has to go through a Service Function Chain, that is several network functions in a specific order. We propose a decomposition model that relies on lightpath configuration to solve the problem. We show that virtualization allows to obtain between 30% to 55% of energy savings for networks of different sizes.

7.1.2. Wireless networks

We study optimization problems on various kinds of wireless networks.

7.1.2.1. Computing and maximizing the exact reliability of wireless backhaul networks

The reliability of a fixed wireless backhaul network is the probability that the network can meet all the communication requirements considering the uncertainty (e.g., due to weather) in the maximum capacity of each link. We provide in [45] an algorithm to compute the exact reliability of a backhaul network, given a discrete probability distribution on the possible capacities available at each link. The algorithm computes a conditional probability tree, where at each leaf in the tree a valid routing for the network is evaluated. Any such tree provides bounds on the reliability, and the algorithm improves these bounds by branching in the tree. We also consider the problem of determining the topology and configuration of a backhaul network that maximizes reliability subject to a limited budget. We provide an algorithm that exploits properties of the conditional probability tree used to calculate reliability of a given network design, and we evaluate its computational efficiency.

7.1.2.2. Analysis of the Failure Tolerance of Linear Access Networks

In [28], we study the disconnection of a moving vehicle from a linear access network composed by cheap WiFi Access Points in the context of the telecommuting in massive transportation systems. In concrete terms, we analyze the probability for a user to experience a disconnection longer than a given time interval (t^*)

such that all on-going communications between the vehicle and the infrastructure network are disrupted. We provide an approximation formula considering two scenarios (intercity bus and train). We then carry out a sensitivity analysis and supply a guide for operators when choosing the parameters of the networks. Last, we show that such systems are viable, as they attain a very low probability of long disconnections with a very low maintenance cost.

7.1.2.3. Efficient Deployment of Connected Unmanned Aerial Vehicles for Optimal Target Coverage

Anytime and anywhere network access can be provided by Unmanned Aerial Vehicles (UAV) with air-to-ground and air-to-air communications using directional antennas for targets located on the ground. Deploying these Unmanned Aerial Vehicles to cover targets is a complex problem since each target should be covered, while minimizing (i) the deployment cost and (ii) the UAV altitudes to ensure good communication quality. We also consider connectivity between the UAVs and a base station in order to collect and send information to the targets, which is not considered in many similar studies. In [40], we provide an efficient optimal program to solve this problem and show the trade-off analysis due to conflicting objectives. We propose a fair trade-off optimal solution and also evaluate the cost of adding connectivity to the UAV deployment.

7.1.3. Routing in the Internet

7.1.3.1. Routing at Large Scale: Advances and Challenges for Complex Networks

A wide range of social, technological and communication systems can be described as complex networks. Scale-free networks are one of the well-known classes of complex networks in which nodes degree follow a power-law distribution. The design of scalable, adaptive and resilient routing schemes in such networks is very challenging. In [38], we present an overview of required routing functionality, categorize the potential design dimensions of routing protocols among existing routing schemes and analyze experimental results and analytical studies performed so far to identify the main trends/trade-offs and draw main conclusions. Besides traditional schemes such as hierarchical/shortest-path path-vector routing, the article pays attention to advances in compact routing and geometric routing since they are known to significantly improve the scalability in terms of memory space. The identified trade-offs and the outcomes of this overview enable more careful conclusions regarding the (in-)suitability of different routing schemes to large-scale complex networks and provide a guideline for future routing research. This article concludes the European Project FP7 STREP EULER (2010-2014).

7.1.3.2. Grid spanners with low forwarding index for energy efficient networks

A routing R of a connected graph G is a collection that contains simple paths connecting every ordered pair of vertices in G. The *edge-forwarding index with respect to* R (or simply the forwarding index with respect to R) $\pi(G,R)$ of G is the maximum number of paths in R passing through any edge of G. The *forwarding index*$\pi(G)$ of G is the minimum $\pi(G,R)$ over all routings R's of G. This parameter has been studied for different graph classes. Motivated by energy efficiency, we look in [30] for different numbers of edges, at the best spanning graphs of a square grid, namely those with a low forwarding index.

7.1.4. Live streaming in distributed systems

Peer to peer networks are an efficient way to carry out video live streaming as the forwarding load is distributed among peers. These systems can be of two types: unstructured and structured. In unstructured overlays, the peers obtain the video in an opportunistic way. The advantage is that such systems handle churn well. However, they are less bandwidth efficient than structured overlays, and the control overhead has a non-negligible impact on the performance. In structured overlays, the diffusion of the video is made via an explicit diffusion tree. The advantage is that the peer bandwidth can be optimally exploited. The drawback is that the departure of peers may break the diffusion tree.

In [29], we propose and analyze a simple local algorithm to balance a tree. In this distributed repair algorithm, each node carries out local operations based on its degree and on the subtree sizes of its children. In a synchronous setting, we first prove that starting from any n-node tree our process converges to a balanced binary tree in $O(n^2)$ rounds. We then describe a more restrictive model, adding a small extra information to each node, under which we adapt our algorithm to converge in $\Theta(n \log n)$ rounds.

In [58], we propose new simple distributed repair protocols for video live streaming structured systems. We show, through simulations with real traces, that structured systems can be very efficient and robust to failures, even for high churn and when peers have very heterogeneous upload bandwidth capabilities.

7.2. Graph Algorithms

Participants: Julien Bensmail, Nathann Cohen, David Coudert, Guillaume Ducoffe, Valentin Garnero, Frédéric Giroire, Frédéric Havet, Fionn Mc Inerney, Nicolas Nisse, Stéphane Pérennes, Rémi Watrigant.

COATI is interested in the algorithmic aspects of Graph Theory. In general we try to find the most efficient algorithms to solve various problems of Graph Theory and telecommunication networks. We use Graph Theory to model various network problems. We study their complexity and then we investigate the structural properties of graphs that make these problems hard or easy.

7.2.1. Complexity of graph problems

We also investigate several graph problems coming from various applications. We mainly consider their complexity in general or particular graph classes. When possible, we present polynomial-time (approximation) algorithms or Fixed Parameter Tractable algorithms.

7.2.1.1. Parameterized complexity of polynomial optimization problems (FPT in P)

Parameterized complexity theory has enabled a refined classification of the difficulty of NP-hard optimization problems on graphs with respect to key structural properties, and so to a better understanding of their true difficulties. More recently, hardness results for problems in P were established under reasonable complexity theoretic assumptions such as: Strong Exponential Time Hypothesis (SETH), 3SUM and All-Pairs Shortest-Paths (APSP). According to these assumptions, many graph theoretic problems do not admit truly subquadratic algorithms, nor even truly subcubic algorithms (Williams and Williams, FOCS 2010 [82] and Abboud *et al.* SODA 2015 [70]). A central technique used to tackle the difficulty of the above mentioned problems is fixed-parameter algorithms for polynomial-time problems with *polynomial dependency* in the fixed parameter (P-FPT). This technique was rigorously formalized by Giannopoulou et al. (IPEC 2015) [75], [76]. Following that, it was continued by Abboud *et al.* (SODA 2016) [71], by Husfeldt (IPEC 2016) [78] and Fomin *et al.* (SODA 2017) [74], using the treewidth as a parameter. Applying this technique to *clique-width*, another important graph parameter, remained to be done.

In [55] we study several graph theoretic problems for which hardness results exist such as *cycle problems* (triangle detection, triangle counting, girth), *distance problems* (diameter, eccentricities, Gromov hyperbolicity, betweenness centrality) and *maximum matching*. We provide hardness results and fully polynomial FPT algorithms, using clique-width and some of its upper-bounds as parameters (split-width, modular-width and P_4-sparseness). We believe that our most important result is an $\mathcal{O}(k^4 \cdot n + m)$-time algorithm for computing a maximum matching where k is either the modular-width or the P_4-sparseness. The latter generalizes many algorithms that have been introduced so far for specific subclasses such as cographs, P_4-lite graphs, P_4-extendible graphs and P_4-tidy graphs. Our algorithms are based on preprocessing methods using modular decomposition, split decomposition and primeval decomposition. Thus they can also be generalized to some graph classes with unbounded clique-width.

7.2.1.2. Finding cut-vertices in the square roots of a graph

The square of a given graph $H = (V, E)$ is obtained from H by adding an edge between every two vertices at distance two in H. Given a graph class $\mathcal{H}$, the $\mathcal{H}$-SQUARE ROOT PROBLEM asks for the recognition of the squares of graphs in $\mathcal{H}$. In [56], [46], we answer positively to an open question of Golovach *et al.* (IWOCA'16) [77] by showing that the squares of *cactus-block graphs* can be recognized in polynomial time. Our proof is based on new relationships between the decomposition of a graph by cut-vertices and the decomposition of its square by clique cutsets. More precisely, we prove that the closed neighbourhoods of cut-vertices in H induce maximal subgraphs of $G = H^2$ with no clique-cutset. Furthermore, based on this relationship, we can compute from a given graph G the block-cut tree of a desired square root (if any). Although the latter tree is not uniquely defined, we show surprisingly that it can only differ marginally between two different roots. Our

approach not only gives the first polynomial-time algorithm for the $\mathcal{H}$-SQUARE ROOT PROBLEM for several graph classes $\mathcal{H}$, but it also provides a unifying framework for the recognition of the squares of trees, block graphs and cactus graphs — among others.

7.2.1.3. Graph hyperbolicity

The Gromov hyperbolicity is an important parameter for analyzing complex networks which expresses how the metric structure of a network looks like a tree (the smaller gap the better). It has recently been used to provide bounds on the expected stretch of greedy-routing algorithms in Internet-like graphs, and for various applications in network security, computational biology, the analysis of graph algorithms, and the classification of complex networks.

In [44], we answer open questions of Verbeek and Suri [81] on the relationships between Gromov hyperbolicity and the optimal stretch of graph embeddings in Hyperbolic space. Then, based on the relationships between hyperbolicity and Cops and Robber games, we turn necessary conditions for a graph to be Cop-win into sufficient conditions for a graph to have a large hyperbolicity (and so, no low-stretch embedding in Hyperbolic space). In doing so we derive lower-bounds on the hyperbolicity in various graph classes – such as Cayley graphs, distance-regular graphs and generalized polygons, to name a few. It partly fills in a gap in the literature on Gromov hyperbolicity, for which few lower-bound techniques are known.

In [23] we study practical improvements for the computation of hyperbolicity in large graphs. Precisely, we investigate relations between the hyperbolicity of a graph G and the hyperbolicity of its *atoms*, that are the subgraphs output by the clique-decomposition invented by Tarjan [80] and Leimer [79]. We prove that the maximum hyperbolicity taken over the atoms is at most one unit off from the hyperbolicity of G and the bound is sharp. We also give an algorithm to slightly modify the atoms, called the "substitution method", which is at no extra cost than computing the clique-decomposition, and so that the maximum hyperbolicity taken over the resulting graphs is *exactly* the hyperbolicity of the input graph G. Experimental evaluation on collaboration networks and biological networks shows that our method provides significant computation time savings. Finally, on a more theoretical side, we deduce from our results the first *linear-time* algorithm for computing the hyperbolicity of an outerplanar graph.

7.2.1.4. Computing metric hulls in graphs

Convexity in graphs generalises the classical convexity in Euclidean spaces. The *hull-number* of a graph is the minimum number k such that there exists a set of k vertices whose convex hull is the graph. Computing the hull-number is NP-hard even in very restricted graph classes such as partial cubes (isometric subgraphs of hypercubes). One challenging question in this area is the status of the parameterized complexity of this problem. We further investigate the complexity of a more general problem.

In [60], we prove that, given a closure function the smallest preimage of a closed set can be calculated in polynomial time in the number of closed sets. This confirms a conjecture of Albenque and Knauer and implies that there is a polynomial time algorithm to compute the convex hull-number of a graph, when all its convex subgraphs are given as input. We then show that computing if the smallest preimage of a closed set is logarithmic in the size of the ground set is LOGSNP-complete if only the ground set is given. A special instance of this problem is computing the dimension of a poset given its linear extension graph, that was conjectured to be in P.

The intent to show that the latter problem is LOGSNP-complete leads to several interesting questions and to the definition of the isometric hull, i.e., a smallest isometric subgraph containing a given set of vertices S. While for $|S| = 2$ an isometric hull is just a shortest path, we show that computing the isometric hull of a set of vertices is NP-complete even if $|S| = 3$. Finally, we consider the problem of computing the isometric hull-number of a graph and show that computing it is Σ_2^P-complete.

7.2.1.5. Application to bioinformatics

For a (possibly infinite) fixed family of graphs F, we say that a graph G overlays F on a hypergraph H if $V(H)$ is equal to $V(G)$ and the subgraph of G induced by every hyperedge of H contains some member of F as a spanning subgraph. While it is easy to see that the complete graph on $|V(H)|$ overlays F on a

hypergraph H whenever the problem admits a solution, the Minimum F-Overlay problem asks for such a graph with the minimum number of edges. This problem allows to generalize some natural problems which may arise in practice. For instance, if the family F contains all connected graphs, then Minimum F-Overlay corresponds to the Minimum Connectivity Inference problem (also known as Subset Interconnection Design problem) introduced for the low-resolution reconstruction of macro-molecular assembly in structural biology, a problem that has been studied jointly by COATI and ABS [72], [73], or for the design of networks. In [41], we show a strong dichotomy result regarding the polynomial vs. NP-hard status with respect to the considered family F. Roughly speaking, we show that the easy cases one can think of (e.g. when edge-less graphs of the right sizes are in F, or if F contains only cliques) are the only families giving rise to a polynomial problem: all others are NP-complete. We then investigate the parameterized complexity of the problem and give similar sufficient conditions on F that give rise to W[1]-hard, W[2]-hard or FPT problems when the parameter is the size of the solution. This yields an FPT/W[1]-hard dichotomy for a relaxed problem, where every hyperedge of H must contain some member of F as a (non necessarily spanning) subgraph.

7.2.1.6. Matchings for the recovery of disrupted airline operations

In an informal collaboration with Amadeus' members (A. Salch and V. Weber), we have studied the following problem. When an aircraft is approaching an airport, it gets a short time interval (called *slot*) that it can use to land. If the landing of the aircraft is delayed (because of bad weather, or if it arrives late, or if other aircrafts have to land first), it looses its slot and Air traffic controllers have to assign it a new slot. However, slots for landing are a scare resource of the airports and, to avoid that an aircraft waits too much time, Air traffic controllers have to regularly modify the assignment of the slots of the aircrafts. Unfortunately, for legal and economical reasons, Air traffic controllers can modify the slot-assignment only through specific kind of operations. The problem is then the following. Precisely, let $k \geq 1$ be an odd integer, a graph G and a matching M (set of pairwise disjoint edges) of G. What is the maximum size of a matching that can be obtained from M by *using* only augmenting paths of length *at most* k?

By Berge's theorem, finding a *maximum matching* in a graph relies on *the use* of *augmenting paths*. When no further constraint is added (k unbounded), Edmonds' algorithm allows to compute a maximum matching in polynomial time by sequentially augmenting such paths. In [39], we first prove that this problem can be solved in polynomial time for $k \leq 3$ in any graph and that it is *NP*-complete for any fixed $k \geq 5$ in the class of planar bipartite graphs of degree at most 3 and arbitrarily large girth. We then prove that this problem is in *P*, for any k, in several subclasses of trees such as caterpillars or trees with all vertices of degree at least 3 "far appart". Moreover, this problem can be solved in time $O(n)$ in the class of n-node trees when k and the maximum degree are fixed parameters. Finally, we consider a more constrained problem where only paths of length *exactly*k can be augmented. We prove that this latter problem becomes *NP*-complete for any fixed $k \geq 3$ and in trees when k is part of the input.

In [51], we perform a deeper analysis of the complexity of this problem for trees. On the positive side, we first show that it can be solved in polynomial time for more classes of trees, namely bounded-degree trees (via a dynamic programming approach), caterpillars and trees where the nodes with degree at least 3 are sufficiently far apart. On the negative side, we show that, when only paths of length *exactly*k can be augmented, the problem becomes *NP*-complete already for $k = 3$, in the class of planar bipartite graphs with maximum degree 3 and arbitrary large girth. We also show that the latter problem is *NP*-complete in trees when k is part of the input.

7.2.2. Graph decompositions and graph searching

It is well known that many NP-hard problems are tractable in the class of bounded treewidth graphs. In particular, tree-decompositions of graphs are an important ingredient of dynamic programming algorithms for solving such problems. This also holds for other width-parameters of graphs. Therefore, computing these widths and associated decompositions of graphs has both a theoretical and practical interest.

7.2.2.1. Minimum size tree-decompositions

We study in [31] the problem of computing a tree-decomposition of a graph with width at most k and minimum number of bags. More precisely, we focus on the following problem: given a fixed $k \geq 1$, what

is the complexity of computing a tree-decomposition of width at most k with minimum number of bags in the class of graphs with treewidth at most k? We prove that the problem is NP-complete for any fixed $k \geq 4$ and polynomial for $k \leq 2$; for $k = 3$, we show that it is polynomial in the class of trees and 2-connected outerplanar graphs.

7.2.2.2. Exclusive Graph Searching and pathwidth.

An algorithmic interpretation of tree/path-decomposition is the well known *graph searching* problem, where a team of searchers aims at capturing an intruder in a network, modeled as a graph. All variants of this problem assume that any node can be simultaneously occupied by several searchers. This assumption may be unrealistic, e.g., in the case of searchers modeling physical searchers, or may require each individual node to provide additional resources, e.g., in the case of searchers modeling software agents.

We thus introduce and investigate in [22] *Exclusive Graph Searching*, in which no two or more searchers can occupy the same node at the same time. As for the classical variants of graph searching, we study the minimum number of searchers required to capture the intruder. This number is called the *exclusive search number* of the considered graph. Exclusive graph searching appears to be considerably more complex than classical graph searching, for at least two reasons: (1) it does not satisfy the *monotonicity property*, and (2) it is not *closed under minor*. Moreover, we observe that the exclusive search number of a tree may differ exponentially from the values of classical search numbers (e.g., pathwidth). Nevertheless, we design a polynomial-time algorithm which, given any n-node tree T, computes the exclusive search number of T in time $O(n^3)$. Moreover, for any integer k, we provide a characterization of the trees T with exclusive search number at most k. Finally, we prove that the ratio between the exclusive search number and the pathwidth of a graph is bounded by its maximum degree.

In [32], we study the complexity of this new variant and show that there are graph classes where its complexity differs from the complexity of pathwidth. We show that the problem is NP-hard in planar graphs with maximum degree 3 and it can be solved in linear-time in the class of cographs. We also show that *monotone* Exclusive Graph Searching is NP-complete in split graphs where Pathwidth is known to be solvable in polynomial time. Moreover, we prove that monotone Exclusive Graph Searching is in P in a subclass of star-like graphs where Pathwidth is known to be NP-hard. Hence, the computational complexities of monotone Exclusive Graph Searching and Pathwidth cannot be compared. This is the first variant of Graph Searching for which such a difference is proved.

7.2.2.3. Distributed Graph Searching.

We then study exclusive graph searching in a distributed setting. Consider a set of mobile robots placed on distinct nodes of a discrete, anonymous, and bidirectional ring. Asynchronously, each robot takes a snapshot of the ring, determining the size of the ring and which nodes are either occupied by robots or empty. Based on the observed configuration, it decides whether to move to one of its adjacent nodes or not. In the first case, it performs the computed move, eventually. This model of computation is known as *Look-Compute-Move*. The computation depends on the required task. In [25], we solve both the well-known *Gathering* and *Exclusive Searching* tasks. In the former problem, all robots must simultaneously occupy the same node, eventually. In the latter problem, the aim is to clear all edges of the graph. An edge is cleared if it is traversed by a robot or if both its endpoints are occupied. We consider the *exclusive* searching where it must be ensured that two robots never occupy the same node. Moreover, since the robots are oblivious, the clearing is *perpetual*, i.e., the ring is cleared infinitely often.

In the literature, most contributions are restricted to a subset of initial configurations. Here, we design two different algorithms and provide a characterization of the initial configurations that permit the resolution of the problems under very weak assumptions. More precisely, we provide a full characterization (except for few pathological cases) of the initial configurations for which Gathering can be solved. The algorithm relies on the necessary assumption of the local-weak multiplicity detection. This means that during the Look phase a robot detects also whether the node it occupies is occupied by other robots, without acquiring the exact number.

For the exclusive searching, we characterize all (except for few pathological cases) aperiodic configurations from which the problem is feasible. We also provide some impossibility results for the case of periodic configurations.

7.2.3. Combinatorial games on graphs

We study several two-player games on graphs. Some of these games allow to model real-life applications. In the case of the Spy-game presented below, we propose a successful new approach by studying fractional relaxation of such games.

7.2.3.1. Localization Game on Geometric and Planar Graphs

Motivated by a localization problem in cellular networks, we introduce in [52] a model based on a pursuit graph game that resembles the famous Cops and Robbers game. It can be considered as a game theoretic variant of the *metric dimension* of a graph. Given a graph G we want to localize a walking agent by checking his distance to as few vertices as possible. We provide upper bounds on the related graph invariant $\zeta(G)$, defined as the least number of cops needed to localize the robber on a graph G, for several classes of graphs (trees, bipartite graphs, etc). Our main result is that, surprisingly, there exists planar graphs of treewidth 2 and unbounded $\zeta(G)$. On a positive side, we prove that $\zeta(G)$ is bounded by the pathwidth of G. We then show that the algorithmic problem of determining $\zeta(G)$ is NP-hard in graphs with diameter at most 2. Finally, we show that at most one cop can approximate (arbitrarily close) the location of the robber in the Euclidean plane.

7.2.3.2. Spy-Game on graphs

We define and study the following two-player game on a graph G. Let $k \in \mathbb{N}^*$. A set of k guards is occupying some vertices of G while one spy is standing at some vertex. At each turn, first the spy may move along at most s edges, where $s \in \mathbb{N}^*$ is his speed. Then, each guard may move along one edge. The spy and the guards may occupy the same vertices. The spy has to escape the surveillance of the guards, i.e., must reach a vertex at distance more than $d \in \mathbb{N}$ (a predefined distance) from every guard. Can the spy win against k guards? Similarly, what is the minimum distance d such that k guards may ensure that at least one of them remains at distance at most d from the spy? This game generalizes two well-studied games: Cops and robber games (when $s = 1$) and Eternal Dominating Set (when s is unbounded).

In [53], we consider the computational complexity of the problem, showing that it is NP-hard (for every speed s and distance d) and that some variant of it is PSPACE-hard in DAGs. Then, we establish tight tradeoffs between the number of guards, the speed s of the spy and the required distance d when G is a path or a cycle.

In order to determine the smallest number of guards necessary for this task, we analyze in [42], [43], [54] the game through a Linear Programming formulation and the *fractional strategies* it yields for the guards. We then show the equivalence of fractional and integral strategies in trees. This allows us to design a polynomial-time algorithm for computing an optimal strategy in this class of graphs. Using duality in Linear Programming, we also provide non-trivial bounds on the fractional guard-number of grids and torus. We believe that the approach using fractional relaxation and Linear Programming is promising to obtain new results in the field of combinatorial games.

7.2.3.3. Hyperopic Cops and Robbers

We introduce in [68] a new variant of the game of Cops and Robbers played on graphs, where the robber is invisible when located in the neighbor set of a cop. The hyperopic cop number is the corresponding analogue of the cop number, and we investigate bounds and other properties of this parameter. We characterize the cop-win graphs for this variant, along with graphs with the largest possible hyperopic cop number. We analyze the cases of graphs with diameter 2 or at least 3, focusing on when the hyperopic cop number is at most one greater than the cop number. We show that for planar graphs, as with the usual cop number, the hyperopic cop number is at most 3. The hyperopic cop number is considered for countable graphs, and it is shown that for connected chains of graphs, the hyperopic cop density can be any real number in $[0, 1/2]$.

7.3. Graph theory

Participants: Julien Bensmail, Guillaume Ducoffe, Frédéric Havet, William Lochet, Nicolas Nisse, Bruce Reed.

COATI studies theoretical problems in graph theory. If some of them are directly motivated by applications (see Subsection 7.3.3), others are more fundamental. In particular, we are putting an effort on understanding better directed graphs (also called *digraphs*) and partionning problems, and in particular colouring problems. We also try to better the understand the many relations between orientation and colourings. We study various substructures and partitions in (di)graphs. For each of them, we aim at giving sufficient conditions that guarantee its existence and at determining the complexity of finding it.

7.3.1. Substructures in (di)graphs

We study various conditions that ensure a (di)graph to contain certain substructures.

In [17], we study the question of finding a set of k vertex-disjoint cycles (resp. directed cycles) of distinct lengths in a given graph (resp. digraph). In the context of undirected graphs, we prove that, for every $k \geq 1$, every graph with minimum degree at least $\frac{k^2+5k-2}{2}$ has k vertex-disjoint cycles of different lengths, where the degree bound is the best possible. We also consider other cases such as when the graph is triangle-free, or the k cycles are required to have different lengths modulo some value r. In the context of directed graphs, we consider a conjecture of Lichiardopol concerning the least minimum out-degree required for a digraph to have k vertex-disjoint directed cycles of different lengths. We verify this conjecture for tournaments, and, by using the probabilistic method, for some regular digraphs and digraphs of small order.

A $(k_1 + k_2)$-*bispindle* is the union of $k_1(x, y)$-dipaths and $k_2(y, x)$-dipaths, all these dipaths being pairwise internally disjoint. Recently, Cohen et al. showed that for every $(1, 1)$-bispindle B, there exists an integer k such that every strongly connected digraph with chromatic number greater than k contains a subdivision of B. In [24], we investigate generalisations of this result by first showing constructions of strongly connected digraphs with large chromatic number without any $(3, 0)$-bispindle or $(2, 2)$-bispindle. Then we show that strongly connected digraphs with large chromatic number contains a $(2, 1)$-bispindle, where at least one of the (x, y)-dipaths and the (y, x)-dipath are long.

Let $\mathcal{H}$ be a family of graphs and let d be large enough. For every d-regular graph G, we study the existence of a spanning $\mathcal{H}$-free subgraph of G with large minimum degree. This problem is well understood if $\mathcal{H}$ does not contain bipartite graphs. In [35] we provide asymptotically tight results for many families of bipartite graphs such as cycles or complete bipartite graphs. To prove these results, we study a locally injective analogue of the question.

An *even pair* (resp. *odd pair*) in a graph is a pair of non-adjacent vertices such that every chordless path between them has even (resp. odd) length. Even and odd pairs are important tools in the study of perfect graphs and were instrumental in the proof of the Strong Perfect Graph Theorem. We suggest that such pairs impose a lot of structure also in arbitrary, not just perfect graphs. To this end, we show in [36] that the presence of even or odd pairs in graphs imply a special structure of the stable set polytope. In fact, we give a polyhedral characterization of even and odd pairs.

7.3.2. Colourings and partitioning (di)graphs

7.3.2.1. Colouring graphs with constraints on connectivity

A graph G has maximal local edge-connectivity k if the maximum number of edge-disjoint paths between every pair of distinct vertices x and y is at most k. We prove in [11] Brooks-type theorems for k-connected graphs with maximal local edge-connectivity k, and for any graph with maximal local edge-connectivity 3. We also consider several related graph classes defined by constraints on connectivity. In particular, we show that there is a polynomial-time algorithm that, given a 3-connected graph G with maximal local connectivity 3, outputs an optimal colouring for G. On the other hand, we prove, for $k \geq 3$, that k-colourability is NP-complete when restricted to minimally k-connected graphs, and 3-colourability is NP-complete when restricted to $(k-1)$-connected graphs with maximal local connectivity k. Finally, we consider a parameterization of k-colourability based on the number of vertices of degree at least $k + 1$, and prove that, even when k is part of the input, the corresponding parameterized problem is FPT.

7.3.2.2. Sum-distinguishing edge-weightings

A k-*edge-weighting* of a graph G is an application from $E(G)$ into $\{1, \cdots, k\}$. An edge-weigthing is *sum-distinguishing* if for every two adajcent vertices u and v, the sum of weights of edges incident to u is distinct from the sum of of weights of edges incident to v. The celebrated 1-2-3-Conjecture (raised in 2004 by Karoński, Luczak and Thomason) asserts that every connected graph (except K_2, the complete graph on two vertices) admits a sum-distinguishing 3-edge-weighting. This conjecture attracted much attention and many variants are now studied. We study several of them.

Towards the 1-2-3-Conjecture, the best-known result to date is due to Kalkowski, Karoński and Pfender, who proved that it holds when relaxed to 5-edge-weightings. Their proof builds upon a weighting algorithm designed by Kalkowski for a total version (where also the vertices are weighted) of the problem. In [67], we present new mechanisms for using Kalkowski's algorithm in the context of the 1-2-3 Conjecture. As a main result we prove that every 5-regular graph admits a 4-edge-weighting that permits to distinguish its adjacent vertices via their incident sums.

In [66], we investigate the consequences on the 1-2-3 Conjecture of requiring a stronger distinction condition. Namely, we consider two adjacent vertices distinguished when their incident sums differ by at least 2. As a guiding line, we conjecture that every graph with no connected component isomorphic to K_2 admits a 5-edge-weighting permitting to distinguish the adjacent vertices in this stronger way. We verify this conjecture for several classes of graphs, including bipartite graphs and cubic graphs. We then consider algorithmic aspects, and show that it is *NP*-complete to determine the smallest k such that a given bipartite graph admits such a k-edge-weighting. In contrast, we show that the same problem can be solved in polynomial time for a given tree.

In [13], we consider the following question, which stands as a directed analogue of the 1-2-3 Conjecture: Given any digraph D with no arc $\overrightarrow{uv}$ verifying $d^+(u) = d^-(v) = 1$, is it possible to weight the arcs of D with weights among $\{1, 2, 3\}$ so that, for every arc $\overrightarrow{uv}$ of D, the sum of incident weights out-going from u is different from the sum of incident weights in-coming to v? We answer positively to this question, and investigate digraphs for which even the weights among $\{1, 2\}$ are sufficient. In relation with the so-called 1-2 Conjecture, we also consider a total version of the problem, which we prove to be false. Our investigations turn to have interesting relations with open questions related to the 1-2-3 Conjecture.

In [21], we study the following question: Is it always possible to injectively assign the weights $1, ..., |E(G)|$ to the edges of any given graph G (with no component isomorphic to K_2) so that every two adjacent vertices of G get distinguished by their sums of incident weights? One may see this question as a combination of the well-known 1-2-3 Conjecture and the Antimagic Labelling Conjecture. We exhibit evidence that this question might be true. Benefiting from the investigations on the Antimagic Labelling Conjecture, we first point out that several classes of graphs, such as regular graphs, indeed admit such assignments. We then show that trees also do, answering a recent conjecture of Arumugam, Premalatha, Bača and Semaničová-Feňovčíková. Towards a general answer to the question above, we then prove that claimed assignments can be constructed for any graph, provided we are allowed to use some number of additional edge weights. For some classes of sparse graphs, namely 2-degenerate graphs and graphs with maximum average degree 3, we show that only a small (constant) number of such additional weights suffices.

7.3.2.3. Variants of vertex- or edge-colouring

A colouring of a graph G is *properly connected* if every two vertices of G are the ends of a properly coloured path. In [57], [47], we study the *complexity* of computing the *proper connection number* (minimum number of colours in a properly connected colouring) for edge and vertex colourings, in undirected and directed graphs, respectively. First we disprove some conjectures of Magnant et al. (2016) on characterizing the strong digraphs with *proper arc connection number* at most two. Then, we prove that deciding whether a given digraph has proper arc connection number at most two is NP-complete. Furthermore, we show that there are infinitely many such digraphs with no even-length dicycle. We initiate the study of proper vertex connectivity in digraphs and we prove similar results as for the arc version. Finally, we present polynomial-time recognition algorithms for *bounded-treewidth* graphs and *bipartite* graphs with *proper edge connection number* at most two.

A graph is *locally irregular* if no two adjacent vertices have the same degree. The *irregular chromatic index* $\chi'_{\text{irr}}(G)$ of a graph G is the smallest number of locally irregular subgraphs needed to edge-decompose G. Not all graphs have such a decomposition, but Baudon, Bensmail, Przybyło, and Woźniak conjectured that if G can be decomposed into locally irregular subgraphs, then $\chi'_{\text{irr}}(G) \leq 3$. In support of this conjecture, Przybyło showed that $\chi'_{\text{irr}}(G) \leq 3$ holds whenever G has minimum degree at least 10^{10}. In [19] we prove that every bipartite graph G which is not an odd length path satisfies $\chi'_{\text{irr}}(G) \leq 10$. This is the first general constant upper bound on the irregular chromatic index of bipartite graphs. Combining this result with Przybyło's result, we show that $\chi'_{\text{irr}}(G) \leq 328$ for every graph G which admits a decomposition into locally irregular subgraphs. Finally, we show that $\chi'_{\text{irr}}(G) \leq 2$ for every 16-edge-connected bipartite graph G.

An (m, n)-coloured mixed graph is a mixed graph with arcs assigned one of m different colours and edges one of n different colours. A homomorphism of an (m, n)-coloured mixed graph G to an (m, n)-coloured mixed graph H is a vertex mapping such that if uv is an arc (edge) of colour c in G, then $f(u)f(v)$ is also an arc (edge) of colour c. The (m, n)-coloured mixed chromatic number, denoted $\chi_{m,n}(G)$, of an (m, n)-coloured mixed graph G is the order of a smallest homomorphic image of G. An (m, n)-clique is an (m, n)-coloured mixed graph C with $\chi_{m,n}(C) = |V(C)|$. In [16], we study the structure of (m, n)-cliques. We show that almost all (m, n)-coloured mixed graphs are (m, n)-cliques, prove bounds for the order of a largest outerplanar and planar (m, n)-clique and resolve an open question concerning the computational complexity of a decision problem related to $(0, 2)$-cliques. Additionally, we explore the relationship between $\chi_{1,0}$ and $\chi_{0,2}$.

An edge colouring of a graph G is called acyclic if it is proper and every cycle contains at least three colours. We show in [33] that for every $\varepsilon > 0$, there exists a $g = g(\varepsilon)$ such that if G has maximum degree Δ and girth at least g then G admits an acyclic edge colouring with $(1 + \varepsilon)\Delta + O(1)$ colours.

7.3.3. Identifying codes

Let G be a graph G. The *neighborhood* of a vertex v in G, denoted by $N(v)$, is the set of vertices adjacent to v in G. Its *closed neighborhood* is the set $N[v] = N(v) \cup \{v\}$. A set $C \subseteq V(G)$ is an *identifying code* in G if (i) for all $v \in V(G)$, $N[v] \cap C \neq \varnothing$, and (ii) for all $u, v \in V(G)$, $N[u] \cap C \neq N[v] \cap C$. The problem of finding low-density identifying codes was introduced in [Karpovsky et al., IEEE Trans. Inform. Theory 44, 1998] in relation to fault diagnosis in arrays of processors. Here the vertices of an identifying code correspond to controlling processors able to check themselves and their neighbors. Thus the identifying property guarantees location of a faulty processor from the set of "complaining" controllers. Identifying codes are also used in [Ray et al., IEEE Journal on Selected Areas in Communications 22, 2004] to model a location detection problem with sensor networks.

A particular interest was dedicated to grids as many processor networks have a grid topology. There are several types of standard regular infinite grids, in particular the hexagonal grids, the square grids, the triangular grids and the king grids. For such graphs G, the problem consists in finding the minimum density $d^*(G)$ of an identifying code of G.

In [26], we study the infinite triangular grid T_k with k rows. We show $d^*(T_1) = d^*(T_2) = 1/2$, $d^*(T_3) = d^*(T_4) = 1/3$, $d^*(T_5) = 3/10$, $d^*(T_6) = 1/3$ and $d^*(T_k) = 1/4 + 1/(4k)$ for all odd $k \geq 7$. In addition, we show that $1/4 + 1/(4k) \leq d^*(T_k) \leq 1/4 + 1/(2k)$ for all even $k \geq 8$.

In [27], we study the density of king grids which are strong product of two paths. We show that for every king grid G, $d^*(G) \geq 2/9$. In addition, we show this bound is attained only for king grids which are strong products of two infinite paths. Given $k \geq 3$, we denote by K_k the (infinite) king strip with k rows. We prove that $d^*(K_3) = 1/3$, $d^*(K_4) = 5/16$, $d^*(K_5) = 4/15$ and $d^*(K_6) = 5/18$. We also prove that $2/9 + 8/81k \leq d^*(K_k) \leq 2/9 + 4/9k$ for every $k \geq 7$.

7.3.4. Miscellaneous

7.3.4.1. A proof of the Barát-Thomassen conjecture

The Barát-Thomassen conjecture asserts that for every tree T on m edges, there exists a constant k_T such that every k_T-edge-connected graph with size divisible by m can be edge-decomposed into copies of T. So far this

conjecture has only been verified when T is a path or when T has diameter at most 4. In [18], we prove the full statement of the conjecture.

7.3.4.2. Recursively partitionable graphs

A connected graph G is said to be *arbitrarily partitionable* (AP for short) if for every partition $(n_1, ..., n_p)$ of $|V(G)|$ there exists a partition $(V_1, ..., V_p)$ of $V(G)$ such that each V_i induces a connected subgraph of G on n_i vertices. Some stronger versions of this property were introduced, namely the ones of being *online arbitrarily partitionable* and *recursively arbitrarily partitionable* (OL-AP and R-AP for short, respectively), in which the subgraphs induced by a partition of G must not only be connected but also fulfil additional conditions. In [14], we point out some structural properties of OL-AP and R-AP graphs with connectivity 2. In particular, we show that deleting a cut pair of these graphs results in a graph with a bounded number of components, some of whom have a small number of vertices. We obtain these results by studying a simple class of 2-connected graphs called *balloons*

7.3.4.3. On oriented cliques with respect to push operation

An oriented graph is a directed graph without any directed cycle of length at most 2. An oriented clique is an oriented graph whose non-adjacent vertices are connected by a directed 2-path. To push a vertex v of a directed graph $\overrightarrow{G}$ is to change the orientations of all the arcs incident to v. A push clique is an oriented clique that remains an oriented clique even if one pushes any set of vertices of it. We show in [20] that it is NP-complete to decide if an undirected graph is the underlying graph of a push clique or not. We also prove that a planar push clique can have at most 8 vertices and provide an exhaustive list of planar push cliques.

7.3.4.4. On q-power cycles in cubic graphs

In the context of a conjecture of Erdős and Gyárfás, we consider in [15], for any $q \geq 2$, the existence of q-power cycles (*i.e.* with length a power of q) in cubic graphs. We exhibit constructions showing that, for every $q \geq 3$, there exist arbitrarily large cubic graphs with no q-power cycles. Concerning the remaining case $q = 2$ (which corresponds to the conjecture of Erdős and Gyárfás), we show that there exist arbitrarily large cubic graphs whose only 2-power cycles have length 4 only, or 8 only.

7.3.4.5. How to determine if a random graph with a fixed degree sequence has a giant component

For a fixed degree sequence $\mathcal{D} = (d_1, ..., d_n)$, let $G(\mathcal{D})$ be a uniformly chosen (simple) graph on $\{1, \cdots, n\}$ where the vertex i has degree d_i. In [34] we determine whether $G(\mathcal{D})$ has a giant component with high probability, essentially imposing no conditions on $\mathcal{D}$. We simply insist that the sum of the degrees in $\mathcal{D}$ which are not 2 is at least $\lambda(n)$ for some function λ going to infinity with n. This is a relatively minor technical condition, and when $\mathcal{D}$ does not satisfy it, both the probability that $G(\mathcal{D})$ has a giant component and the probability that $G(\mathcal{D})$ has no giant component are bounded away from 1.

7.3.4.6. A proof of the Erdős-Sands-Sauer-Woodrow conjecture

A very nice result of Barany and Lehel asserts that every finite subset X of R^d can be covered by $f(d)X$-boxes (i.e. each box has two antipodal points in X). As shown by Gyárfás and Pálvőlgyi this result would follow from the following conjecture : If a tournament admits a partition of its arc set into k partial orders, then its domination number is bounded in terms of k. This question is in turn implied by the Erdős-Sands-Sauer-Woodrow conjecture : If the arcs of a tournament T are colored with k colors, there is a set X of at most $g(k)$ vertices such that for every vertex v of T, there is a monochromatic path from X to v. We give in [69] a short proof of this statement. We moreover show that the general Sands-Sauer-Woodrow conjecture (which as a special case implies the stable marriage theorem) is valid for directed graphs with bounded stability number. This conjecture remains however open.

DANTE Project-Team

7. New Results

7.1. Graph & Signal Processing

Participants: Paulo Gonçalves, Éric Fleury, Sarra Ben Alaya, Esteban Bautista Ruiz, Gaëtan Frusque, Sarah de Nigris, Mikhail Tsitsvero.

7.1.1. Fractional Semi-Supervised Machine Learning

Graph-based semi-supervised learning for classification endorses a nice interpretation in terms of diffusive random walks, where the regularisation factor in the original optimisation formulation plays the role of a restarting probability. Recently, a new type of biased random walks for characterising certain dynamics on networks have been defined and rely on the γ-th power of the standard Laplacian matrix $\mathbf{L}^\gamma$, with $\gamma > 0$. In particular, these processes embed long range transitions, the Lévy flights, that are capable of one-step jumps between far-distant states (nodes) of the graph. In a series of two articles [28] and [29], we envisioned to build upon these volatile random walks to propose two new versions of graph based semi-supervised learning algorithms: one called fractional SSL corresponds to the case where $0 < \gamma < 1$ whose classification outcome could benefit from the dynamics induced by the fractional transition matrix, and the other less straightforwardly connected to random walks, derives from $\gamma > 1$.

7.1.2. Design of graph filters and filterbanks

Basic operations in graph signal processing consist in processing signals indexed on graphs either by filtering them or by changing their domain of representation, in order to better extract or analyze the important information they contain. The aim of our chapter [58] is to review general concepts underlying such filters and representations of graph signals. We first recall the different Graph Fourier Transforms that have been developed in the literature, and show how to introduce a notion of frequency analysis for graph signals by looking at their variations. Then, we move to the introduction of graph filters, that are defined like the classical equivalent for 1D signals or 2D images, as linear systems which operate on each frequency of a signal. Some examples of filters and of their implementations are given. Finally, as alternate representations of graph signals, we focus on multiscale transforms that are defined from filters. Continuous multiscale transforms such as spectral wavelets on graphs are reviewed, as well as the versatile approaches of filterbanks on graphs. Several variants of graph filterbanks are discussed, for structured as well as arbitrary graphs, with a focus on the central point of the choice of the decimation or aggregation operators.

7.1.3. GraSP: A Matlab Toolbox for Graph Signal Processing

In [30], we publicised the recent developments and new functionalities of our Graph Signal Processing Toolbox (GraSP).

7.2. Performance analysis and networks protocols

Participants: Mohammed Amer, Thomas Begin, Anthony Busson, Éric Fleury, Yannick Leo, Isabelle Guerin Lassous, Philippe Nain, Huu Nghi Nguyen, Laurent Reynaud.

7.2.1. Network Softwarization

We have developed a modelling framework to analytically evaluate the performance of DPDK-based virtual switches in the context of NFV (Network Function Virtualisation) networks. In [34], we extended our previous work [82] to enable non-null switch-over times that account for a delay overhead whenever a CPU starts polling a different queue. More recently, in [35], we refined our framework to let it deal with batches of packets (i.e. several packets on the same queue are processed together) that tends to speed up the performance of the virtual switches. These works were partly funded by the French ANR REFLEXION under the "ANR-14-CE28-0019" project.

7.2.2. Wi-Fi optimization

Densification of Wi-Fi networks has led to the possibility for a station to choose between several access points (APs). On the other hand, the densification of APs generates interference, contention and decreases the global throughput as APs have to share a limited number of channels. Optimizing the association step between APs and stations can alleviate this problem and increase the overall throughput and fairness between stations. We have proposed an original solution to this optimization problem based on a mathematical model and introduce a local search algorithm to solve this problem through a suitable neighborhood structure. Our evaluation, based on simulations, shows that the proposed solution improves the overall throughput and the fairness of the network. We are currently working on variant of this problem where the traffic to the stations is taken into account in the model and the optimization formulation.

7.2.3. Caching

In [72] we focus on the LRU cache where requests for distinct contents are described by independent stationary and ergodic processes. We extend a TTL-based approximation of the cache hit probability first proposed by R. Fagin in 1977 for the independence reference model to this more general workload model. We show that under very general conditions this approximation is exact as the cache size and the number of contents go to infinity. Moreover, we establish this not only for the aggregate cache hit probability but also for every individual content. Last, we obtain a rate of convergence.

In [70] we consider the problem of allocating cache resources among multiple content providers. The cache can be partitioned into slices and each partition can be dedicated to a particular content provider, or shared among a number of them. It is assumed that each partition employs the LRU policy for managing content. We propose utility-driven partitioning, where we associate with each content provider a utility that is a function of the hit rate observed by the content provider. We consider two scenarios: i) content providers serve disjoint sets of files, ii) there is some overlap in the content served by multiple content providers. In the first case, we prove that cache partitioning outperforms cache sharing as cache size and numbers of contents served by providers go to infinity. In the second case, It can be beneficial to have separate partitions for overlapped content. In the case of two providers it is usually always beneficial to allocate a cache partition to serve all overlapped content and separate partitions to serve the non-overlapped contents of both providers. We establish conditions when this is true asymptotically but also present an example where it is not true asymptotically. We develop online algorithms that dynamically adjust partition sizes in order to maximize the overall utility and prove that they converge to optimal solutions, and through numerical evaluations we show they are effective.

7.2.4. Mobile networks

The development of analytical models to analyze the behavior of vehicular ad hoc networks (VANETs) is a challenging aim. Adaptive methods are suitable for many algorithms (*e.g.* choice of forwarding paths, dynamic resource allocation, channel control congestion) and services (*e.g.* provision of multimedia services, message dissemination). These adaptive algorithms help the network to maintain a desired performance level. However, this is a difficult goal to achieve, especially in VANETs due to fast position changes of the VANET nodes. Adaptive decisions should be taken according to the current conditions of the VANET. Therefore, evaluation of transient measures is required for the characterization of VANETs. In the literature, different works address the characterization and measurement of the idle (or busy) time to be used in different proposals to attain a more efficient usage of wireless network. We have developed an analytical model based on a straightforward Markov reward chain (MRC) to obtain transient measurements of the idle time of the link between two VANET nodes. We have shown that numerical results from the analytical model fit well with simulation results [20].

In another study, we have investigated the application of an adapted controlled mobility strategy on self-propelling nodes, which could efficiently provide network resource to users scattered on a designated area. We have designed a virtual force-based controlled mobility scheme (called VFPc) and evaluated its ability to be jointly used with a dual packet-forwarding and epidemic routing protocol. In particular, we have studied the possibility for end-users to achieve synchronous communications at given times of the considered scenarios. On this basis, we have studied the delay distribution for such user traffic and show the advantages of our

solution compared to other packet-forwarding and packet-replication schemes, and highlighted that VFPc-enabled applications could take benefit of both schemes to yield a better user experience, despite challenging network conditions [21].

7.3. Modeling of Dynamics of Complex Networks

Participants: Jean Pierre Chevrot, Christophe Crespelle, Sicheng Dai, Éric Fleury, Eric, Philippe Guichard, Márton Karsai, Yannick Leo, Sebastien Lerique, Jacob Levy Abitbol, Jean-Philippe Magué, Matteo Morini, Samuel Unicomb, Samuel Unicomb.

7.3.1. Multilayer networks

In [67] we introduce a new class of stochastic multilayer networks. A stochastic multilayer network is the aggregation of M networks (one per layer) where each is a subgraph of a foundational network G. Each layer network is the result of probabilistically removing links and nodes from G. The resulting network includes any link that appears in at least K layers. This model is an instance of a non-standard site-bond percolation model. Two sets of results are obtained: first, we derive the probability distribution that the M-layer network is in a given configuration for some particular graph structures (explicit results are provided for a line, an algorithm is provided for a tree), where a configuration is the collective state of all links (each either active or inactive). Next, we show that for appropriate scalings of the node and link selection processes in a layer, links are asymptotically independent as the number of layers goes to infinity, and follow a Poisson distribution. Numerical results are provided to highlight the impact of having several layers on some metrics of interest (including expected size of the cluster a node belongs to in the case of the line). This model finds applications in wireless communication networks with multichannel radios, multiple social networks with overlapping memberships, transportation networks, and, more generally, in any scenario where a common set of nodes can be linked via co-existing means of connectivity.

7.3.2. Models of time varying networks

In terms of modelling temporal networks we had the following main contributions in 2017.

A book on Bursty Human Dynamics, written by M. Karsai as a the leading author. Bursty dynamics is a common temporal property of various complex systems in Nature but it also characterises the dynamics of human actions and interactions. At the phenomenological level it is a feature of all systems that evolve heterogeneously over time by alternating between periods of low and high event frequencies. In such systems, bursts are identified as periods in which the events occur with a rapid pace within a short time-interval while these periods are separated by long periods of time with low frequency of events. As such dynamical patterns occur in a wide range of natural phenomena, their observation, characterisation, and modelling have been a long standing challenge in several fields of research. However, due to some recent developments in communication and data collection techniques it has become possible to follow digital traces of actions and interactions of humans from the individual up to the societal level. This led to several new observations of bursty phenomena in the new but largely unexplored area of human dynamics, which called for the renaissance to study these systems using research concepts and methodologies, including data analytics and modelling. As a result, large amount of new insight and knowledge as well as innovations have been accumulated in the field, which provided the timely opportunity to write a monograph book [56] to make an up-to-date review and summary of the observations, appropriate measures, modelling, and applications of heterogeneous bursty patterns occurring in the dynamics of human behaviour.

In another contribution M. Karsai and collaborators introduced a new representation of temporal networks [73]. The dynamics of diffusion-like processes on temporal networks are influenced by correlations in the times of contacts. This influence is particularly strong for processes where the spreading agent has a limited lifetime at nodes: disease spreading (recovery time), diffusion of rumors (lifetime of information), and passenger routing (maximum acceptable time between transfers). We introduce weighted event graphs as a powerful and fast framework for studying connectivity determined by time-respecting paths where the allowed waiting times between contacts have an upper limit. We study percolation on the weighted

event graphs and in the underlying temporal networks, with simulated and real-world networks. We show that this type of temporal-network percolation is analogous to directed percolation, and that it can be characterized by multiple order parameters.

M. Karsai also contributed to a new definition to better quantify attention distributed in dynamical egocentric social networks [64]. Granovetter's weak tie theory of social networks is built around two central hypotheses. The first states that strong social ties carry the large majority of interaction events; the second maintains that weak social ties, although less active, are often relevant for the exchange of especially important information (e.g., about potential new jobs in Granovetter's work). While several empirical studies have provided support for the first hypothesis, the second has been the object of far less scrutiny. A possible reason is that it involves notions relative to the nature and importance of the information that are hard to quantify and measure, especially in large scale studies. Here, we search for empirical validation of both Granovetter's hypotheses. We find clear empirical support for the first. We also provide empirical evidence and a quantitative interpretation for the second. We show that attention, measured as the fraction of interactions devoted to a particular social connection, is high on weak ties — possibly reflecting the postulated informational purposes of such ties — but also on very strong ties. Data from online social media and mobile communication reveal network-dependent mixtures of these two effects on the basis of a platform's typical usage. Our results establish a clear relationships between attention, importance, and strength of social links, and could lead to improved algorithms to prioritize social media content.

7.3.3. Dynamical processes on networks

Another field which has been intensively studied during the last year addresses dynamical processes on temporal and static networks.

In a book chapter M. Karsai summarised his recent findings on temporal network immunisation [57]. The vast majority of strategies aimed at controlling contagion processes on networks consider a timescale separation between the evolution of the system and the unfolding of the process. However, in the real world, many networks are highly dynamical and evolve, in time, concurrently to the contagion phenomena. Here, we review the most commonly used immunization strategies on networks. In the first part of the chapter, we focus on controlling strategies in the limit of timescale separation. In the second part instead, we introduce results and methods that relax this approximation. In doing so, we summarize the main findings considering both numerical and analytically approaches in real as well as synthetic time-varying networks.

With the PhD student S. Unicomb, M. Karsai and a collaborator developed a new formalism, which is capable to precisely capture and predict the non-monotonous dependence of threshold driven dynamics on weight heterogeneities in networks [76]. Weighted networks capture the structure of complex systems where interaction strength is meaningful. This information is essential to a large number of processes, such as threshold dynamics, where link weights reflect the amount of influence that neighbours have in determining a node's behaviour. Despite describing numerous cascading phenomena, such as neural firing or social contagion, the modelling of threshold dynamics on weighted networks has been largely overlooked. We fill this gap by studying a dynamical threshold model over synthetic and real weighted networks with numerical and analytical tools. We show that the time of cascade emergence depends non-monotonously on weight heterogeneities, which accelerate or decelerate the dynamics, and lead to non-trivial parameter spaces for various networks and weight distributions. Our methodology applies to arbitrary binary state processes and link properties, and may prove instrumental in understanding the role of edge heterogeneities in various natural and social phenomena.

With other co-authors, M. Karsai published another book chapter about his recent findings on the modelling threshold driven dynamics on networks [55]. The collective behaviour of people adopting an innovation, product or online service is commonly interpreted as a spreading phenomenon throughout the fabric of society. This process is arguably driven by social influence, social learning and by external effects like media. Observations of such processes date back to the seminal studies by Rogers and Bass, and their mathematical modelling has taken two directions: One paradigm, called simple contagion, identifies adoption spreading with an epidemic process. The other one, named complex contagion, is concerned with behavioural thresholds

and successfully explains the emergence of large cascades of adoption resulting in a rapid spreading often seen in empirical data. The observation of real world adoption processes has become easier lately due to the availability of large digital social network and behavioural datasets. This has allowed simultaneous study of network structures and dynamics of online service adoption, shedding light on the mechanisms and external effects that influence the temporal evolution of behavioural or innovation adoption. These advancements have induced the development of more realistic models of social spreading phenomena, which in turn have provided remarkably good predictions of various empirical adoption processes. In this chapter we review recent data-driven studies addressing real-world service adoption processes. Our studies provide the first detailed empirical evidence of a heterogeneous threshold distribution in adoption. We also describe the modelling of such phenomena with formal methods and data-driven simulations. Our objective is to understand the effects of identified social mechanisms on service adoption spreading, and to provide potential new directions and open questions for future research.

Y. Leo, E. Fleury and M. Karsai is in the final stage to publish a study on a unique mobile call/banking dataset on the dynamics of purchasing patterns. We analyse a coupled dataset collecting the mobile phone communications and bank transactions history of a large number of individuals living in a Latin American country. After mapping the social structure and introducing indicators of socioeconomic status, demographic features, and purchasing habits of individuals we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of stratification in the social structure. In addition we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Finally, by analysing individual consumption histories, we detect dynamical patterns in purchase behaviour and their correlations with the socioeconomic status, demographic characters and the egocentric social network of individuals. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in resource allocation, marketing, and recommendation system design.

7.3.4. SoSweet

The SoSweet project focuses on the synchronic variation and the diachronic evolution of the variety of French language used on Twitter.

In one paper accepted to WWW'18 we addressed some of the main questions of the project using a unique dataset combining the largest French Twitter dataset and demographic data coming from INSEE [31]. Our usage of language is not solely reliant on cognition but is arguably determined by myriad external factors leading to a global variability of linguistic patterns. This issue, which lies at the core of sociolinguistics and is backed by many small-scale studies on face-to-face communication, is addressed here by constructing a dataset combining the largest French Twitter corpus to date with detailed socioeconomic maps obtained from national census in France. We show how key linguistic variables measured in individual Twitter streams depend on factors like socioeconomic status, location, time, and the social network of individuals. We found that *(i)* people of higher socioeconomic status, active to a greater degree during the daytime, use a more standard language; *(ii)* the southern part of the country is more prone to use more standard language than the northern one, while locally the used variety or dialect is determined by the spatial distribution of socioeconomic status; and *(iii)* individuals connected in the social network are closer linguistically than disconnected ones, even after the effects of status homophily have been removed. Our results inform sociolinguistic theory and may inspire novel learning methods enabling the inference of socioeconomic status of people from the way they tweet.

7.3.5. Relational methods for media studies

A very relevant application of the research that DANTE carries out on networks structures and networks dynamics concerns the field of journalism and media study. Relational analysis may be helpful in these fields in two different way.

On the one hand, the advent of digital media has challenged the established vertical structure of information distribution typical of broadcasting media with a decentralised organisation that facilitates the spreading of contents through all sort of horizontal channels (in the Web and in Social Media). This new type of circulation

is still insufficiently studied and require both quantitative and qualitative investigation. We tried to provide the first in our Field Guide to Fake News already introduced in the highlights of this document [68] and in a forthcoming chapter on the heterogeneous clustering of French Media system for the The Routledge Handbook to Developments in Digital Journalism Studies [60]. As for the qualitative study of the structure of the media system, we published and analysis of the strategies employed by Facebook to steer the evolution of the technology of Live Video Streaming [23].

On the other hand, network analysis can be a powerful tool to investigate and narrate journalistic stories, but its techniques need adapted to the language used by journalists and understood by their audiences. We tried to provide such a translation in a paper for the journal Digital Journalism [13] and in a chapter for the Datafied Society book [59].

The use of network analysis to study vast societal phenomena has also profound implications for the theory of social sciences, which we tried to explore in a paper for the journal Big Data & Society [25] and in a chapter of a book on the Frontiers of Social Science [63], and for their practice [62].

7.3.6. Philosophy of technologies revisited by Internet

The Internet, as a technology of writing, helps us to understand that a technology is not always a mean to reach a goal, nor an application of science. In fact, the Internet does not appear as a revolution, but as a revealer. We understand that a technology can be reflexive (it invites us to think it) and that it cannot be clearly separated from human activities (writing, etc.). For instance, 50 years ago, we imagined we could think with our own mind (and perhaps with a paper and pencil). Now, we know that we cannot think without material stuff (a computer, the internet, etc.). A very few philosophers knew this fact (Leibniz, Boole, etc.). But this evidence transforms completely the philosophy of technologies. Another important point is the effect of technology on epistemology. We realise that we can not ask or imagine some questions if the technology is not there (eg: social cartography or statistics). This fact invites us to insert technologies and methods in the traditional diptych of theory and experience. In synthesis, we also discover strong links between technology and culture; hence the role of engineers in the construction of culture [53], [18], [52], [54], [71].

DIANA Project-Team

6. New Results

6.1. Service Transparency

6.1.1. On active sampling of controlled experiments for QoE modeling

Participants: Muhammad Jawad Khokhar, Nawfal Abbasi Saber, Thierry Spetebroot, Chadi Barakat.
For internet applications, measuring, modeling and predicting the quality experienced by end users as a function of network conditions is challenging. A common approach for building application specific Quality of Experience (QoE) models is to rely on controlled experimentation. For accurate QoE modeling, this approach can result in a large number of experiments to carry out because of the multiplicity of the network features, their large span (e.g., band-width, delay) and the time needed to setup the experiments themselves. However, most often, the space of network features in which experimentations are carried out shows a high degree of uniformity in the training labels of QoE. This uniformity, difficult to predict beforehand, amplifies the training cost with little or no improvement in QoE modeling accuracy. So, in this work, we aim to exploit this uniformity, and propose a methodology based on active learning, to sample the experimental space intelligently, so that the training cost of experimentation is reduced. We prove the feasibility of our methodology by validating it over a particular case of YouTube streaming, where QoE is modeled both in terms of interruptions and stalling duration. This first validation has appeared in [19]. In another paper which is currently under submission, we propose an online version of this methodology together with a set of criterion to stop the experiments when the learner is confident enough.

6.1.2. On the Cost of Measuring Traffic in a Virtualized Environment

Participants: Karyna Gogunska, Chadi Barakat, Guillaume Urvoy-Keller, Dino Lopez Pacheco.
The current trend in application development and deployment is to package applications and services within containers or virtual machines. This results in a blend of virtual and physical resources with complex interconnection network schemas mixing virtual and physical switches along with specific protocols to build virtual networks spanning over several servers. While the complexity of this set-up is hidden by private/public cloud management solutions, e.g. OpenStack, this constitutes a challenge when it comes to monitor and debug performance related issues. In this work, carried out in collaboration with the Signet team of I3S with the support of the UCN@SOPHIA Labex, we introduce the problem of measuring traffic in a virtualized environment and focus on one typical scenario, namely virtual servers interconnected with a virtual switch. For this scenario, we assess the cost of continuously measuring the network traffic activity of the machines. Specifically, we seek to estimate the competition that exists to access the physical resources (CPU, memory, etc.) of the physical substrate between the measurement task and the legacy application activity. The results of this first study are currently under submission.

6.1.3. LISP measurements

Participant: Damien Saucez.
The Locator/Identifier Separation Protocol (LISP) separates classical IP addresses into two categories: one for identifying terminals, the other for routing. To associate identifiers and locators LISP needs a specific mechanism, called mapping system. This technology is still at an early stage but two experimental platforms have already been deployed in the Internet: LISP Beta Network and LISP-Lab. However, only the LISP Beta Network is monitored with LISPmon that partially monitors the mapping system once a day. To accompany the growth of LISP, a dynamic and complete monitoring system is required. Therefore, we propose LISP-Views, a dynamic versatile large scale LISP monitoring architecture. LISP-Views allows to automatically conduct comprehensive and objective measurements. After running LISP-Views in the wild for several months and comparing the monitoring results with LISPmon, we confirm that LISP-Views provides more detailed and accurate information. We observe the different behaviours between every network entity within mapping system, and also explore the current LISP performance for further improvements. A paper on "LISP-Views Monitoring LISP at Large Scale" was published in ITC this year.

6.2. Open Network Architecture

6.2.1. Controller load in SDN networks

Participant: Damien Saucez.

In OpenFlow, a centralized programmable controller installs forwarding rules into switches to implement policies. However, this flexibility comes at the expense of extra overhead in signalling and number of rules to install. The community considered that it was essential to install all rules and strictly respect routing requirements, hence working on making extra fast and large memory switches and controllers. Instead we took an opposite direction and came with a new vision that leverages the SDN concept and considers the network as a black box where tailored rules should be used only for network traffic that really matters while for the rest a good-enough (sub-optimal but cheap) default behaviour should be enough. In the past, we applied this vision to limit the needed memory on network switches in [5]. Lately, we proposed solutions to limit the number of exchanged messages between the switches and the controller. More precisely, in [31], [16] we developed a distributed sampling adaptive algorithm that allows switches to locally decide if they can contact the controller or if instead they should make their own decision locally. Numerical evaluation and emulation in Mininet demonstrate the benefit of the approach. The results were published in the PGMO (Gaspard Monge Program for Optimisation) days, Nov 2017, Paris, France.

6.2.2. Traceroute facility for Content-Centric Network

Participant: Thierry Turletti.

In the context of the UHD-on-5G associated team with our colleagues at NICT, Japan, we have proposed the Contrace tool for Measuring and Tracing Content-Centric Networks (CCNs). CCNs are fundamental evolutionary technologies that promise to form the cornerstone of the future Internet. The information flow in these networks is based on named data requesting, in-network caching, and forwarding – which are unique and can be independent of IP routing. As a result, common IP-based network tools such as ping and traceroute can neither trace a forwarding path in CCNs nor feasibly evaluate CCN performance. We designed Contrace, a network tool for CCNs (particularly, CCNx implementation running on top of IP) that can be used to investigate 1) the Round-Trip Time (RTT) between content forwarder and consumer, 2) the states of in-network cache per name prefix, and 3) the forwarding path information per name prefix. This tool can estimate the content popularity and design more effective cache control mechanisms in experimental networks. We have published an Internet-Draft [30] describing the specification of Contrace.

6.2.3. Message Dissemination in Intelligent Transport Systems

Participant: Thierry Turletti.

We proposed D2-ITS, a flexible and extensible framework to dynamically distribute network control to enable message dissemination in Intelligent Transport Systems (ITS). By decoupling the control from the data plane, D2-ITS leverages network programmability to address ITS scalability, delay intolerance and decentralization. It uses a distributed control plane based on a hierarchy of controllers that can dynamically adjust to environment and network conditions in order to satisfy ITS application requirements. We demonstrate the benefits of D2-ITS through a proof-of-concept prototype using the ns-3 simulation platform. Results indicate lower message delivery latency with minimal additional overhead. This work has been presented at the IEEE/ACM Symposium on Distributed Simulation and Real Time Applications (DS-RT) in October 2017 [18].

6.2.4. Peer-assisted Information-Centric Network

Participant: Thierry Turletti.

Information-Centric Networking (ICN) is a promising solution for most of Internet applications where the content represents the core of the application. However, the proposed solutions for the ICN architecture are associated with many complexities including pervasive caching in the Internet and incompatibility with legacy IP networks, so the deployment of ICN in real networks is still an open problem. In this work, we proposed a backward compatible ICN architecture to address the caching issue in particular. The key idea is implementing edge caching in ICN, using a coalition of end clients and edge servers. Our solution can be deployed in IP networks with HTTP requests. We performed a trace-driven simulation for analyzing PICN benefits using IRCache and Berkeley trace files. The results showed that in average, PICN decreases the latency for 78% and increases the content retrieval speed for 69% compared to a direct download from the original web servers. When comparing PICN with a solution based on central proxy servers, we showed that the hit ratio obtained using a small cache size in each PICN client is almost 14% higher than the hit ratio obtained with a central proxy server using an unlimited cache storage. This work has been published in the IEEE Access journal [14].

6.2.5. Streaming using In-Network Coding and Caching

Participant: Thierry Turletti.

With the rapid growth in high-quality video streaming over the Internet, preserving high-level robustness against data loss and low latency, while maintaining higher data transmission rates, is becoming an increasingly important issue for high-quality real-time delay-sensitive streaming. We have proposed a low latency, low loss streaming mechanism, L4C2, specialized for high-quality delay-sensitive streaming. Using L4C2, nodes in a network estimate the acceptable delay and packet loss probability in their uplinks, aiming at retrieving lost data packets from in-network cache and/or coded data packets using in-network coding within an acceptable delay, by extending the Content-Centric Networking (CCN) approach. Further, L4C2 naturally provides multiple path and multicast technologies to efficiently utilize network resources while sharing network resources fairly with competing data flows by adjusting the video quality as necessary. We validate through comprehensive simulations that L4C2 achieves a high success probability of data transmission considering the acceptable one-way delay and outperforms the existing solution. This work has been presented at the IEEE Infocom conference in May 2017 [23].

6.2.6. Scalable Multicast Service in Software Defined ISP networks

Participants: Hardik Soni, Thierry Turletti, Walid Dabbous.

In the context of the SDN-based multicast mechanisms activity, we designed an architectural solution to provide scalable multicast service in ISP networks. In fact, new applications where anyone can broadcast video are becoming very popular on smartphones. With the advent of high definition video, ISP providers may take the opportunity to propose new high quality broadcast services to their clients. Because of its centralized control plane, Software Defined Networking (SDN) seems an ideal way to deploy such a service in a flexible and bandwidth-efficient way. But deploying large scale multicast services on SDN requires smart group membership management and a bandwidth reservation mechanism to support QoS guarantees that should neither waste bandwidth nor impact too severely best effort traffic. We have proposed a Network Function Virtualization based solution for Software Defined ISP networks to implement scalable multicast group management. We also proposed a routing algorithm called Lazy Load balancing Multicast (L2BM) for sharing the network capacity in a friendly way between guaranteed-bandwidth multicast traffic and best-effort traffic. Our implementation of the framework made on Floodlight controllers and Open vSwitches has been used to study the performance of L2BM. This work has been presented at the IEEE ICC conference [24] in May 2017 and an extended version has been published to the IEEE TNSM journal [13].

6.2.7. Placement of Virtual Network Function Chains in 5G

Participants: Osama Arouk, Thierry Turletti.

We proposed a novel algorithm, namely Multi-Objective Placement (MOP), for the efficient placement of Virtualized Network Function (VNF) chains in future 5G systems. Real datasets are used to evaluate the performance of MOP in terms of acceptance ratio and embedding time when placing the time critical radio access network (RAN) functions as a chain. In addition, we rely on a realistic infrastructure topology to assess the performance of MOP with two main objectives: maximizing the number of base stations that could be embedded in the Cloud and load balancing. The results reveal that the acceptance ratio of embedding RAN functions is only 5% less than the one obtained with the optimal solution for the majority of considered scenarios, with a speedup factor of up to 2000 times. This work has been presented at the IEEE CloudNet conference in September 2017 [17].

6.2.8. P4Bricks: Enabling multiprocessing using Linker-based network data plane architecture

Participants: Hardik Soni, Thierry Turletti, Walid Dabbous.

In order to realize NFV-based multicast service as an add-on network capability without having knowledge of implementation level details of other network functions, we proposed a novel data plane architecture, P4Bricks, for modularized control and packet processing in the network... We propose P4Bricks, a system which aims to deploy and execute multiple independently developed and compiled P4 programs on the same reconfigurable hardware device. P4Bricks is based on a Linker component that merges the programmable parsers/deparsers and restructures the logical pipeline of P4 programs by refactoring, decomposing and scheduling the pipelines' tables. It merges P4 programs according to packet processing semantics (parallel or sequential) specified by the network operator and runs the programs on the stages of the same hardware pipeline, thereby enabling multiprocessing. A paper presenting the initial design of our system with an ongoing implementation and studies P4 language's fundamental constructs facilitating merging of independently written programs was submitted to SOSR and publisehd as a research report [37].

6.2.9. Vehicles as a Mobile Cloud: Modelling, Optimization and Performance Analysis

Participants: Luigi Vigneri, Thrasyvoulos Spyropoulos, Chadi Barakat.

The large diffusion of handheld devices is leading to an exponential growth of the mobile traffic demand which is already overloading the core network. To deal with such a problem, several works suggest to store content (files or videos) in small cells or user equipments. In this work, done in collaboration with Eurecom with the support of the UCN@SOPHIA Labex, we push the idea of caching at the edge a step further, and we propose to use public or private transportation as mobile small cells and caches. In fact, vehicles are widespread in modern cities, and the majority of them could be readily equipped with network connectivity and storage. The adoption of such a mobile cloud, which does not suffer from energy constraints (compared to user equipments), reduces installation and maintenance costs (compared to small cells). In our work, a user can opportunistically download chunks of a requested content from nearby vehicles, and be redirected to the cellular network after a deadline (imposed by the operator) or when her playout buffer empties. The main goal of the work is to suggest to an operator how to optimally replicate content to minimize the load on the core network. Our main contributions are: (i) Modelling: We model the above scenario considering heterogeneous content size, generic mobility and a number of other system parameters. (ii) Optimization: We formulate some optimization problems to calculate allocation policies under different models and constraints. (iii) Performance analysis. We build a MATLAB simulator to validate the theoretical findings through real trace-based simulations. We show that, even with low technology penetration, the proposed caching policies are able to offload more than 50 percent of the mobile traffic demand. The results of this work has been published in several papers, and are currently the subject of two submissions to journals. In particular, in [25] we consider the case of per-chunk caching for video streaming, whereas in [26] we consider the case of Quality of Experience-aware caching of files where the Quality of Experience is modeled as the slowdown in the file download time. A thorough presentation of this work and of our contributions can be found in the PhD thesis of Luigi Vigneri defended in July 2017 and available at [12].

6.3. Experimental Evaluation

6.3.1. The Reproducibility'17 workshop

Participant: Damien Saucez.

Recently, the ACM highlighted that the lack of reproducibility tended to be general in computer science and proposed normalised artifact reviewing and badging definitions [0] with the hope that the various ACM communities would perform artifact reviews based on these definition. We organized a special workshop on reproducibility in conjunction with the ACM SIGCOMM 2017 conference to produce a set of recommendations on how to assess the reproducibility of research published in ACM SIGCOMM'related conferences and journals and ways to promote reproducibility. The proceedings of the workshop is available in [27] and we have produced a set of recommendations to the community in [34] with the following conclusions:

The workshop pointed out that there are several hurdles concerning reproducibility, namely the absence of incentives and the bad habit that our community has grown accustomed to. This is evident in the current typical review process which is not adapted to handle reproducibility. Furthermore, there is no general way to share and preserve artifacts (and related documentation), every author does it in their own way.

The workshop focused on the two most important points to be tackled, namely, i) how to provide incentives for reproducible papers and ii) how to share artifacts.

For the first, a promising approach is to put in place a Reproducibility Committee, which will run in parallel with the normal Technical Program Committee of conferences and workshops, which will assess the level of reproducibility of papers accepted for publication by the TPC. Such approach will solve some of the privacy and anonymity issues while reducing the volume of work for the reviewers that volunteer in assessing the reproducibility level.

For the second, a gradual approach has been suggested. The ACM digital library has been suggested as place to start sharing artifacts, which will be also identified via a DOI number. Beside the artifact itself it is important to share all of the meta-information necessary to actually reproduce prior work, as well as a way to provide feedback in order to make the community learn which meta-information is actually important and build guidelines on how to provide such information.

6.3.2. *Towards Realistic Software-Defined Wireless Networking Experiments*

Participants: Mohamed Naoufal Mahfoudi, Walid Dabbous, Thierry Turletti.
Software-Defined Wireless Networking (SDWN) is an emerging approach based on decoupling radio control functions from the radio data plane through programmatic interfaces. Despite diverse ongoing efforts to realize the vision of SDWN, many questions remain open from multiple perspectives such as means to rapid prototype and experiment candidate software solutions applicable to real world deployments. To this end, emulation of SDWN has the potential to boost research and development efforts by re-using existing protocol and application stacks while mimicking the behavior of real wireless networks. In this work, we provided an in-depth discussion on that matter focusing on the Mininet-WiFi emulator design to fill a gap in the experimental platform space. We showcased the applicability of our emulator in an SDN wireless context by illustrating the support of a number of use cases aiming to address the question on how far we can go in realistic SDWN experiments, including comparisons to the results obtained in a wireless testbed. Finally, we discussed the ability to replay packet-level and radio signal traces captured in the real testbed towards a virtual yet realistic emulation environment in support of SDWN research. This works has been published in a Special Issue on Software Defined Wireless Networks of the Computer Journal [15].

6.3.3. *ORION: Orientation Estimation Using Commodity Wi-Fi*

Participants: Mohamed Naoufal Mahfoudi, Thierry Turletti, Thierry Parmentelat, Walid Dabbous.
With MIMO, Wi-Fi led the way to the adoption of antenna array signal processing techniques for finegrained localization using commodity hardware. These techniques, previously exclusive to specific domains of applications, open the road to reach beyond localization, and now allow to consider estimating the device's orientation in space, that once required other sources of information. Wi-Fi's popularity and the availability of metrics related to channel propagation (CSI), makes it a candidate readily available for experimentation. We have recently proposed the ORION system to estimate the orientation (heading and yaw) of a MIMO Wi-Fi equipped object, relying on a joint estimation of the angle of arrival and the angle of departure. Although

[0] Artifact Review and Badging, https://www.acm.org/publications/policies/artifact-review-badging, December 2017

the CSI's phase data is plagued by several phase inconsistencies, we demonstrate that an appropriate phase compensation strategy significantly improves estimation accuracy. By feeding the estimation to a Kalman filter, we further improve the overall system accuracy, and lay the ground for an efficient tracking. Our technique allows estimating orientations within high precision. The results of the study were presented at a IEEE specialized workshop on Network Localization on Navigation [22].

6.3.4. Lessons Learned while Trying to Reproduce the OpenRF Experiment

Participants: Mohamed Naoufal Mahfoudi, Thierry Turletti, Thierry Parmentelat, Walid Dabbous.
Evaluating and comparing performance of wireless systems, like for any other scientific area, requires the ability to reproduce experimental results. In this work, we described the specific issues that we encountered when focusing on reproducing the experiments described in a paper related to wireless systems. We selected the OpenRF paper published in SIGCOMM 2013, a very interesting research work allowing to perform beamforming on commodity WiFi devices. We illustrated how reproducibility is strongly dependent on the used hardware, and why an extensive knowledge of the used hardware and its design is necessary. On the basis of this experience, we proposed some recommendations and lessons for the design of reproducible wireless experiments. This work has been presented at the ACM SIGCOMM 2017 Reproducibility Workshop in August 2017 [21].

6.3.5. Deploying a 5G network in less than 5 minutes

Participants: Mohamed Naoufal Mahfoudi, Thierry Parmentelat, Thierry Turletti, Walid Dabbous.
We proposed a demonstration run on R2lab, an anechoic chamber located at Inria Sophia Antipolis, France. This demonstration consists in deploying a standalone 5G network in less than 5 minutes. All the network components (base station, subscriber management, serving and packet gateways, network traac analyzers) were run automatically using the nepi-ng experiment orchestration tool. Download and upload performance to the Internet from a commercial phone located in the anechoic chamber are shown. This demo has been presented at the ACM SIGCOMM conference in August 2017 [33].

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation

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

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

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

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies.

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

Grand Challenge. Since 2011, the ACM International Conference on Distributed Event-based Systems (DEBS) launched the Grand Challenge series to increase the focus on these systems as well as provide common benchmarks to evaluate and compare them. The ACM DEBS 2017 Grand Challenge focused on (soft) real-time anomaly detection in manufacturing equipment. To handle continuous monitoring, each machine is fitted with a vast array of sensors, either digital or analog. These sensors provide periodic measurements, which are sent to a monitoring base station. The latter receives then a large collection of observations. Analyzing in an efficient and accurate way, this very-high-rate – and potentially massive – stream of events is the core of the Grand Challenge. Although, the analysis of a massive amount of sensor reading requires an on-line analytics pipeline that deals with linked-data, clustering as well as a Markov model training and querying. The FlinkMan system [62] proposes a solution to the 2017 Grand Challenge, making use of a publicly available streaming engine and thus offering a generic solution that is not specially tailored for this or for another challenge. We offer an efficient solution that maximally utilizes available cores, balances the load among the cores, and avoids to the extent possible tasks such as garbage collection that are only indirectly related to the task at hand.

Health big data processing. Sharing and exploiting efficiently Health Big Data (HBD) lead to tackle great challenges: data protection and governance taking into account legal, ethical and deontological aspects which enables a trust, transparent and win-to-win relationship between researchers, citizen and data providers. Lack of interoperability: data are compartmentalized and are so syntactically and semantically heterogeneous. Variable data quality with a great impact on data management and statistical analysis. The objective of the INSHARE project [41] is to explore, through an experimental proof of concept, how recent technologies could overcome such issues. It aims at demonstrating the feasibility and the added value of an IT platform based on CDW, dedicated to collaborative HBD sharing for medical research.

The consortium includes 6 data providers: 2 academic hospitals, the SNIIRAM (the French national reimbursement database) and 3 national or regional registries. The platform is designed following a three steps approach: (1) to analyze use cases, needs and requirements, (2) to define data sharing governance and secure access to the platform, (3) to define the platform specifications. Three use cases (healthcare trajectory analysis, epidemiological registry enrichment, signal detection) were analyzed to design the platform corresponding to five studies and using eleven data sources. The governance was derived from the SCANNER model and adapted to data sharing. As a result, the platform architecture integrates the following tools and services: data repository and hosting, semantic integration services, data processing, aggregate computing, data quality and integrity monitoring, id linking, multi-source query builder, visualization and data export services, data governance, study management service and security including data watermarking.

Throughput prediction in cellular networks. Downlink data rates can vary significantly in cellular networks, with a potentially non-negligible effect on the user experience. Content providers address this problem by using different representations (*e.g.*, picture resolution, video resolution and rate) of the same content and by switching among these based on measurements collected during the connection. If it were possible to know the achievable data rate before the connection establishment, content providers could choose the most appropriate representation from the very beginning. We have conducted a measurement campaign involving 60 users connected to a production network in France, to determine whether it is possible to predict the achievable data rate using measurements collected, before establishing the connection to the content provider, on the operator's network and on the mobile node. We show that it is indeed possible to exploit these measurements to predict, with a reasonable accuracy, the achievable data rate [53].

Population protocol model. We consider in [50] a large system populated by n anonymous nodes that communicate through asynchronous and pairwise interactions. The aim of these interactions is, for each node, to converge toward a global property of the system that depends on the initial state of the nodes. We focus on both the counting and proportion problems. We show that for any $\delta \in (0,1)$, the number of interactions needed per node to converge is $O(\ln(n/\delta))$ with probability at least $1-\delta$. We also prove that each node can determine, with any high probability, the proportion of nodes that initially started in a given state without knowing the number of nodes in the system. This work provides a precise analysis of the convergence bounds, and shows that using the 4-norm is very effective to derive useful bounds.

The context of [71] is the well studied dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called *rumor mongering*, and then *rumor spreading* has mainly been investigated in the synchronous model, which relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In this paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where, at random times, nodes successively interact by pairs exchanging their information on the rumor. In a previous paper, we performed a study of the total number of interactions needed for all the nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow us to compare different protocols, we provided a thorough analysis of the distribution of this total number of interactions together with its asymptotic behavior. In this paper we extend this discrete-time analysis by solving a conjecture proposed previously and we consider the continuous-time case, where a Poisson process is associated with each node to determine the instants at which interactions occur. The rumor spreading time is thus more realistic since it is the time needed for all the nodes of the network to discover the rumor. Once again, as most of the existing results involve huge constants, we provide a tight bound and equivalent of the complementary distribution of the rumor spreading time. We also give the exact asymptotic behavior of the complementary distribution of the rumor spreading time around its expected value when the number of nodes tends to infinity.

Transient analysis. Last, in two keynotes ([35] and [34]), we described part of our previous analytical results concerning the transient behavior of well-structured Markov processes, mainly on performance models (queueing systems), and we presented recent new results that extend those initial findings. The heart of the novelties lie on an extension of the concept of duality proposed by Anderson in [73] that we call pseudo-dual. The dual of a stochastic process needs strong monotonicity conditions to exist. Our proposed pseudo-dual always exist, and is directly defined on a linear system of differential equations with constant coefficients, that can be, in particular, the system of Chapman-Kolmogorov equations corresponding to a Markov process, but not necessarily. This allows, for instance, to prove the validity of closed-forms expressions of the transient distribution of a Markov process in cases where the dual doesn't exist. The keynote [35] was presented to a public oriented toward differential equations and dynamical systems; [34] has a more modeling flavour. A paper is under preparation with the technical details.

7.2. Distributed deep learning on edge-devices

Participants: Corentin Hardy, Gerardo Rubino, Bruno Sericola

A large portion of data mining and analytic services use modern machine learning techniques, such as deep learning. The state-of-the-art results related to deep learning come at the price of an intensive use of computing resources. The leading frameworks (*e.g.*, TensorFlow) are executed on GPUs or on high-end servers in data centers. On the other end, there is a proliferation of personal devices with possibly free CPU cycles; this can enable services to run in users' homes, embedding machine learning operations. In [66] and [43], we ask the following question: *Is distributed deep learning computation on WAN connected devices feasible, in spite of the traffic caused by learning tasks*? We show that such a setup rises some important challenges, most notably the ingress traffic that the servers hosting the up-to-date model have to sustain. In order to reduce this stress, we propose *AdaComp*, a novel algorithm for compressing worker updates to the model on the server. Applicable to stochastic gradient descent based approaches, it combines efficient gradient selection and learning rate modulation. We experiment and measure the impact of compression, device heterogeneity and reliability on the accuracy of learned models, with an emulator platform that embeds TensorFlow into Linux containers. We report a reduction of the total amount of data sent by workers to the server by two order of magnitude (*e.g.*, 191-fold reduction for a convolutional network on the MNIST dataset), when compared to a standard asynchronous stochastic gradient descent, while preserving model accuracy. The extension of the AdaComp algorithm to Random Neural Networks started with the introduction of Random Neural Layers, see [65].

7.3. Network Economics

Participants: Bruno Tuffin, Patrick Maillé, Pierre L'Ecuyer

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team. The whole problem of network economics, from theory to practice, describing all issues and challenges, is described in our book [7].

Roaming. In October 2015, the European parliament has decided to forbid roaming charges among EU mobile phone users, starting June 2017, as a first step toward the unification of the European digital market. We have investigated the consequences of such a measure from an economic perspective. In [47], we analyze the effect of the willingness-to-pay heterogeneity among users (also due to wealth heterogeneity), and the fact that the roaming behavior is positively correlated with wealth. Our analysis suggests that imposing free roaming degrades the revenues of the operator but can also deter some users from subscribing; hence we conclude that such (apparently beneficial) regulatory decisions must be taken with care. In [47], we particularly focus on the strategies on transit payments between ISPs in different countries. We highlight that scrutiny is also required since, depending on parameters, consumer surplus or subscription penetration are not necessarily maximized if free roaming is enforced.

Network neutrality. Most of our activity has been devoted to the vivid network neutrality debate, going beyond the traditional for or against neutrality, and trying to tackle it from different angles.

Network neutrality has been a very sensitive topic of discussion all over the world. In the keynote talk [59], we first introduce the elements of the debate and how the problem can be modeled and analyzed through game theory. With an Internet ecosystem much more complex now than the simple delivery chain Content-ISP-User, we highlight, in a second step, how neutrality principles can be bypassed in various ways without violating the rules currently evoked in the debate, for example via Content Delivery Networks (CDNs), or via search engines which can affect the visibility and accessibility of content. We describe some other grey zones requiring to be dealt with and spend some time on discussing the (potential) implications for clouds.

The impact of CDNs on the debate has been detailed in [18]. Content Delivery Networks (CDN) have become key telecommunication actors. They contribute to improve significantly the quality of services delivering content to end users. However, their impact on the ecosystem raises concerns about their fairness, and therefore the question of their inclusion in the neutrality debates becomes relevant. We analyze the impact of a revenue-maximizing CDN on some other major actors, namely, the end-users, the network operators and the content providers, at comparing the outcome with that of a fair behavior, and at providing tools to investigate whether some regulation should be introduced.We present a mathematical model and show that there exists a unique optimal revenue-maximizing policy for a CDN actor, in terms of dimensioning and allocation of its storage capacity, and depending on parameters such as prices for service/transport/storage. Numerical experiments are then performed with both synthetic data and real traces obtained from a major Video-on-Demand provider. In addition, using the real traces, we compare the revenue-based policy with policies based on several fairness criteria.

Network neutrality is often advocated by content providers, stressing that side payments to Internet Service Providers would hinder innovation. However, we also observe some content providers actually paying those fees. In [24], we intend to explain such behaviors through economic modeling, illustrating how side payments can be a way for an incumbent content provider to prevent new competitors from entering the market. We investigate the conditions under which the incumbent can benefit from such a barrier-to-entry, and the consequences of that strategic behavior on the other actors: content providers, users, and the Internet Service Provider. We also describe how the Nash bargaining solution concept can be used to determine the side payment.

Similarly, major content/service providers are publishing grades they give to ISPs about the quality of delivery of their content. The goal is to inform customers about the “best” ISPs. But this could be an incentive for, or even a pressure on, ISPs to differentiate service and provide a better quality to those big content providers in order to be more attractive. Instead of the traditional vision of ISPs pressing content providers, we face here the opposite situation, still possibly at the expense of small content providers though. We design in [48] a model describing the various actors and their strategies, analyzes it using non-cooperative game theory tools, and quantifies the impact of those advertised grades with respect to the situation where no grade is published. We illustrate that a non-neutral behavior, differentiating traffic, is not leading to a desirable situation.

Sponsored data. With wireless sponsored data, a third party, content or service provider, can pay for some of your data traffic so that it is not counted in your plan's monthly cap. This type of behavior is currently under scrutiny, with telecommunication regulators wondering if it could be applied to prevent competitors from entering the market, and what the impact on all telecommunication actors can be. To answer those questions, we design and analyze in [69] a model where a Content Provider (CP) can choose the proportion of data to sponsor and a level of advertisement to get a return on investment, and several Internet Service Providers (ISPs) in competition. We distinguish three scenarios: no sponsoring, the same sponsoring to all users, and a different sponsoring depending on the ISP you have subscribed to. This last possibility may particularly be considered an infringement of the network neutrality principle. We see that sponsoring can be beneficial to users and ISPs, especially with identical sponsoring. We also discuss the impact of zero-rating where an ISP offers free data to a CP to attract more customers, of and vertical integration where a CP and an ISP are the same company.

Online platforms and search engines. The search neutrality debate is about whether search engines should or should not be allowed to uprank certain results among the organic content matching a query. This debate is related to that of network neutrality, which focuses on whether all bytes being transmitted through the Internet should be treated equally. In a previous paper, we had formulated a model that formalizes this question and characterized an optimal ranking policy for a search engine. The model relies on the trade-off between short-term revenues, captured by the benefits of highly-paying results, and long-term revenues which can increase by providing users with more relevant results to minimize churn. In [21], we apply that model to investigate the relations between search neutrality and innovation. We illustrate through a simple setting and computer simulations that a revenue-maximizing search engine may indeed deter innovation at the content level. Our simple setting obviously simplifies reality, but this has the advantage of providing better insights on how optimization by some actors impacts other actors.

Sponsored auctions. Advertisement in dedicated webpage spaces or in search engines sponsored slots is usually sold using auctions, with a payment rule that is either per impression or per click. But advertisers can be both sensitive to being viewed (brand awareness effect) and being clicked (conversion into sales). In [23], we generalize the auction mechanism by including both pricing components: the advertisers are charged when their ad is displayed, and pay an additional price if the ad is clicked. Applying the results for Vickrey-Clarke-Groves (VCG) auctions, we show how to compute payments to ensure incentive compatibility from advertisers as well as maximize the total value extracted from the advertisement slot(s). We provide tight upper bounds for the loss of efficiency due to applying only pay-per-click (or pay-per-view) pricing instead of our scheme. Those bounds depend on the joint distribution of advertisement visibility and population likelihood to click on ads, and can help identify situations where our mechanism yields significant improvements. We also describe how the commonly used generalized second price (GSP) auction can be extended to this context.

7.4. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino, Pierre L'Ecuyer

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types. We have published an introduction to Monte Carlo methods on Insterstices, including animations https://interstices.info/jcms/int_69164/la-simulation-de-monte-carlo.

Rare event simulation. The mean time to failure (MTTF) of a stochastic system is often estimated by simulation. One natural estimator, which we call the direct estimator, simply averages independent and identically distributed copies of simulated times to failure. When the system is regenerative, an alternative approach is based on a ratio representation of the MTTF. The purpose of [42] is to compare the two estimators. We first analyze them in the setting of crude simulation (i.e., no importance sampling), showing that they are actually asymptotically identical in a rare-event context. The two crude estimators are inefficient in different but closely related ways: the direct estimator requires a large computational time because times to failure often include many transitions, whereas the ratio estimator entails estimating a rare-event probability. We then

discuss the two approaches when employing importance sampling; for highly reliable Markovian systems, we show that using a ratio estimator is advised.

Another problem studied in [40] is the estimation of the tail of the distribution of the sum of correlated log-normal random variables. While a number of theoretically efficient estimators have been proposed for this setting, using a few numerical examples we illustrate that these published proposals may not always be useful in practical simulations. As a remedy to this defect, we propose a new estimator and we demonstrate that, not only is our novel estimator theoretically efficient, but, more importantly, its practical performance is significantly better than that of its competitors.

Random variable generation. Random number generators were invented before there were symbols for writing numbers, and long before mechanical and electronic computers. All major civilizations through the ages found the urge to make random selections, for various reasons. Today, random number generators, particularly on computers, are an important (although often hidden) ingredient in human activity. In the invited paper [32], we give a historical account on the design, implementation, and testing of uniform random number generators used for simulation.

We study in [68] the lattice structure of random number generators of the specific MIXMAX family, a class of matrix linear congruential generators that produce a vector of random numbers at each step. These generators were initially proposed and justified as close approximations to certain ergodic dynamical systems having the Kolmogorov K-mixing property, which implies a chaotic (fast-mixing) behavior. But for a K-mixing system, the matrix must have irrational entries, whereas for the MIXMAX it has only integer entries. As a result, the MIXMAX has a lattice structure just like linear congruential and multiple recursive generators. We study this lattice structure for vectors of successive and non-successive output values in various dimensions. We show in particular that for coordinates at specific lags not too far apart, in three dimensions, all the nonzero points lie in only two hyperplanes. This is reminiscent of the behavior of lagged-Fibonacci and AWC/SWB generators. And even if we skip the output coordinates involved in this bad structure, other highly structured projections often remain, depending on the choice of parameters.

Quasi-Monte Carlo (QMC). In [5], which appeared in 2017, we survey basic ideas and results on randomized quasi-Monte Carlo (RQMC) methods, discuss their practical aspects, and give numerical illustrations. RQMC can improve accuracy compared with standard Monte Carlo (MC) when estimating an integral interpreted as a mathematical expectation. RQMC estimators are unbiased and their variance converges at a faster rate (under certain conditions) than MC estimators, as a function of the sample size. Variants of RQMC also work for the simulation of Markov chains, for function approximation and optimization, for solving partial differential equations, etc. In this introductory survey, we look at how RQMC point sets and sequences are constructed, how we measure their uniformity, why they can work for high-dimensional integrals, and how can they work when simulating Markov chains over a large number of steps.

General presentations. Finally, in two general presentations, we described state-of-the-art technologies available to deal with rare events by means of Monte Carlo techniques, including several methods produced inside Dionysos. In the tutorial [33], we gave an overview of the field, with a focus on dependability analysis applications. The keynote [36] described specific procedures taken from our monograph [72], that were adapted to the needs of the micro-simulation community.

7.5. Wireless Networks

Participants: Yue Li, Imad Alawe, Quang Pham, Patrick Maillé, Yassine Hadjadj-Aoul, César Viho, Gerardo Rubino

Mobile wireless networks' improvements. Software Defined Networking (SDN) is one of the key enablers for evolving mobile network architecture towards 5G. SDN involves the separation of control and data plane functions, which leads, in the context of 5G, to consider the separation of the control and data plane functions of the different gateways of the Evolved Packet Core (EPC), namely Serving and Packet data Gateways (S and P-GW). Indeed, the envisioned solutions propose to separate the S/P-GW into two entities: the S/P-GW-C, which integrates the control plane functions and the S/P-GW-U that handles the User Equipment (UE)

data plane traffic. There are two major approaches to create and update user plane forwarding rules for such a partition: (i) considering an SDN controller for the S/P-GW-C (SDNEPC) or (ii) using a direct specific interface to control the S/P-GW-U (enhancedEPC). In [38], we evaluate, using a testbed, those two visions against the classical virtual EPC (vEPC), where all the elements of the EPC are virtualized. Besides evaluating the capacity of the vEPC to manage and scale to UE requests, we compare the performances of the solutions in terms of the time needed to create the user data plane. The obtained results allow drawing several remarks, which may help to dimension the vEPC's components as well as to improve the S/P-GW-U management procedure.

One of the requirements of 5G is to support a massive number of connected devices, considering many use-cases such as IoT and massive Machine Type Communication (MTC). While this represents an interesting opportunity for operators to grow their business, it will need new mechanisms to scale and manage the envisioned high number of devices and their generated traffic. Particularity, the signaling traffic, which will overload the 5G core Network Function (NF) in charge of authentication and mobility, namely Access and Mobility Management Function (AMF). The objective of [37] is to provide an algorithm based on Control Theory allowing: (i) to equilibrate the load on the AMF instances in order to maintain an optimal response time with limited computing latency; (ii) to scale out or in the AMF instance (using NFV techniques) depending on the network load to save energy and avoid wasting resources. Obtained results indicate the superiority of our algorithm in ensuring fair load balancing while scaling dynamically with the traffic load. In [64] we are going further by using new advances on machine learning, and more specifically Recurrent Neural Networks (RNN), to predict accurately the arrival traffic pattern of devices. The main objective of the proposed approach is to early react to congestion by pro-actively scaling the AMF VNF in a way to absorb such congestion while respecting the traffic constraints.

Energy consumption improvements. Recently in cellular networks, the focus has been moved to seeking ways to increase the energy efficiency by better adapting to the existing users behaviors. In [17], we are going a step further in studying a new type of disruptive service by trying to answer the question "What are the potential energy efficiency gains if some of the users are willing to tolerate delays?". We present an analytical model of the energy usage of LTE base stations, which provides lower bounds of the possible energy gains under a decentralized, noncooperative setup. The model is analyzed in six different scenarios (such as micro-macro cell interaction and coverage redundancy) for varying traffic and user-tolerable delays. We show that it is possible to reduce the power consumption by up to 30%.

Computation offloading in mobile network. Mobile edge computing (MEC) emerges as a promising paradigm that extends the cloud computing to the edge of pervasive radio access networks, in near vicinity to mobile users, reducing drastically the latency of end-to-end access to computing resources. Moreover, MEC enables the access to up-to-date information on users' network quality via the radio network information service (RNIS) application programming interface (API), allowing to build novel applications tailored to users' context. In [25] and [49], we present a novel framework for offloading computation tasks, from a user device to a server hosted in the mobile edge (ME) with highest CPU availability. Besides taking advantage of the proximity of the MEC server, the main innovation of the proposed solution is to rely on the RNIS API to drive the user equipment (UE) decision to offload or not computing tasks for a given application. The contributions are twofold. First, we propose the design of an application hosted in the ME, which estimates the current value of the round trip time (RTT) between the UE and the ME, according to radio quality indicators available through RNIS API, and provides it to the UE. Second, we present a novel computation algorithm which, based on the estimated RTT coupled with other parameters (e.g., energy consumption), decide when to offload UE's applications computing tasks to the MEC server. The effectiveness of the proposed framework is demonstrated via testbed experiments featuring a face recognition application.

Services improvement in wireless heterogeneous networks. With the rapid growth of HTTP-based Adaptive Streaming (HAS) multimedia video services on the Internet, improving the Quality of Experience (QoE) of video delivery will be highly requested in wireless heterogeneous networks. Various access technologies such as 3G/LTE and Wi-Fi with overlapping coverage is the main characteristic of network heterogeneity. Since contemporary mobile devices are usually equipped with multiple radio interfaces, mobile users are enabled to

utilize multiple access links simultaneously for additional capacity or reliability. However, network and video quality selection can have notable impact on the QoE of DASH clients facing the video service's requirements, the wireless channel profiles and the costs of the different links. In this context, the emerging Multi-access Edge Computing (MEC) standard gives new opportunities to improve DASH performance, by moving IT and cloud computing capabilities down to the edge of the mobile network. In [45], we propose a MEC-assisted architecture for improving the performance of DASH-based streaming, a standard implementation of a HAS framework in wireless heterogeneous networks. With the proposed algorithm running as a MEC service, the overall QoE and fairness of DASH clients are improved in a real time manner in case of network congestion.

QoE aware routing in wireless networks. This year we continued our research on QoE-based optimization routing for wireless mesh networks. The difficulties of the problem are analyzed and centralized and decentralized algorithms are proposed. The quality of the solution, the computational complexity of the proposed algorithm, and the fairness are our main concerns. Several centralized approximation algorithms have been already proposed in order to address the complexity and the quality of possible solutions. This year, we focused mainly on distributed algorithm to complement of the existing centralized algorithms. We propose decentralized heuristic algorithms based on the well-known Optimized Link-State Routing (OLSR) protocol. Control packets of OLSR are modified so as to be able to convey QoE-related information. The routing algorithm chooses the paths heuristically. After that, we studied message passing algorithms in order to find near optimal routing solutions in cooperative distributed networks. These algorithms have been published in [27], [13].

Sensors networks. In the literature, it is common to consider that sensor nodes in a clustered-based event-driven Wireless Sensor Network (WSN) use a Carrier Sense Multiple Access (CSMA) protocol with a fixed transmission probability to control data transmission. However, due to the highly variable environment in these networks, a fixed transmission probability may lead to a significant amount of extra energy consumption. In view of this, three different transmission probability strategies for event-driven WSNs were studied in [51]: the optimal one, the "fixed" approach and a third "adaptive" method. As expected, the optimum strategy achieves the best results in terms of energy consumption but its implementation in a practical system is not feasible. The commonly used fixed transmission strategy (the probability for any node to attempt transmission is a constant) is the simplest approach but it does not adapt to changes in the system's conditions and achieves the worst performance. In the paper, we find that our proposed adaptive transmission strategy, where that probability is changed depending on specific conditions and in a very precise way, is pretty easy to implement and achieves results very close to the optimal method. The three strategies are analyzed in terms of energy consumption but also regarding the cluster formation latency. In [28], we also investigate cluster head selection schemes. Specifically, we consider two intelligent schemes based on the fuzzy C-means and k-medoids algorithms, and a random selection with no intelligence. We show that the use of intelligent schemes greatly improves the performance of the system, but their use entails higher complexity and some selection delay. The main performance metrics considered in this work are energy consumption, successful transmission probability and cluster formation latency. As an additional feature of this work, we study the effect of errors in the wireless channel and the impact on the performance of the system under the different considered transmission probability schemes.

Transmission delay, throughput and energy are also important criteria to consider in wireless sensor networks (WSNs). The IEEE 802.15.4 standard was conceived with the objective of reducing resource's consumption in both WSNs and Personal Area Networks (WPANs). In such networks, the slotted CSMA/CA still occupies a prominent place as a channel control access mechanism with its inherent simplicity and reduced complexity. In [26], we propose to introduce a network allocation vector (NAV) to reduce energy consumption and collisions in IEEE 802.15.4 networks. A Markov chain-based analytical model of the fragmentation mechanism, in a saturated traffic, is given as well as a model of the energy consumption using the NAV mechanism. The obtained results show that the fragmentation technique improves at the same time the throughput, the access delay and the bandwidth occupation. They also show that using the NAV allows reducing significantly the energy consumption when applying the fragmentation technique in slotted CSMA/CA under saturated traffic conditions.

7.6. Optical Networks

Participants: Nicolás Jara, Gerardo Rubino

The rapid increase in demand for bandwidth in communication networks has caused a growth in the use of technologies based on WDM optical infrastructures. Nevertheless, in this last decade many researchers have recognized a "Capacity Crunch" associated with this technology, a transmission capacity limit on optical fibers, that is close to be reached pretty soon. This situation claims for an evolution on the currently used WDM optical architectures, in order to satisfy this relentless exponential growth in bandwidth demand. Following this trend, research started to examine in some detail specific aspects of the present functioning, and in particular, the way these networks are operated. Currently, optical networks are operated statically, but this is known to be inefficient in the usage of network resources, and with the previously mentioned upcoming risk of capacity collapse, it is of pressing matter to upgrade it. To this purpose, several proposals have been addressed and researched so far. Among these solutions, dynamic optical networks is the one closest to be implemented, but it has not been considered yet since the network cost savings are not enough to convince enterprises. This has been the focus of our research effort in the area.

The design of dynamic optical networks decomposes into different tasks, where the engineers must basically organize the way the main system's resources are used, minimizing the design and operation costs and respecting critical performance constraints. These tasks must guarantee certain level of quality of service (QoS) pre-established in the Service Level Agreement. In order to provide a proper quality of service measurement, we propose a new fast and accurate analytical method to evaluate the blocking probability that is at the heart of the path toward solving all the mentioned design problems. Blocking probability is the main QoS metric considered in the field. This work has been done in [20], where an analytical procedure has been proposed that combines efficiency and accuracy.

Next, the different tasks that must be addressed to find a good global design have been addressed in [19]. These are: which wavelength is going to be used by each user (the Wavelength Assignment Problem), how many wavelengths will be needed on each network link (the Wavelength Dimensioning Problem), and which set of paths enabling each network user to transmit (known as the Routing Problem) are to be established in order to minimize costs and to deal with link failures when the network is operating (this is the Fault Tolerance Problem). Two types of innovations and presented in this last paper. First, each of the problems receives a solution shown to be highly efficient. Second, and this is also new, we solve all the design problems simultaneously, using a single global algorithm (the usual way is to isolate them and to solve them one at a time, in a specific order). This work may provide a strategy to finally achieve sufficient cost savings, and thus, to contribute to make the decision to migrate from static to dynamic resource allocation easier. A preliminary version of a part of these results was presented previously in [44].

7.7. Future networks and architectures

Participants: Jean-Michel Sanner, Hamza Ben Ammar, Louiza Yala, Yassine Hadjadj-Aoul, Gerardo Rubino

SDN and NFV placement. Mastering the increasing complexity of current and future networks, while reducing the operational and investments costs, is one of the major challenges faced by network operators (NOs). This explains in large part the recent enthusiasm of NOs towards Software Defined Networking (SDN) and Network Function Virtualization (NFV). Indeed, on the one hand, SDN makes it possible to get rid of the control plane distribution complexity, by centralizing it logically, while allowing its programmability. On the other hand, the NFV allows virtualizing the network functions, which considerably facilitates the deployment and the orchestration of the network resources. Providing a carrier grade network involves, however, several requirements such as providing a robust network meeting the constraints of the supported services. In order to achieve this objective, it is clearly necessary to scale network functions while placing them strategically in a way to guarantee the system's responsiveness.

The placement in TelCo networks are generally multi-objective and multi-constrained problems. The solutions proposed in the literature usually model the placement problem by providing a mixed integer linear program (MILP). Their performances are, however, quickly limited for large sized networks, due to the significant increase in the computational delays. In order to avoid the inherent complexity of optimal approaches and the

lack of flexibility of heuristics, we propose in [54] a genetic algorithm designed from the NSGA II framework that aims to deal with the controller placement problem. Genetic algorithms can be both multi-objective, multi-constraints and can be designed to be implemented in parallel. They constitute a real opportunity to find good solutions to this category of problems. Furthermore, the proposed algorithm can be easily adapted to manage dynamic placements scenarios. In [55], our main focus were devoted to maximize the clusters average connectivity and to balance the control's load between clusters, in a way to improve the networks' reliability.

We focus, in [60], on the problem of optimal computing resource allocation and placement for the provision of a virtualized Content Delivery Network (CDN) service over a telecom operator's Network Functions Virtualization (NFV) infrastructure. Starting from a Quality of Experience (QoE)-driven decision on the necessary amount of CPU resources to allocate in order to satisfy a virtual CDN deployment request with QoE guarantees, we address the problem of distributing these resources to virtual machines and placing the latter to physical hosts, optimizing for the conflicting objectives of management cost and service availability, while respecting physical capacity, availability and cost constraints. We present a multi-objective optimization problem formulation, and provide efficient algorithms to solve it by relaxing some of the original problem's assumptions. Numerical results demonstrate how our solutions address the trade-off between service availability and cost, and show the benefits of our approach compared with resource placement algorithms which do not take this trade-off into account.

Real-time NFV placement in edge cloud. Sometimes, the placement of NFV can not be planned in advance and therefore requires real-time placement as requests arrive. The placement is particularly challenging with the recent development of geographically distributed mini data centers, also referred to as cloudlets, at the edge of the network (i.e., typically at Points of Presence (PoPs) level). These edge data centers have rather small capacities in terms of storage, computing and networking resources, when compared with the huge centralized data centers deployed today.

All these radical changes in NOs' infrastructures raise many new issues (especially in terms of resource allocation), which so far have not been considered in the cloud literature. Traditionally, resources in cloud platforms are considered as to be infinite and request blocking is most of the time ignored when evaluating resources' allocation algorithms, precisely because of this infinite capacity assumption. However, if we assume that the NO's infrastructure will very likely be composed of small data centers with limited capacities, and deployed at the edge of network, the congestion of such a system may occur, notably if the demand is sufficiently high and exceeds what the infrastructure can handle at a given time.

We proposed in [57] an analytical model for the blocking analysis in a multidimensional cloud system, which was validated using discrete events' simulations. Besides, we conducted a comparative analysis of the most popular placement's strategies. The proposed model, as well as the comparative study, reveal practical insights into the performance evaluation of resource allocation and capacity planning for distributed edge cloud with limited capacities.

In [58] we set design principles of future distributed edge clouds in order to meet application requirements. We precisely introduce a costless distributed resource allocation algorithm, named *CLOSE*, which considers local information only. We compare via simulations the performance of *CLOSE* against those obtained by using mechanisms proposed in the literature, notably the Tricircle project within OpenStack. It turns out that the proposed distributed algorithm yields better performance while requiring less overhead.

In the context of the Open Network Automation Platform (ONAP), we develop in [56] a resource allocation strategy for deploying Virtualized Network Functions (VNFs) on distributed data centers. For this purpose, we rely on a three-level data center hierarchy exploiting co-location facilities available within Main and Core Central Offices. We precisely propose an active VNFs' placement strategy, which dynamically offloads requests on the basis of the load observed within a data center. We compare via simulations the performance of the proposed solution against mechanisms so far proposed in the literature, notably the centralized approach of the multi-site project within OpenStack, currently adopted by ONAP. Our algorithm yields better performance in terms of both data center occupancy and overhead. Furthermore, it allows extending the applicability of ONAP in the context of distributed cloud, without requiring any modification.

Content Centric Networking. Content-Centric Networking (CCN) has been proposed to address the challenges raised by the Internet usage evolution over the last years. One key feature provided by CCN to improve the efficiency of content delivery is the in-network caching, which has major impact on the system performance. In order to improve caching effectiveness in such systems, the study of the functioning of CCN in-network storage must go deeper. In [39], we propose MACS, a Markov chain-based Approximation of CCN caching Systems. We start initially by modeling a single cache node. Then, we extend our model to the case of multiple nodes. A closed-form expression is then derived to define the cache hit probability of each content in the caching system. We compare the results of MACS to those obtained with simulations. The conducted experiments show clearly the accuracy of our model in estimating the cache hit performance of the system.

In [16], we present the design and implementation of a Content-Delivery-Network-as-a-Service (CDNaaS) architecture, which allows a telecom operator to open up its cloud infrastructure for content providers to deploy virtual CDN instances on demand, at regions where the operator has presence. Using northbound REST APIs, content providers can express performance requirements and demand specifications, which are translated into an appropriate service placement on the underlying cloud substrate. Our architecture is extensible, supporting various different CDN flavors, and, in turn, different schemes for cloud resource allocation and management. In order to decide on the latter in an optimal manner from an infrastructure cost and a service quality perspective, knowledge of the performance capabilities of the underlying technologies and computing resources is critical. Therefore, to gain insight which can be applied to the design of such mechanisms, but also with further implications on service pricing and SLA design, we carry out a measurement campaign to evaluate the capabilities of key enabling technologies for CDNaaS provision. In particular, we focus on virtualization and containerization technologies for implementing virtual CDN functions to deliver a generic HTTP service, as well as an HTTP video streaming one, empirically capturing the relationship between performance and service workload, both from a system operator and a user-centric viewpoints.

New tools for network design. In the efforts for designing future networks' topologies, the inclusion of dependability aspects has been recently enriched with finer criteria, and one relatively new family of metrics consider diameter-constrained parameters that capture more accurately reliability aspects of communication infrastructures. This is done by taking into account not only connectivity properties but also delays when nodes are connected. Paper [15] deals with factorization theory in diameter-constrained reliability, when terminal nodes are further required to be connected by d hops or fewer (d is a given strictly positive parameter of the metric, called its diameter). This metric was defined in 2001, inspired by delay-sensitive applications in telecommunications. Factorization theory is fundamental for classical network reliability evaluation, and today it is a mature area. However, its extension to the diameter-constrained context requires at least the recognition of irrelevant links, which is an open problem. In this paper, irrelevant links are efficiently determined in the most used case, where we consider the communication between a given pair of nodes in the network. The article also proposes a Factoring algorithm that includes the way series-parallels substructures can be handled.

Quality of Experience activities. We continue to develop tools for Quality of Experience assessment, and applications of this quantitative evaluation.

Predicting time series. For the future of the PSQA project, we intend to integrate the capability of *predicting* the Perceptual Quality and not only evaluating its current value. With this goal in mind, we explored this year the idea of combining a Reservoir Computing architecture (whose good performances have been reported many times, when used to predict sequences of numbers or of vectors) with Recurrent Random Neural Networks, that belong to a class of Neural Networks that have some nice properties. Both have been very successful in many applications. In [29] we propose a new model belonging to the first class, taking the structure of the second for its dynamics. The new model is called Echo State Queuing Network. The paper positions the model in the global Machine Learning area, and provides examples of its use and performances. We show on largely used benchmarks that it is a very accurate tool, and we illustrate how it compares with standard Reservoir Computing models. In [31] we presented some preliminary results to the Random Neural Network community.

QoE and P2P design. In [30] we describe a Peer-to-Peer (P2P) network that was designed to support Video on Demand (VoD) services. The network is based on a video-file sharing mechanism that classifies peers

according to the window (segment of the file) that they are downloading. This classification easily allows identifying peers that are able to share windows among them, so one of our major contributions is the definition of a mechanism that could be implemented to efficiently distribute video content in future 5G networks. Considering that cooperation among peers can be insufficient to guarantee an appropriate system performance, we also propose that this network must be assisted by upload bandwidth coming from servers; since these resources represent an extra cost to the service provider, especially in mobile networks, we complement our work by defining a scheme that efficiently allocates them only to those peers that are in windows with resources scarcity (we called it *prioritized windows distribution scheme*). On the basis of a fluid model and a Markov chain, we also develop a methodology that allows us to select the system parameters values (e.g., windows sizes or minimum servers upload bandwidth) that satisfy a set of Quality of Experience (QoE) parameters.

DYOGENE Project-Team

7. New Results

7.1. Reversibility and further properties of FCFS infinite bipartite matching

[3] The model of FCFS infinite bipartite matching was introduced in Caldentey, Kaplan, & Weiss Adv. Appl. Probab., 2009. In this model, there is a sequence of items that are chosen i.i.d. from a finite set C and an independent sequence of items that are chosen i.i.d. from a finite set S, and a bipartite compatibility graph G between C and S. Items of the two sequences are matched according to the compatibility graph, and the matching is FCFS, meaning that each item in the one sequence is matched to the earliest compatible unmatched item in the other sequence. In Adan & Weiss, Operations Research, 2012, a Markov chain associated with the matching was analyzed, a condition for stability was derived, and a product form stationary distribution was obtained. In the current paper, we present several new results that unveil the fundamental structure of the model. First, we provide a pathwise Loynes' type construction which enables to prove the existence of a unique matching for the model defined over all the integers. Second, we prove that the model is dynamically reversible: we define an exchange transformation in which we interchange the positions of each matched pair, and show that the items in the resulting permuted sequences are again independent and i.i.d., and the matching between them is FCFS in reversed time. Third, we obtain product form stationary distributions of several new Markov chains associated with the model. As a by product, we compute useful performance measures, for instance the link lengths between matched items.

7.2. Point-map-probabilities of a point process and Mecke's invariant measure equation

[4] A compatible point-shift F maps, in a translation invariant way, each point of a stationary point process Φ to some point of Φ. It is fully determined by its associated point-map, f, which gives the image of the origin by F. It was proved by J. Mecke that if F is bijective, then the Palm probability of Φ is left invariant by the translation of $-f$. The initial question motivating this paper is the following generalization of this invariance result: in the nonbijective case, what probability measures on the set of counting measures are left invariant by the translation of $-f$? The point-map-probabilities of Φ are defined from the action of the semigroup of point-map translations on the space of Palm probabilities, and more precisely from the compactification of the orbits of this semigroup action. If the point-map-probability exists, is uniquely defined and if it satisfies certain continuity properties, it then provides a solution to this invariant measure problem. Point-map-probabilities are objects of independent interest. They are shown to be a strict generalization of Palm probabilities: when F is bijective, the point-map-probability of Φ boils down to the Palm probability of Φ. When it is not bijective, there exist cases where the point-map-probability of Φ is singular with respect to its Palm probability. A tightness based criterion for the existence of the point-map-probabilities of a stationary point process is given. An interpretation of the point-map-probability as the conditional law of the point process given that the origin has F-pre-images of all orders is also provided. The results are illustrated by a few examples.

7.3. Gibbsian on-line distributed content caching strategy for cellular networks

[7] We develop Gibbs sampling based techniques for learning the optimal content placement in a cellular network. A collection of base stations are scattered on the space, each having a cell (possibly overlapping with other cells). Mobile users request for downloads from a finite set of contents according to some popularity distribution. Each base station can store only a strict subset of the contents at a time; if a requested content is not available at any serving base station, it has to be downloaded from the backhaul. Thus, there arises the problem of optimal content placement which can minimize the download rate from the backhaul, or equivalently maximize the cache hit rate. Using similar ideas as Gibbs sampling, we propose imple sequential

content update rules that decide whether to store a content at a base station based on the knowledge of contents in neighbouring base stations. The update rule is shown to be asymptotically converging to the optimal content placement for all nodes. Next, we extend the algorithm to address the situation where content popularities and cell topology are initially unknown, but are estimated as new requests arrive to the base stations. Finally, improvement in cache hit rate is demonstrated numerically.

7.4. State estimation for the individual and the population in mean field control with application to demand dispatch

[10] This paper concerns state estimation problems in a mean field control setting. In a finite population model, the goal is to estimate the joint distribution of the population state and the state of a typical individual. The observation equations are a noisy measurement of the population. The general results are applied to demand dispatch for regulation of the power grid, based on randomized local control algorithms. In prior work by the authors it is shown that local control can be designed so that the aggregate of loads behaves as a controllable resource, with accuracy matching or exceeding traditional sources of frequency regulation. The operational cost is nearly zero in many cases. The information exchange between grid and load is minimal, but it is assumed in the overall control architecture that the aggregate power consumption of loads is available to the grid operator. It is shown that the Kalman filter can be constructed to reduce these communication requirements, and to provide the grid operator with accurate estimates of the mean and variance of quality of service (QoS) for an individual load.

7.5. Distributed spectrum management in TV white space networks

[11] In this paper, we investigate the spectrum management problem in TV White Space (TVWS) Cognitive Radio Networks using a game theoretical approach, accounting for adjacent-channel interference. TV Bands Devices (TVBDs) compete to access available TV channels and choose idle blocks that optimize some objective function. Specifically, the goal of each TVBD is to minimize the price paid to the Database operator and a cost function that depends on the interference between unlicensed devices. We show that the proposed TVWS management game admits a potential function under general conditions. Accordingly, we use a Best Response algorithm to converge in few iterations to the Nash Equilibrium (NE) points. We evaluate the performance of the proposed game, considering both static and dynamic TVWS scenarios and taking into account users' mobility. Our results show that at the NE, the game provides an interesting tradeoff between efficient TV spectrum use and reduction of interference between TVBDs.

7.6. A spectral method for community detection in moderately sparse degree-corrected stochastic block models

[12] We consider community detection in degree-corrected stochastic block models. We propose a spectral clustering algorithm based on a suitably normalized adjacency matrix. We show that this algorithm consistently recovers the block membership of all but a vanishing fraction of nodes, in the regime where the lowest degree is of order log(n) or higher. Recovery succeeds even for very heterogeneous degree distributions. The algorithm does not rely on parameters as input. In particular, it does not need to know the number of communities.

7.7. Non-backtracking spectrum of degree-corrected stochastic block models

[25] Motivated by community detection, we characterise the spectrum of the non-backtracking matrix B in the Degree-Corrected Stochastic Block Model. Specifically, we consider a random graph on n vertices partitioned into two asymptotically equal-sized clusters. The vertices have i.i.d. weights $\{\phi_u\}_{u=1}^n$ with second moment $\Phi^{(2)}$. The intra-cluster connection probability for vertices u and v is $\frac{\phi_u\phi_v}{n}a$ and the inter-cluster connection probability is $\frac{\phi_u\phi_v}{n}b$. We show that with high probability, the following holds: The leading eigenvalue of the non-backtracking matrix B is asymptotic to $\rho = \frac{a+b}{2}\Phi^{(2)}$. The second eigenvalue is asymptotic to $\mu_2 = \frac{a-b}{2}\Phi^{(2)}$ when $\mu_2^2 > \rho$, but asymptotically bounded by $\sqrt{\rho}$ when $\mu_2^2 \leq \rho$. All the remaining eigenvalues

are asymptotically bounded by $\sqrt{\rho}$. As a result, a clustering positively-correlated with the true communities can be obtained based on the second eigenvector of B in the regime where $\mu_2^2 > \rho$. In a previous work we obtained that detection is impossible when $\mu_2^2 < \rho$, meaning that there occurs a phase-transition in the sparse regime of the Degree-Corrected Stochastic Block Model. As a corollary, we obtain that Degree-Corrected Erdős-Rényi graphs asymptotically satisfy the graph Riemann hypothesis, a quasi-Ramanujan property. A by-product of our proof is a weak law of large numbers for local-functionals on Degree-Corrected Stochastic Block Models, which could be of independent interest.

7.8. A spectral algorithm with additive clustering for the recovery of overlapping communities in networks

[13] This paper presents a novel spectral algorithm with additive clustering designed to identify overlapping communities in networks. The algorithm is based on geometric properties of the spectrum of the expected adjacency matrix in a random graph model that we call stochastic blockmodel with overlap (SBMO). An adaptive version of the algorithm, that does not require the knowledge of the number of hidden communities, is proved to be consistent under the SBMO when the degrees in the graph are (slightly more than) logarithmic. The algorithm is shown to perform well on simulated data and on real-world graphs with known overlapping communities.

7.9. Optimal geographic caching in cellular networks with linear content coding

[14] We state and solve a problem of the optimal geographic caching of content in cellular networks, where linear combinations of contents are stored in the caches of base stations. We consider a general content popularity distribution and a general distribution of the number of stations covering the typical location in the network. We are looking for a policy of content caching maximizing the probability of serving the typical content request from the caches of covering stations. The problem has a special form monotone sub-modular set function maximization. Using dynamic programming, we find a deterministic policy solving the problem. We also consider two natural greedy caching policies. We evaluate our policies considering two popular stochastic geometric coverage models: the Boolean one and the Signal-to-Interference-and-Noise-Ratio one, assuming Zipf popularity distribution. Our numerical results show that the proposed deterministic policies are in general not worst than some randomized policy considered in the literature and can further improve the total hit probability in the moderately high coverage regime.

7.10. Online mobile user speed estimation: performance and tradeoff considerations

[15] This paper presents an online algorithm for mobile user speed estimation in 3GPP Long Term Evolution (LTE)/LTE-Advanced (LTE-A) networks. The proposed method leverages on uplink (UL) sounding reference signal (SRS) power measurements performed at the base station, also known as eNodeB (eNB), and remains effective even under large sampling period. Extensive performance evaluation of the proposed algorithm is carried out using field traces from realistic environment. The on-line solution is proven highly efficient in terms of computational requirement, estimation delay, and accuracy. In particular, we show that the proposed algorithm can allow for the first speed estimation to be obtained after 10 seconds and with an average speed underestimation error of 14 kmph. After the first speed acquisition, subsequent speed estimations can be obtained much faster (e.g., each second) with limited implementation cost and still provide high accuracy.

7.11. Self-similarity in urban wireless networks: Hyperfractals

[18] In this work we study a Poisson patterns of fixed and mobile nodes distributed on straight lines designed for 2D urban wireless networks. The particularity of the model is that, in addition to capturing the irregularity and variability of the network topology, it exploits self-similarity, a characteristic of urban wireless networks.

The pattern obeys to " Hyperfractal " measures which show scaling properties corresponding to an apparent dimension larger than 2. The hyperfractal pattern is best suitable for capturing the traffic over the streets and highways in a city. The scaling effect depends on the hyperfractal dimensions. Assuming radio propagation limited to streets, we prove results on the scaling of routing metrics and connectivity graph.

7.12. Optimizing spatial throughput in device-to-device networks

[19] Results are presented for optimizing device-to-device communications in cellular networks, while maintaining spectral efficiency of the base-station-to-device downlink channel. We build upon established and tested stochastic geometry models of signal-to-interference ratio in wireless networks based on the Poisson point process, which incorporate random propagation effects such as fading and shadowing. A key result is a simple formula, allowing one to optimize the device-to-device spatial throughput by suitably adjusting the proportion of active devices. These results can lead to further investigation as they can be immediately applied to more sophisticated models such as studying multi-tier network models to address coverage in closed access networks.

7.13. Demand dispatch with heterogeneous intelligent loads

[20] A distributed control architecture is presented that is intended to make a collection of heterogeneous loads appear to the grid operator as a nearly perfect battery. Local control is based on randomized decision rules advocated in prior research, and extended in this paper to any load with a discrete number of power states. Additional linear filtering at the load ensures that the input-output dynamics of the aggregate has a nearly flat input-output response: the behavior of an ideal, multi-GW battery system.

7.14. Energy savings for virtual MISO in fractal sensor networks

[21] We design a model of wireless terminals, i.e. transmitters and receivers, obtained from a Poisson point process with support in an embedded fractal map. The terminals form a virtual MISO (Multiple Input Single Output) system with successful reception under SNR (signal-to-noise ratio) capture condition in a single hop transmission. We show that if we omit antennas cross sections, the energy needed to broadcast a packet of information tends to zero when the density of transmitters and receivers increases. This property is a direct consequence of the fact that the support map is fractal and would not hold if the terminal distribution were Poisson uniform, as confirmed by simulations. The result becomes invalid if the cross sections overlap or if we consider a masking effect due to antennas, which would imply an extremely large density of terminals. In the case where the cross sections of the transmitters have a non-zero value, the energy has a non-zero limit which decays to zero when the cross sections tend to zero.

7.15. Distributed control of a fleet of batteries

[22] Battery storage is increasingly important for grid-level services such as frequency regulation, load following, and peak-shaving. The management of a large number of batteries presents a control challenge: How can we solve the apparently combinatorial problem of coordinating a large number of batteries with discrete, and possibly slow rates of charge/discharge? The control solution must respect battery constraints, and ensure that the aggregate power output tracks the desired grid-level signal. A distributed stochastic control architecture is introduced as a potential solution. Extending prior research on distributed control of flexible loads, a randomized decision rule is defined for each battery of the same type. The power mode at each time-slot is a randomized function of the grid-signal and its internal state. The randomized decision rule is designed to maximize idle time of each battery, and keep the state-of-charge near its optimal level, while ensuring that the aggregate power output can be continuously controlled by a grid operator or aggregator. Numerical results show excellent tracking, and low stress to individual batteries.

7.16. Exact Computation and bounds for the coupling time in queueing systems

[23] This paper is a work in progress on the exact computation and bounds of the expected coupling time for finite-state Markov chains. We give an exact formula in terms of generating series. We show how this may help to bound the expected coupling time for queueing networks.

7.17. An online disaggregation algorithm and its application to demand control

[24] The increase of renewable energy has made the supply-demand balance of power more complex to handle. Previous approach designed randomized controllers to obtain ancillary services to the power grid by harnessing inherent flexibility in many loads. However these controllers suppose that we know the consumption of each device that we want to control. This introduce the cost and the social constraint of putting sensors on each device of each house. Therefore, our approach was to use Nonintrusive Appliance Load Monitoring (NALM) methods to solve a disaggregation problem. The latter comes down to estimating the power consumption of each device given the total power consumption of the whole house. We started by looking at the Factorial Hierarchical Dirichlet Process-Hidden Semi-Markov Model (Factorial HDP-HSMM). In our application, the total power consumption is considered as the observations of this state-space model and the consumption of each device as the state variables. Each of the latter is modelled by an HDP-HSMM which is an extension of a Hidden Markov Model. However, the inference method proposed previously is based on Gibbs sampling and has a complexity of $O(T2N + TN2)$ where T is the number of observations and N is the number of hidden states. As our goal is to use the randomized controllers with our estimations, we wanted a method that does not scale with T. Therefore, we developed an online algorithm based on particle filters. Because we worked in a Bayesian setting, we had to infer the parameters of our model. To do so, we used a method called Particle Learning. The idea is to include the parameters in the state space so that they are tied to the particles. Then, for each (re)sampling step, the parameters are sampled from their posterior distribution with the help of Bayesian sufficient statistics. We applied the method to data from Pecan Street. Using their Dataport, we have collected the power consumption of each device from about a hundred houses. We selected the few devices that consume the most and that are present in most houses. We separated the houses in a training set and a test set. For each device of each house from the training set, we estimated the operating modes with a HDP-HSMM and used these estimations to compute estimators of the priors hyperparameters. Finally we applied the particle filters method to the test houses using the computed priors. The algorithm performs well for the device with the highest power consumption, the air compressor in our case. We will discuss ongoing work where we apply the "Thermo-statically Controlled Loads" example using our estimations of this air compressor's operating modes.

7.18. Multiple local community detection

[26] Community detection is a classical problem in the field of graph mining. We are interested in local community detection where the objective is the recover the communities containing some given set of nodes, called the seed set. While existing approaches typically recover only one community around the seed set, most nodes belong to multiple communities in practice. In this paper, we introduce a new algorithm for detecting multiple local communities, possibly overlapping, by expanding the initial seed set. The new nodes are selected by some local clustering of the graph embedded in a vector space of low dimension. We validate our approach on real graphs, and show that it provides more information than existing algorithms to recover the complex graph structure that appears locally.

7.19. A Streaming Algorithm for Graph Clustering

[27] We introduce a novel algorithm to perform graph clustering in the edge streaming setting. In this model, the graph is presented as a sequence of edges that can be processed strictly once. Our streaming algorithm has an extremely low memory footprint as it stores only three integers per node and does not keep any edge in memory. We provide a theoretical justification of the design of the algorithm based on the modularity function,

which is a usual metric to evaluate the quality of a graph partition. We perform experiments on massive real-life graphs ranging from one million to more than one billion edges and we show that this new algorithm runs more than ten times faster than existing algorithms and leads to similar or better detection scores on the largest graphs.

7.20. Discrete probability models and methods: probability on graphs and trees, markov chains and random fields, entropy and coding

[28] The emphasis in this book is placed on general models (Markov chains, random fields, random graphs), universal methods (the probabilistic method, the coupling method, the Stein-Chen method, martingale methods, the method of types) and versatile tools (Chernoff's bound, Hoeffding's inequality, Holley's inequality) whose domain of application extends far beyond the present text. Although the examples treated in the book relate to the possible applications, in the communication and computing sciences, in operations research and in physics, this book is in the first instance concerned with theory. The level of the book is that of a beginning graduate course. It is self-contained, the prerequisites consisting merely of basic calculus (series) and basic linear algebra (matrices). The reader is not assumed to be trained in probability since the first chapters give in considerable detail the background necessary to understand the rest of the book.

7.21. Distributed control design for balancing the grid using flexible loads

[29] nexpensive energy from the wind and the sun comes with unwanted volatility, such as ramps with the setting sun or a gust of wind. Controllable generators manage supply-demand balance of power today, but this is becoming increasingly costly with increasing penetration of renewable energy. It has been argued since the 1980s that consumers should be put in the loop: " demand response " will help to create needed supply-demand balance. However, consumers use power for a reason, and expect that the quality of service (QoS) they receive will lie within reasonable bounds. Moreover, the behavior of some consumers is unpredictable, while the grid operator requires predictable controllable resources to maintain reliability. The goal of this chapter is to describe an emerging science for demand dispatch that will create virtual energy storage from flexible loads. By design, the grid-level services from flexible loads will be as controllable and predictable as a generator or fleet of batteries. Strict bounds on QoS will be maintained in all cases. The potential economic impact of these new resources is enormous. California plans to spend billions of dollars on batteries that will provide only a small fraction of the balancing services that can be obtained using demand dispatch. The potential impact on society is enormous: a sustainable energy future is possible with the right mix of infrastructure and control systems.

7.22. Un classificateur non-supervisé utilisant les complexes simpliciaux avec une application à la stylométrie

[30] Un classificateur non-supervisé utilisant les complexes simpliciaux (avec une application à la stylométrie). Nous nous proposons au cours des quelques pages de ce rapport de présenter au lecteur ce que sont les complexes simpliciaux ainsi qu'une de leurs possibles (et nombreuses !) applications : en classification non-supervisée. Les complexes simpliciaux peuvent s'appréhender comme une généralisation des graphes ; un graphe étant la donnée d'un ensemble de sommets ainsi que d'une relation de voisinage entre des paires de ces sommets (deux points sont voisins si une arête les relie). Les complexes simpliciaux permettent de rendre compte de relations de voisinage plus élaboré (et faisant notamment intervenir un nombre arbitraire de points ; pas seulement deux). La classification non supervisée est une branche du vaste domaine de l'apprentissage automatique. Etant donné un échantillon de données (le plus souvent des points de l'espace euclidien R^d), elle consiste a regrouper ces données en différentes classes de sorte que les données d'une même classe présentent des similarités entre elles tandis que deux données appartenant à deux classes distinctes soient dissemblables. Le présent rapport s'articulera donc en deux parties : la première introduira au lecteur non forcément familier cette notion de complexe simplicial d'un point de vue théorique. On l'illustrera ensuite avec la présentation des complexes de Cech et certaines propriétés mathématiques qui en font un outil puissant et pratique (la

théorie de Morse permet, par exemple, de manier ces complexes de différentes façons). On verra encore quelques résultats des complexes simpliciaux aléatoires (c'est-à-dire que les sommets sont des points générés aléatoirement) dans le cas des régimes dits surcritiques jus-tifiant certains algorithmes d'apprentissage de variétés (une des multiples applications promises des complexes simpliciaux). Enfin, nous présenterons très succinctement l'homologie persistante...

7.23. Phase transitions, optimal errors and optimality of message-passing in generalized linear models

[31] We consider generalized linear models where an unknown nn-dimensional signal vector is observed through the successive application of a random matrix and a non-linear (possibly probabilistic) componentwise function. We consider the models in the high-dimensional limit, where the observation consists of $m \times n$ points, and $m/n \to \alpha$ where α stays finite in the limit $m, n \to \infty$. This situation is ubiquitous in applications ranging from supervised machine learning to signal processing. A substantial amount of work suggests that both the inference and learning tasks in these problems have sharp intrinsic limitations when the available data become too scarce or too noisy. Here, we provide rigorous asymptotic predictions for these thresholds through the proof of a simple expression for the mutual information between the observations and the signal. Thanks to this expression we also obtain as a consequence the optimal value of the generalization error in many statistical learning models of interest, such as the teacher-student binary perceptron, and introduce several new models with remarquable properties. We compute these thresholds (or "phase transitions") using ideas from statistical physics that are turned into rigorous methods thanks to a new powerful smart-path interpolation technique called the stochastic interpolation method, which has recently been introduced by two of the authors. Moreover we show that a polynomial-time algorithm refered to as generalized approximate message-passing reaches the optimal generalization performance for a large set of parameters in these problems. Our results clarify the difficulties and challenges one has to face when solving complex high-dimensional statistical problems.

7.24. Lecture notes on random geometric models — random graphs, point processes and stochastic geometry

[32] The goal of this sequence of lessons is to provide quick access to some popular models of random geometric structures used in many applications: from communication networks, including social, transportation, wireless networks, to geology, material sciences and astronomy. The course is composed of the following 15 lessons: (1) Bond percolation on the square lattice, (2) Galton-Watson tree, (3) Erdős-Rényi graph — emergence of the giant component, (4) Graphs with a given node degree distribution, (5) Typical nodes and random unimodular graphs, (6) Erdős-Rényi graph — emergence of the full connectivity, (7) Poisson point process, (8) Point conditioning and Palm theory for point processes, (9) Hard-core point processes, (10) Stationary point processes and mass transport principle, (11) Stationary Voronoi tessellation, (12) Ergodicity and point-shift invariance, (13) Random closed sets, (14) Boolean model and coverage processes, (15) Connectedness of random sets and continuum percolation. Usually, these subjects are presented in different monographs: random graphs (lessons 2–6), point processes (7-12), stochastic geometry (13-14), with percolation models presented in lesson 1 and 15 often addressed separately. Having them in one course gives us an opportunity to observe some similarities and even fundamental relations between different models. Examples of such connections are:

- Similar phase transitions regarding the emergence of big components observed in different discrete, lattice and continuous euclidean models (lessons 1–4, 15).
- Single isolated nodes being the last obstacle in the emergence of the full connectivity in some discrete and euclidean graphs exhibiting enough independence (lessons 6, 15).
- A mass transport principle as a fundamental property for unimodular random graphs and Palm theory for stationary point processes; with both theories seeking to define the typical node/point of a homogeneous structure (lessons 5, 10–12).

- Poisson-Galton-Watson tree and Poisson process playing a similar role in the theory of random graphs and point processes, respectively: for both models independence and Poisson distribution are the key assumptions, both appear as natural limits, and both rooted/conditioned to a typical node/point preserve the distribution of the remaining part of the structure (lessons 2,5, 7–8).
- Size biased sampling appearing in several, apparently different, conditioning scenarios, as unimodular trees (lesson 5), Palm distributions for point process (lesson 8), zero cell of the stationary tessellations (lessons 11).

The goal of this series of lectures is to present some spectrum of models and ideas. When doing this, we sometimes skip more technical proof details, sending the reader for them to more specialised monographs. Some theoretical and computer exercises are provided after each lesson to let the reader practice his/her skills. Regarding the prerequisites, the reader will benefit from having had some prior exposure to probability and measure theory, but this is not absolutely necessary.

The content of the course has been evolving while the author teaches it within the master programme *Probabilité et modelés aléatoires* at the University Pierre and Marie Curie in Paris. The present notes were thoroughly revised when the author was presenting them as a specially appointed professor at the School of Computing, Tokyo Institute of Technology, in the autumn term 2017.

7.25. Energy trade-offs for end-to-end communications in urban vehicular networks exploiting an hyperfractal model

[34] We present results on the trade-offs between the end-to-end communication delay and energy spent for completing a transmission in vehicular communications in urban settings. This study exploits our innovative model called “hyperfractal” that captures the self-similarity of the topology and vehicle locations in cities. We enrich the model by incorporating roadside infrastructure. We use analytical tools to derive theoretical bounds for the end-to-end communication hop count under two different energy constraints: either total accumulated energy, or maximum energy per node. More precisely, we prove that the hop count is bounded by $O(n1-\alpha/(dm-1))$ where $\alpha < 1$ and d $m > 2$ is the precise hyperfractal dimension. This proves that for both constraints the energy decreases as we allow to chose among paths of larger length. In fact the asymptotic limit of the energy becomes significantly small when the number of nodes becomes asymptotically large. A lower bound on the network throughput capacity with constraints on path energy is also given. The results are confirmed through exhaustive simulations using different hyperfractal dimensions and path loss coefficients.

7.26. Fundamental limits of low-rank matrix estimation: the non-symmetric case

[36] We consider the high-dimensional inference problem where the signal is a low-rank symmetric matrix which is corrupted by an additive Gaussian noise. Given a probabilistic model for the low-rank matrix, we compute the limit in the large dimension setting for the mutual information between the signal and the observations, as well as the matrix minimum mean square error, while the rank of the signal remains constant. We also show that our model extends beyond the particular case of additive Gaussian noise and we prove an universality result connecting the community detection problem to our Gaussian framework. We unify and generalize a number of recent works on PCA, sparse PCA, submatrix localization or community detection by computing the information-theoretic limits for these problems in the high noise regime. In addition, we show that the posterior distribution of the signal given the observations is characterized by a parameter of the same dimension as the square of the rank of the signal (i.e. scalar in the case of rank one). Finally, we connect our work with the hard but detectable conjecture in statistical physics.

EVA Project-Team

7. New Results

7.1. 6TiSCH Standardization and Benchmarking

7.1.1. Minimal Security Solution

Participants: Malisa Vucinic, Thomas Watteyne.

The 6TiSCH standardization effort had, until 2017, a big gap: security. Thanks to the work of Malisa Vucinic, this gap is now filled, with the publication of the Minimal Security solution (draft-ietf-6tisch-minimal-security). Here is a summary of what has been implemented and tested:

- Two implementations of the OSCORE protocol, formerly known as OSCOAP, specified in draft-ietf-core-object-security-03, in C and in Python, supporting both client and server roles, as part of the OpenWSN stack. Updated the test suite of the Python implementation with OSCOAP functional tests.
- Two implementations of Simple Join Protocol for 6TiSCH, specified in draft-ietf-6tisch-minimal-security-03, in C supporting the role of a pledge and in Python, supporting the role of JRC. Written unit tests for the implemented CBOR decoder in C.
- Simulation of the join process in 6TiSCH simulator. Extended the simulator to support shared cells, downwards RPL routing and join traffic. Tested the two implementations of Simple Join Protocol/OSCOAP using the F-Interop tools.

7.1.2. OpenWSN Fresh with full 6TiSCH Support

Participants: Tengfei Chang, Thomas Watteyne.

Thanks to the incredible work of Tengfei Chang, the OpenWSN project was refocused on being the lead reference 6TiSCH implementation. “OpenWSN Fresh” was a 2017 program to separate the protocol stack implementation from the rest of the OpenWSN code, and to have full standards-compliance.

7.1.3. First F-Interop 6TiSCH Interop Event

Participants: Remy Leone, Tengfei Chang, Malisa Vucinic, Thomas Watteyne.

The 6TiSCH WG organized an interoperability event co-located with the IETF meeting in Prague in July 2017. The interop tests focused on the minimal security framework and the 6top protocol. OpenWSN was used as the reference implementation, and F-Interop tools were demonstrated.

7.1.4. Agile Networking

Participants: Jonathan Munoz, Thomas Watteyne.

Today’s low-power wireless devices typically consist of a micro-controller and a radio. The most commonly used radios are IEEE802.15.4 2.4GHz, IEEE802.15.4g sub-GHz and LoRA (SemTech) compliant. Radios offer a different trade-off between range and data-rate, given some energy budget. To make things more complex, standards such IEEE802.15.4g include different modulations schemes (2-FSK, 4-FSK, O-QPSK, OFDM), further expanding the number of options.

The main idea behind agile networking is to redefine a low-power wireless device as having multiple radios, which it can possibly use at the same time. That is, in a TSCH context, for each frame a node sends, it can change the radio it is using, and its setting. If the next hop is close, it sends the frame with a fast data rate thereby reducing the radio on-time and the energy consumption. If the next hop is far, it uses a slower data rate.

We recently design the OpenMote B within the OpenMote company. This board contains both a CC2538 IEEE802.15.4 radio, and an AT86RF215 IEEE802.15.4g radio, offering communication on both 2.4GHz and sub-GHz frequency bands, 4 modulations schemes, and data rates from 50 kbps to 800 kbps. The first prototypes started being tested in December 2017.

The second challenge is to redesign the protocol stack in a standards-compliant way. We are working with Jonathan Munoz on a 6TiSCH design in which neighbor discovery happens independently on each radio, and the same neighbor node can appear as many times in the neighbor table as it has radios. The goal is to standardize an "Agile 6TiSCH" profile, without having to touch the core specifications. This is been implemented in OpenWSN. The next step is to evaluate the performance of the solution on an 80-node OpenMote B testbed we are putting together. We hope to show that a single device running the same stack can satisfy both building-size and campus-size deployment, with the same industrial requirements.

7.2. SolSystem Deployments

SolSystem (http://solsystem.io/) is a complete sensor-to-cloud solution, which the Inria-EVA team uses to federate the different real-world deployments it is conducting.

7.2.1. SmartMarina

Participants: Ziran Zhang, Keoma Brun-Laguna, Thomas Watteyne.

Marinas are quickly evolving from sailing spots to floating neighborhoods. It is now common for people to live on their boat year-round, and for boats to be rented for just a week-end through online platforms. Today, living or staying on a boat is often cheaper that buying or renting an apartment. Similarly, in coastal areas, the marina is often the center of the city, so an ideal location for lodging. As a result, the trend is not going to end any time soon. Today's marinas are tomorrow's smart cities.

And as the marina is evolving, so are its needs.

- From a marina management point of view, automatic mooring management and electricity/water monitoring allows personnel to free up to welcome visitors and focus entirely on their well-being.
- Year-road boat owners and occasional marina visitors now can enjoy new services, from increased mooring availability to remote monitoring and alerts about the state of their boat.

The combination of embedded micro-controllers, low-power wireless communication and sensors/actuators offers tremendous opportunities for marinas. Off-the-shelf "Internet of Things" technology can now be used to detect the presence of boats in moorings, track usage of water and electricity on a per-boat basis, track a boat in real-time as it enters the marina, etc. Because no wires need to be installed – neither for power, nor communication – installation can be done in a matter of hours in a peal-and-stick fashion. Pontoons can be moved, rearranged or removed, without having to worry about the smart devices mounted on it.

The goal of the SmartMarina project (http://smartmarina.org/) is to build a system composed of sensors deployed all over the marina, and advanced software to monitor the occupation of moorings, and the electricity and water consumption on each spot. The result is a system that allows more efficient management and new services. The first sensor was installed in April 2017, and the Inria-EVA team is looking at turning this activity into a startup company.

7.2.2. SaveThePeaches

Participants: Keoma Brun-Laguna, Thomas Watteyne.

In 2013, 85% of the peach production in the Mendoza region (Argentina) was lost because of frost. Because less fruit was produced in the region, 600.000 less work days were needed to process the harvest between November 2013 and March 2014, a reduction in work force of 10.600 people. Across the Mendoza region, frost has caused a loss of revenue of 950 million Argentine pesos roughly 100 million USD in the peach business alone.

A frost event happens when the temperature is so low that the crops cannot recover their tissue or internal structure from the effects of water freezing inside or outside the plant. For the peach production, a critical period is when the trees are in bloom and fruit set (Aug./Sept. in Mendoza), during which the temperature needs to be kept above 3 C. Even a few hours below that temperature causes flowers to fall, preventing fruits to grow.

Because of the huge economic impact, countermeasures exist and are used extensively. Today, virtually all industrial peach orchards are equipped with a small number of meteorological stations which monitor temperature and humidity. If the temperature drops dangerously low, the most effective countermeasures is to install a number of furnaces in the orchard (typically coalfueled) and fly helicopters above the orchard to distribute the heat and avoid cold spots. This countermeasure is effective, but suffers from false positives (the helicopters are called in, but there is no frost event) and false negatives (the meteorological stations don't pick up a frost event happening in some part of the orchard).

What the SaveThePeaches project (http://www.savethepeaches.com/) has developed in 2016-2017 is a dense 120-sensor real-time monitoring solution deployed in the orchard, and feeding a frost prediction model. A node is the size of a deck of cards, is self-contained and battery-operated. When switched on, nodes form a multi-hop low-power wireless network, automatically. Rather than being installed at a fixed location, these nodes can be hung directly in the trees. A network is deployed in an orchard in a matter of hours, and if needed, sensing points can be moved to improve the accuracy of the prediction model in minutes. We use machine learning and pattern recognition to build an micro-climate predictive model by continuously analyzing the gathered sensor data in real time. This model generates early frost warnings. Ones demonstrated, the solution can be extended to other crops, and other regions.

7.2.3. SnowHow

Participants: Keoma Brun-Laguna, Thomas Watteyne.

Between 2012 and 2015, California suffered from the highest water drought since recordings started in this state. Up to 2/3 of its water resources are coming from the Sierra Nevada snowpack. Understanding the effect of the droughts on the mountain snowpack is crucial.

Historically, the study of mountain hydrology and the water cycle has been largely observational, with variables extrapolated from a few infrequent manual measurements. Low-power wireless mesh networking technology has evolved significantly over recent years. With this technology, a node is the size of a deck of cards, is self-contained and battery-operated. When switched on, nodes form a multi-hop low-power wireless network, automatically. Next-generation hydrologic science and monitoring requires real-time, spatially distributed measurements of key variables including: soil moisture, air/soil temperature, snow depth, and air relative humidity.

The SnowHow project (http://snowhow.io/) provides these measurements by deploying low-power mesh networks across the California Sierra Nevada. Off-the-shelf commercial solutions are available today which offer >99.999% end-to-end data reliability and a decade of battery lifetime. A new wireless network can be deployed in a couple of hours and report sensor data minutes after it was measured.

7.3. IoT and Wireless Sensor Networks

More than 50 billions of devices will be connected in 2020. This huge infrastructure of devices, which is managed by highly developed technologies, is called Internet of Things (IoT). The latter provides advanced services, and brings economical and societal benefits. This is the reason why engineers and researchers of both industry and scientific communities are interested in this area. The Internet of Things enables the interconnection of smart physical and virtual objects, managed by highly developed technologies. WSN (Wireless Sensor Network), is an essential part of this paradigm. The WSN uses smart, autonomous and usually limited capacity devices in order to sense and monitor their environment.

7.3.1. Deployment of autonomous and mobile wireless sensor nodes

Participants: Ines Khoufi, Pascale Minet.

This work was done in collaboration with Nadia Boufares (ENSI, University of Manouba, Tunisia) and Leila Saidane (ENSI, University of Manouba, Tunisia).

Wireless Sensor Networks (WSNs) are used in a wide range of applications due to their monitoring and tracking abilities. Depending on the applications goals, sensor nodes are deployed either in a two-dimensional (2D) area or in a three-dimensional (3D) area. In addition, WSN deployment can be either in a distributed or a centralized manner. In 2017, we were interested in a fully distributed deployment of WSN in several 3D-flat-surface configurations using autonomous and mobile nodes. Our goal was to ensure full 3D flat surfaces coverage and maintain network connectivity for these surfaces. To reach our goal we proposed 3D-DVFA-FSC, a distributed deployment algorithm based on virtual forces strategy to move sensor nodes over different 3D-flat-surface shapes. Initially, nodes were randomly deployed. Full coverage was reached in the given configurations and maintained up to the end of simulation. We also evaluated the total distance traveled by nodes. Simulation results show that sensor nodes still move even when full 3D-surface coverage is reached. This is due to the node oscillations problem. This problem will be tackled in our future work. We will also focus on how to stop nodes when full coverage is reached and consider 3D surface complex shapes where the challenges of coverage and connectivity are more complicated. This work was presented at the IWCMC 2017 conference, see [15].

7.3.2. Collision avoidance on shared slots in wireless slotted networks

Participants: Ines Khoufi, Pascale Minet, Paul Muhlethaler.

We propose an analysis of slotted based protocols designed for devices of the Internet of Thing (IoT). In contrast to other TDMA-based protocols this scheme uses a random technique to access shared slots which presents similarities with CSMA protocols. In practice the transmissions are scheduled in a given back-off window of slots whose duration allows the transmission of a packet and its acknowledgment. Therefore this protocol can be analyzed according to the methodology introduced by Bianchi for the IEEE 802.11 protocol even if the protocol studied differs in many aspects. The model we use is also particular because we succeed in obtaining a Markov model even if the scheme used to send a packet (in a node) may depend on the transmission of the previous packet(s). We distinguish two protocols. In the first one, at the initial stage or after a successful transmission, the packets are transmitted without any back-off, whereas in the second protocol each transmission is always preceded by the count down of a random back-off. Extensive simulations show a very good match between the model and the simulation results, see [22]. For moderate medium load, the protocol performing a backoff before each transmission outperforms the TSCH protocol, when the number of neighboring nodes is greater than or equal to 8. For a smaller number of neighboring nodes, the TSCH protocol provides a higher throughput. For high medium load, the TSCH protocol provides the highest normalized throughput at the cost of some unfairness in the transmission opportunities.

7.3.3. Security in the OCARI wireless sensor network

Participant: Pascale Minet.

Wireless Sensor Networks and Industrial Internet of Things use smart, autonomous and usually limited capacity devices in order to sense and monitor industrial environments. The devices in a wireless sensor network are managed by a controller, also called CPAN, which should authenticate them before they join the network. OCARI is a promising wireless sensor network technology providing optimized protocols in order to reduce the energy consumption and support pedestrian mobility. However, it needs to be secured against the different threats, especially those that concern confidentiality, data integrity, and entities authentication. This challenge was addressed in a joint work with Mohammed Tahar Hammi (Telecom ParisTech), Erwan Livolant (AFNet, Boost technologies), Patrick Bellot (Telecom ParisTech), Ahmed Serhouchni (Telecom ParisTech) and **Pascale Minet** (Inria). The main results have been published in two papers.

A robust mutual authentication is the challenge addressed in the paper presented at the ICMWT 2017 conference [28]. We proposed a lightweight, robust, and energy efficient WSN mutual authentication protocol. This protocol is especially designed to be implemented on devices with low storage and computing capacities. It has been implemented on OCARI. All nodes wanting to access the network should be authenticated at the MAC sub-layer of OCARI. This solution provides a protection against "replay attacks", because the exchanged OTPs are based on random numbers, therefore, they are valid only for one transaction. Using the blacklisting mechanism we can secure our systems against "some DoS" attacks. Finally it is flexible and does not decrease the scalability of the system, and can be deployed in different WSNs technologies, while keeping the same level of robustness. In our future work we aim to ensure the confidentiality of the transmitted messages exchanged after the MAC sub-layer association and authentication procedure. And thus we will have a secure system which ensures the "Confidentiality", "Integrity, and "Authentication" services.
In the paper presented at CSNet 2017 ([27] , we designed a security protocol that enables to secure most of the WSNs thanks to its lightness and energy efficiency. It ensures a mutual authentication of the communicating entities and a protection of both the integrity and the confidentiality of the exchanged data. The "personalization" mechanism solves the problem of the internal identity usurpation. The proposed key management allows a safe and secure keys exchange between the concerned entities. Furthermore, this protocol provides a very fast establishment of a secure channel based on a robust, fast, and lightweight symmetric encryption algorithm (AES GCM/CCM). Finally, this solution is resilient against the cryptanalysis and the replay attacks. In our future works, we aim to create a secure communicating system between different CPANs and to facilitate a secure migration of devices from a network managed by a CPAN to a network managed by another CPAN.

7.3.4. Security in Wireless Sensor Networks

Participants: Selma Boumerdassi, Paul Muhlethaler.

Sensor networks are often used to collect data from the environment where they are located. These data can then be transmitted regularly to a special node called a *sink*, which can be fixed or mobile. For critical data (like military or medical data), it is important that sinks and simple sensors can mutually authenticate so as to avoid data to be collected and/or accessed by fake nodes. For some applications, the collection frequency can be very high. As a result, the authentication mechanism used between a node and a sink must be fast and efficient both in terms of calculation time and energy consumption. This is especially important for nodes which computing capabilities and battery lifetime are very low. Moreover, an extra effort has been done to develop alternative solutions to secure, authenticate, and ensure the confidentiality of sensors, and the distribution of keys in the sensor network. Specific researches have also been conducted for large-scale sensors. At present, we work on an exchange protocol between sensors and sinks based on low-cost shifts and xor operations. After this publication, we have been working on the performance evaluation of the solution to determine the memory overhead together with both computing and communication latencies.

7.3.5. Massive MIMO Cooperative Communications for Wireless Sensor Networks

Participants: Nadjib Achir, Paul Muhlethaler.

This work is done in collaboration with Mérouane Debbah (Supelec, France).

The objective of this work is to propose a framework for massive MIMO cooperative communications for Wireless Sensor Networks. Our main objective is to analyze the performances of the deployment of a large number of sensors. This deployment should cope with a high demand for real time monitoring and should also take into account energy consumption. We have assumed a communication protocol with two phases: an initial training period followed by a second transmit period. The first period allows the sensors to estimate the channel state and the objective of the second period is to transmit the data sensed. We start analyzing the impact of the time devoted to each period. We study the throughput obtained with respect to the number of sensors when there is one sink. We also compute the optimal number of sinks with respect to the energy spent for different values of sensors. This work is a first step to establish a complete framework to study energy efficient Wireless Sensor Networks where the sensors collaborate to send information to a sink. Currently, we are exploring the multi-hop case.

7.4. Industry 4.0 and Wireless Sensor Networks

By the year 2020, it is expected that the number of connected objects will exceed several billions devices. These objects will be present in everyday life for a smarter home and city as well as in future smart factories that will revolutionize the industry organization. This is actually the expected fourth industrial revolution, more known as Industry 4.0. In which, the Internet of Things (IoT) is considered as a key enabler for this major transformation. IoT will allow more intelligent monitoring and self-organizing capabilities than traditional factories. As a consequence, the production process will be more efficient and flexible with products of higher quality.

Several standards have been designed for industrial wireless sensor (IoT) networks such as WirelessHart and ISA100. Both of them are based on the IEEE 802.15.4 standard for the lower layers. More recently, Time Slotted Channel Hopping (TSCH) which is specified in amendment e of the IEEE 802.15.4 standard, uses a time slotted medium access operating on several channels simultaneously. In addition, radio perturbations are mitigated by frequency hopping. TSCH supports star and mesh topologies, as well as multi-hop communication. It has been designed for process automation, process control, equipment monitoring and more generally the Internet of Things. It is a candidate technology for the Industry 4.0. In fact, Industry 4.0 will use more and more the on-demand manufacturing in a highly fexible and widespread environment. Different supply chains located in various regions need to coordinate their actions in a real-time basis with high fidelity. The IoT communicating in a wireless manner will play a major role to achieve this target. Time Slotted Channel Hopping (TSCH) networks are emerging as a promising technology for the Internet of Things and the Industry 4.0 where ease of deployment, reliability, short latency, flexibility and adaptivity are required. However, the strong requirements in terms of short latency and high reliability of such applications are obstacles to its penetration in the Industry 4.0. That is why in 2017 we made three contributions dealing with:

- how to quickly build a TSCH network;
- how to increase the reliability of end-to-end communications;
- how to efficiently schedule the transmissions made for data gathering.

7.4.1. Building an IEEE 802.15.4e TSCH network

Participants: Ines Khoufi, Pascale Minet.

The IEEE 802.15.4e amendment has been designed to meet the requirements of industrial applications with regard to the wireless sensor networks supporting them. Because of its scheduled medium access and multichannel transmissions, the TSCH mode has received much attention. In this study, we focus on the time needed by a node to detect a beacon sent by a TSCH network, as well as on the time needed to build a TSCH network. These times are important for industrial applications where new nodes are inserted progressively, or when failed nodes are replaced. Both times highly depend on the beacon advertisement policy, policy that is not specified in the standard and is under the responsibility of a layer upper than the MAC one. Since beacons are broadcast, they are lost in case of collisions: the vital information they carry is lost. The main problem is how to avoid collisions between two devices that are not neighbors. That is why we propose DBA, a Deterministic Beacon Advertisement algorithm that ensures a regular transmission of beacons without collisions. The goal of DBA is to ensure that beacons are transmitted on all frequencies used by the TSCH network, regularly and without collision. With DBA, the exact value for the maximum time for a joining node to detect a beacon can be computed easily. We use the NS3 Simulator to evaluate this time as well as the the number of message losses, considering different network topologies (star or multihop). Simulation results show that DBA clearly outperforms existing solutions such as Random Vertical and Random Horizontal, two algorithms existing in the state of the art. In addition, DBA is able to provide the exact value of the maximum joining time.These results have been presented at the EUCASS 2017 conference, see [31].

7.4.2. Increasing the reliability of an IEEE 802.15.4e TSCH network

Participants: Ines Khoufi, Pascale Minet.

Our goal is to improve reliability of data gathering in such wireless sensor networks. We present three redundancy patterns to build a reliable path from a source to a destination. The first one is the well-known two node-Disjoint paths. The second one is based on a Triangular pattern, and the third one on a Braided pattern. The reliability provided by each pattern, the delivery time and the overhead in terms of the number of transmissions generated by each pattern as well as the amount of energy consumed by an end-to-end transmission allows us to conclude that the Braided pattern provides the highest reliability but with an overhead approximately twice the overhead of the Disjoint-path pattern and 4/3 the overhead of the Triangular pattern. These performance results are corroborated by simulations performed with NS3 for various configurations. This result has been presented at the NCA 2017 conference ([21]).

7.4.3. Scheduling transmissions in an IEEE 802.15.4e TSCH network

Participants: Ines Khoufi, Pascale Minet.

TSCH provides a multichannel slotted medium access ruled by a periodic schedule and supports multihop communications. This schedule is repeated every slotframe. A slotframe consists of a set of cells, each cell is identified by a (time slot offset, channel offset) pair. The size of a timeslot (e.g. 10 ms by default) allows the transmission of a point-to-point frame and its immediate acknowledgment. The schedule defines for each cell the nodes allowed to transmit and those that should receive. The channel offset is translated into a physical channel depending on the channel hopping sequence of the TSCH network. Channel hopping allows the TSCH to increase its robustness against external perturbations of the radio signal.

In the paper presented at VTC-Fall 2017 [20], we study how applications with data delivery constraints can be supported by a TSCH network. We first propose a framework based on a multislotframe that allows the coexistence of Data Slotframes and Control Slotframes. We then determine a lower bound on the minimum number of slots required to perform data gathering, taking into account the number of channels, the number of interfaces of the sink, the number of packets generated by each sensor node as well as the number of children of the sink. These feasibility conditions are established for two cases: with spatial reuse and without. We propose a debt-based scheduler that for simple topologies, provides a schedule minimizing the slotframe size. We determine the conditions for which an increase in the number of channels or sink's interfaces leads to a shorter data delivery delay. We compare the number of slots needed by data gathering with and without spatial reuse for small configurations. Finally, we consider a network configuration representative of an industrial application and evaluate the performance of the TSCH network in terms of data delivery delay and queue size for each sensor node, using the NS-3 simulator, where the multislotframe has been integrated. Simulation results Simulation results showed that the maximum theoretical delivery delay is never exceeded and the number of messages in the Transmit queue of each sensor node remains small. In addition, the debt-based scheduler builds a valid schedule with the minimum number of slots for the industrial application considered. we can conclude that TSCH with its time-slotted and multichannel medium access provides an efficient support for data gathering.

7.5. Machine Learning for an efficient and dynamic management of network resources and services

7.5.1. Machine Learning in Networks

Participants: Nesrine Ben Hassine, Dana Marinca, Pascale Minet.

This work was done in collaboration with Dominique Barth (UVSQ) .

Content Delivery Networks (CDNs) are faced with an increasing and time varying demand of video contents. Their ability to promptly react to this demand is a success factor. Caching helps, but the question is: which contents to cache? We need to know which resources are needed before they are requested. This anticipation is made possible by using prediction computed by learning techniques.

Machine learning techniques can be used to improve the quality of experience for the end users of Content Delivery Networks (CDNs). In a CDN, the most popular video contents are cached near the end-users in order to minimize the contents delivery latency. Classically, machine learning techniques are classified as supervised or unsupervised. In 2017, we addressed two challenges:

- as a supervised learning, the use of prediction techniques based on regression to evaluate the future popularity of video contents in order to decide which ones should be cached. The popularity of a video content is evaluated by the number of daily requests for this content.
- as an unsupervised learning, the use of clustering techniques to put together videos with similar features. This clustering will reduce the number of prediction methods, called experts, used to provide an accurate prediction.

7.5.2. Prediction of video content popularity

Participants: Nesrine Ben Hassine, Dana Marinca, Pascale Minet.

This work was done in collaboration with Dominique Barth (UVSQ).

We consider various experts, coming from different fields (e.g. statistics, control theory). To evaluate the accuracy of the experts' popularity predictions, we assess these experts according to three criteria: cumulated loss, maximum instantaneous loss and best ranking. The loss function expresses the discrepancy between the prediction value and the real number of requests. We use real traces extracted from YouTube to compare different prediction methods and determine the best tuning of their parameters. The goal is to find the best trade-off between complexity and accuracy of the prediction methods used.
We also show the importance of a decision maker, called forecaster, that predicts the popularity based on the predictions of a selection of several experts. The forecaster based on the best K experts outperforms in terms of cumulated loss the individual experts' predictions and those of the forecaster based on only one expert, even if this expert varies over time.
The paper presented at the Wireless days 2017 conference ([29] is the result of a joint work done in collaboration with Ruben Milocco (Universidad Nacional Comahue, Buenos Aires, Argentina) and Selma Boumerdassi (CNAM, Paris). We focused on predicting the popularity of video contents using Auto-Regressive Moving Average (ARMA) methods applied on a sliding window. These predictions are used to put the most popular video contents into caches. After having identified the parameters of ARMA experts, we compare them with an expert predicting the same number of requests as the previous day. Results show that ARMA experts improve the accuracy of the predictions. Nevertheless, there is no ARMA model that provides the best prediction for all the video contents over all their lifetime. We combine these statistical experts with a higher level of experts, called forecasters. By combining the experts prediction, some forecasters succeed in predicting more accurate values which helped to increase the hit ratio while keeping a correct update ratio. Hence, improving the accuracy of the predictions succeeds in improving the hit ratio. To summarize, we proposed an original solution combining the predictions of several ARMA models. This solution achieves a better Hit Ratio and a smaller Update Ratio than the classical Least Frequently Used (LFU) caching technique.

7.5.3. Clustering of video contents

Participants: Nesrine Ben Hassine, Pascale Minet.

With regard to video content clustering, we proposed an original solution based on game theory that was presented at the CCNC 2017 conference ([30]. This is a joint work with Mohammed-Amine Koulali (Mohammed I University Oujda, Morocco), Mohammed Erradi (Mohammed I University Rabat, Morocco), Dana Marinca (University of Versailles Saint-Quentin) and Dominique Barth (University of Versailles Saint-Quentin). Game theory is a powerful tool that has recently been used in networks to improve the end users' quality of experience (e.g. decreased response time, higher delivery rate). In this paper, the original idea consists in using game theory in the context of Content Delivery Networks (CDNs) to organize video contents into clusters having similar request profiles. The popularity of each content in the cluster can be determined from the popularity of the representative of the cluster and used to store the most popular contents close to end users. A group of experts and a decision-maker predict the popularity of the representative of the cluster. This considerably reduces the number of experts used. More precisely, we model the clustering problem as a

hedonic coalition formation game where the players are the video contents. We proved that this game always converges to a stable partition consisting of different clusters. We determined the best size of the observation window and showed that the play order minimizing the maximum distance to the representative of the cluster is the Rich-to-Poor order, whatever the number of video contents in the interval [20; 200]. The complexity of the coalition game remains very light. Convergence is obtained in a small number of rounds (i.e. less than 35 rounds for 200 video contents). We compare the results of this approach with the clustering obtained by the K-means algorithm, using real traces extracted from YouTube. We also evaluate the complexity of the proposed algorithm. The coalition game outperforms K-means in terms of the average and maximum distances to the representative of the cluster. The execution time is also in favor of the coalition game when the number of contents is higher than or equal to 50. Furthermore, the coalition game can be used to quickly determine the best value of K that is required as an input parameter of the K-means algorithm. Simulation results show that the coalition game provides very good performances.

7.6. Protocols and Models for Wireless Networks - Application to VANETs

7.6.1. Protocols for VANETs

7.6.1.1. TRPM: a TDMA-aware routing protocol for multi-hop communications in VANETs

Participants: Mohamed Elhadad Or Hadded, Paul Muhlethaler, Anis Laouiti.

The main idea of TRPM is to select the next hop using the vehicle position and the time slot information from the TDMA scheduling. Like the GPSR protocol, we assume that each transmitting vehicle knows the position of the packet's destination. In TRPM, the TDMA scheduling information and the position of a packet's destination are sufficient to make correct forwarding decisions at each transmitting vehicle. Specifically, if a source vehicle is moving in area x_i, the locally optimal choice of next hop is the neighbor geographically located in area x_{i+1} or x_{i-1} according to the position of the packet's destination. As a result, the TDMA slot scheduling obtained by DTMAC can be used to determine the set of next hops that are geographically closer to the destination. In fact, each vehicle that is moving in the area x_i can know the locally optimal set of next hops that are located in adjacent areas x_{i+1} or x_{i-1} by observing the set of time slots $S_{(i+3)\%3}$ or $S_{(i+1)\%3}$, respectively. We consider the same example presented above when vehicle G as the destination vehicle that will broadcast a message received from vehicle A. As shown in Figure 3 , only two relay vehicles are needed to ensure a multi-hop path between vehicle A and G (one relay node in the area x_2 and another one in the area x_3).

In the following, the DTMAC protocol has been used by the vehicles to organize the channel access. The TDMA slot scheduling obtained by DTMAC is illustrated in Figure 3 . Firstly, vehicle A forwards a packet to B, as vehicle A uses its frame information to choose a vehicle that is accessing the channel during the set S_1. Upon receiving the packet for forwarding, vehicle B will choose by using its frame information a vehicle that's accessing the channel during the set of time slots S_2 (say vehicle D). Then, vehicle D will forward the packet to G, as G is moving in area x_4 (accessing the channel during the set S_0) and it is the direct neighbor of vehicle D. By using DTMAC as the MAC layer, we can note that the path A-B-D-G is the shortest, in terms of the number of hops as well as the end-to-end delay which is equal to 6 time slots (2 time slots between t_0 and t_2 as t_2 is the transmission slot for vehicle B, then 2 time slots between t_2 and t_4 as t_4 is the transmission slot for vehicle D and finally 2 time slots between t_4 and t_0 as t_0 is the transmission slot in which vehicle G will broadcast the message received from vehicle A).

The idea of TRPM [16] is the following . Whenever a vehicle i accessing the channel during the set S_k wants to send/forward an event-driven safety message, it constructs two sets of candidate forwarders based on its frame information FI as follows, where $TS(j)$ indicates the time slot reserved by vehicle j.

- $A_i = \{j \in N(i) \mid TS(j) \in S_{(k+1)\%3}\}$ *// The set of vehicles that are moving in the adjacent right-hand area.*
- $B_i = \{j \in N(i) \mid TS(j) \in S_{(k+2)\%3}\}$ *// The set of vehicles that are moving in the adjacent left-hand area.*

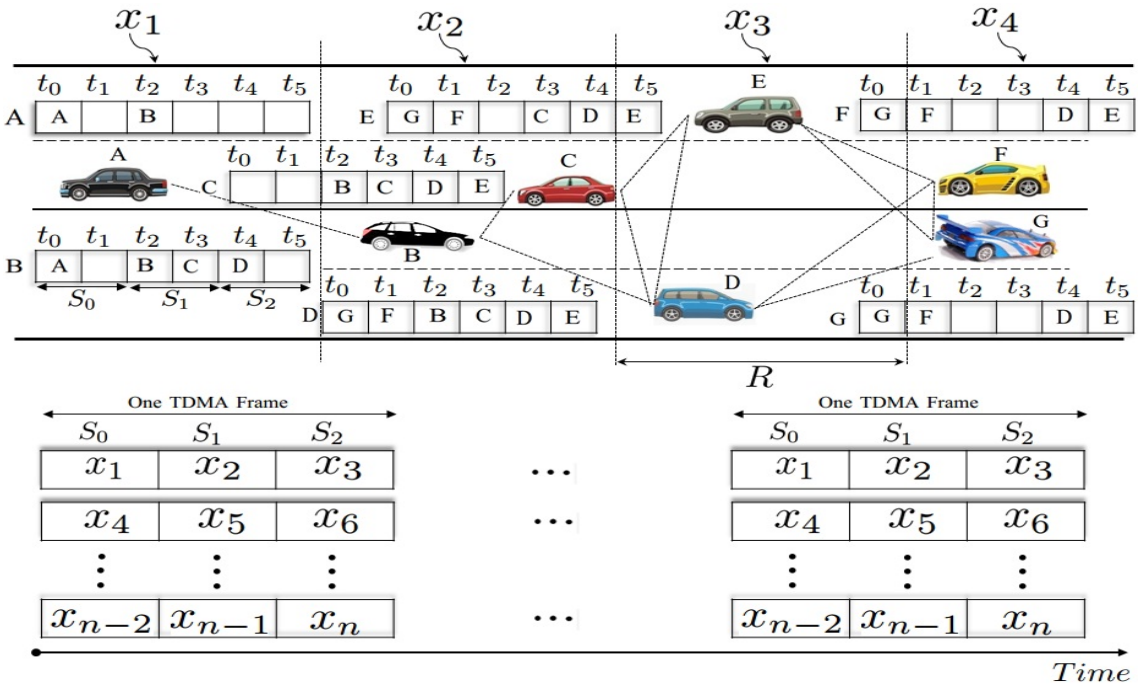

Figure 3. VANET network using DTMAC scheduling scheme.

Each source vehicle uses the position of a packet's destination and the TDMA scheduling information to make packet forwarding decisions. In fact, when a source vehicle i is moving behind the destination vehicle, it will select a next hop relay that belongs to set B_i; when the transmitter is moving in front of the destination vehicle, it will select a forwarder vehicle from those in set A_i. For each vehicle i that will send or forward a message, we define the normalized weight function WHS (Weighted next-Hop Selection) which depends on the delay and the distance between each neighboring vehicle j. WHS is calculated as follows:

$$WHS_{i,j} = \alpha * \frac{\Delta t_{i,j}}{\tau} - (1-\alpha) * \frac{d_{i,j}}{R} \qquad (1)$$

Where:

- τ is the length of the TDMA frame (in number of time slots).
- j is one of the neighbors of vehicle i, which represents the potential next hop that will relay the message received from vehicle i.
- $\Delta t_{i,j}$ is the gap between the sending slot of vehicle i and the sending slot of vehicle j.
- $d_{i,j}$ is the distance between the two vehicles i and j, and R is the communication range.
- α is a weighted value in the interval $[0, 1]$ that gives more weight to either distance or delay. When α is high, more weight is given to the delay. Otherwise, when α is small, more weight is given to the distance.

We note that the two weight factors $\frac{\Delta t_{i,j}}{\tau}$ and $\frac{d_{i,j}}{R}$ are in conflict. For simplicity, we assume that all the factors should be minimized. In fact, the multiplication of the second weight factor by (-1) allows us to transform a maximization to a minimization. Therefore, the forwarding vehicle for i is the vehicle j that is moving in an adjacent area for which $WHS_{i,j}$ is the lowest value.

The simulation results reveal that our routing protocol significantly outperforms other protocols in terms of average end-to-end delay, average number of relay vehicles and the average delivery ratio.

We have developed an analytical model to evaluate the packet loss rate and the end-to-end delay for safety messages transmitted in vehicular networks over long distances when TRPM is used as a routing protocol, see refhadded:hal-01617924. Comparisons of realistic simulation results, carried out using ns-2 and MOVE/SUMO, and analytical results show that the analytical model proposed provides close approximations for the end-to-end delay and packet loss rate for the different scenarios considered.

7.6.1.2. Trust-CTMAC: A Trust Based Scheduling Algorithm

Participants: Mohamed Elhadad Or Hadded, Paul Muhlethaler, Anis Laouiti.

In Vehicular Ad hoc NETworks, communication is possible both between the vehicles themselves and between the vehicles and the infrastructure. These applications need a reliable and secure broadcast system that takes into consideration the security issues in VANETs, the high speed of nodes and the strict QoS requirements. For these reasons, we propose a trust-based and centralized TDMA-based MAC protocol. Our solution will permit Road Side Units (RSUs) to manage time slot assignment by avoiding malicious nodes and by minimizing message collision. The experiments carried out and the results obtained prove the effectiveness of our approach.

We present a trust based centralized TDMA scheduling mechanism which aims to isolate and prevent malicious vehicles from accessing the channel. This is accomplished by serving only the slot reservation requests of vehicles that have trust values greater than a trust threshold. In Trust-CTMAC, each RSU maintains additional data structure called Trust Counters Table (TCT) and Malicious Vehicles Table (MVT) for all vehicles within its communication range based on the list of properties shown in Table 1. The TCT and the FI information are periodically broadcasted by the RSU for ecah time interval of $100ms$. So each vehicle can identify and isolate malicious vehicles among all neighboring nodes based on the TCT information received from its RSU, which can protect the radio channel from any potential damage caused by the malicious vehicles. An RSU declares a vehicle as a malicious node if the corresponding trust value falls below a trust threshold.

7.6.1.3. A Flooding-Based Location Service in VANETs

Participants: Selma Boumerdassi, Paul Muhlethaler.

Table 1. Threat lists that are checked in our trust platform

Threat Name	Description	Level
Message Saturation	A huge number of a vehicle packets do not include any form of identification information	3 (high)
False GNSS (Global Navigation Satellite System) Signals	A vehicle is sending messages with false geographic information	3 (high)
Slot reservation attack	A vehicle requests different slots during the same frame	3 (high)
Malicious MAC behavior	A vehicle is sending data in another slot different to its reserved one	4 (Critical)
Malicious isolation	Some vehicle functionalities are disabled (create, process, receive and send messages) caused by the installation of a malware	3 (high)
Denial of access to incoming messages	A vehicle may be unlinked if it receives a huge number of messages.	4 (Critical)
Frame information poisoning	The frame information is falsified by a vehicle	3 (high)
Identity spoofing	A vehicle is using a wrong node type in order to act as an RSU	3 (high)

This work has been done in collaboration with Eric Renault, Telecom Sud Paris.

We have designed and analyzed a location service for VANETs; such a service can be used in Location-based routing protocols for VANETs. Our protocol is a proactive flooding-based location service that drastically reduces the number of update packets sent over the network as compared to traditional flooding-based location services. This goal is achieved by partially forwarding location information at each node. A mathematical model and some simulations are proposed to show the effectiveness of this solution. Cases for 1D, 2D and 3D spaces are studied for both deterministic and probabilistic forwarding decisions. We compare our protocol with the Multi-Point Relay (MPR) technique which is used in the OLSR protocol and determine the best technique according to the network conditions.

7.6.2. Models for Wireless Networks and VANETs

7.6.2.1. Performance analysis of IEEE 802.11 broadcast schemes with different inter-frame spacings

Participants: Younes Bouchaala, Paul Muhlethaler, Nadjib Achir.

This work has been done in collaboration with Oyunchimeg Shagdar (Vedecom).

We have started to build a model which analyzes the performance of IEEE 802.11p managing different classes of priorities. The differentiation of traffic streams is obtained with different inter-frame spacings: AIFSs (for Arbitration Inter Frame Spacings) and with different back-off windows: CWs (for Collision Windows). This model is based on a Markov model where the state is the remaining number of idle slots that a packet of a given class has to wait before transmission. However, in addition to this Markov model for which we compute a steady state we also consider the Markov chain which counts the number of idle slots after the smallest AIFS. As a matter of fact the probability these states are not evenly distributed since with different AIFSs the arrival rate is not constant when the number of idle slots experienced after the smallest AIFS varies. The resolution of the steady state of these two inter-mixed Markov chains lead to non linear and intertwined equations that can be easily solved with a software such as Maple. With the model we have obtained, we can compute the delivery rate of packets of different classes and show the influence of system parameters: AIFSs and CWs. The preliminary results show a a very strong influence of different AIFSs on the performance for each traffic streams, see [13].

7.6.2.2. Model and optimization of CSMA

Participants: Younes Bouchaala, Paul Muhlethaler, Nadjib Achir.

This work has been done in collaboration with Oyunchimeg Shagdar (Vedecom).

We have studied the maximum throughput of CSMA in scenarios with spatial reuse. The nodes of our network form a Poisson Point Process (PPP) of a one- or two-dimensional space. The one-dimensional PPP well represents VANETs. To model the effect of Carrier Sense Multiple Access (CSMA), we give random marks to our nodes and to elect transmitting nodes in the PPP we choose the nodes with the smallest marks in their neighborhood, this is the Matern hardcore selection process. To describe the signal propagation, we use a signal with power-law decay and we add a random Rayleigh fading. To decide whether or not a transmission is successful, we adopt the Signal-over-Interference Ratio (SIR) model in which a packet is correctly received if its transmission power divided by the interference power is above a capture threshold. We assume that each node in our PPP has a random receiver at a typical distance. We choose the average distance to its closest neighbor. We also assume that all the network nodes always have a pending packet. With these assumptions, we analytically study the density of throughput of successful transmissions and we show that it can be optimized with the carrier-sense threshold. The model makes it possible to analytically compute the performance of a CSMA system and gives interesting results on the network performance such as the capture probability when the throughput is optimized, and the effect on a non-optimization of the carrier sense threshold on the throughput. We can also study the influence of the parameters and see their effects on the overall performance. We observe a significant difference between 2D an 1D networks.

We have built two models to compare the spatial density of successful transmissions of CSMA and Aloha. To carry out a fair comparison, we optimize both schemes by adjusting their parameters. For spatial Aloha, we can adapt the transmission probability, whereas for spatial CSMA we have to find the suitable carrier sense threshold. The results obtained show that CSMA, when optimized, outperforms Aloha for nearly all the parameters of the network model values and we evaluate the gain of CSMA over Aloha. We also find interesting results concerning the effect of the model parameters on the performance of both Aloha and CSMA. The closed formulas we have obtained provide immediate evaluation of performance, whereas simulations may take minutes to give their results, see [14]. Even if Aloha and CSMA are not recent protocols, this comparison of spatial performance is new and provides interesting and useful results.

For Aloha networks, when we study transmissions over the average distance to the closest neighbor, the optimization does not depend on the density of nodes, which is a very interesting property. Thus in Aloha networks, the density of successful transmissions easily scales linearly in λ when we vary λ whereas in CSMA networks the protocol must be carefully tuned to obtain this scaling.

With CSMA, we have also shown that this density of throughput (when optimized) scales with the density of nodes if we study the throughput is measured between the nodes to their closest neighbors. We have mathematically justified this property.

7.6.2.3. Adaptive CSMA

Participants: Nadjib Achir, Younes Bouchaala, Paul Muhlethaler.

This work has been done in collaboration with Oyunchimeg Shagdar (Vedecom).

Using the model we have built for CSMA, we have shown that when optimized with the carrier sense detection threshold P_{cs}, the probability p^* of transmission for a node in the CSMA network does not depend on the density of nodes λ. In other words when the CSMA is optimized to obtain the largest density of successful transmissions (communication from nodes to their neighbors), p^* is constant. We have verified this statement on several examples and we think that a formal proof of this remark is possible using scaling arguments. The average access delay is a direct function of the probability of transmission p. Thus the average delay when the carrier sense detection threshold is optimized is a constant D_{target} which does not depend on λ. A stabilization algorithm which adapts P_{cs} to reach the D_{target} can thus be envisioned. Another stabilization algorithm adapts P_{cs} so that the mean number of neighbors of a node is N_{target} a given number of nodes which only depends on the network parameters and not on the network density. A third stabilization algorithm adapts P_{cs} so that the channel busy ratio (CBR) is near a given target.

We have justified theoretically all these algorithms and simulated their behavior. The simulations well justify the theoretical analysis.

7.6.2.4. Optimizing spatial throughput in device-to-device networks

Participants: Bartek Blaszczyszyn, Paul Keeler, Paul Muhlethaler.

Results are presented for optimizing device-to-device communications in cellular networks, while maintaining spectral efficiency of the base-station-to-device downlink channel. We build upon established and tested stochastic geometry models of signal-to-interference ratio in wireless networks based on the Poisson point process, which incorporate random propagation effects such as fading and shadowing. A key result is a simple formula, allowing one to optimize the device-to-device spatial throughput by suitably adjusting the proportion of active devices, see [19]. These results can lead to further investigation as they can be immediately applied to more sophisticated models such as studying multi-tier network models to address coverage in closed access networks.

7.6.2.5. Model and analysis of Coded Slotted Aloha (CSA) with capture

Participants: Ebrahimi Khaleghi, Cedric Adjih, Paul Muhlethaler.

This work has been done in collaboration with Amira Alloum, Nokia Bell Labs.

Motivated by scenario requirements for 5G cellular networks, we have studied one among the protocols candidate to the massive random access: the family of random access methods known as Coded Slotted ALOHA (CSA). Recent body of research has explored aspects of such methods in various contexts, but one aspect has not been fully taken into account: the impact of the path loss, which is a major design constraint in long-range wireless networks. We have explored the behavior of CSA, by focusing on the path loss component correlated to the distance to the base station. Path loss provides opportunities for capture, improving the performance of CSA. We have revised methods for estimating CSA behavior. We have provided bounds of performance and derived the achievable throughput. We have extensively explore the key parameters, and their associated gain (experimentally). Our results has shed light on the open question of the optimal distribution of repetitions in actual wireless networks.

7.6.2.6. Mobility Prediction in Vehicular Networks : An Approach through Hybrid Neural Networks under Uncertainty

Participants: Soumya Banerjee, Samia Bouzefrane, Paul Muhlethaler.

Conventionally, the exposure regarding knowledge of the inter vehicle link duration is a significant parameter in *Vehicular Networks* to estimate the delay during the failure of a specific link during the transmission. However, the mobility and dynamics of the nodes is considerably higher in a smart city than on highways and thus could emerge a complex random pattern for the investigation of the link duration, referring all sorts of uncertain conditions. There are existing link duration estimation models, which perform linear operations under linear relationships without imprecise conditions. Anticipating, the requirement to tackle the uncertain conditions in *Vehicular Network*s, this paper presents a hybrid neural network-driven mobility prediction model. The proposed hybrid neural network comprises a *Fuzzy Constrained Boltzmann machine (FCBM*), which allows the random patterns of several vehicles in a single time stamp to be learned. The several dynamic parameters, which may make the contexts of *Vehicular Networks* uncertain, could be vehicle speed at the moment of prediction, the number of leading vehicles, the average speed of the leading vehicle, the distance to the subsequent intersection of traffic roadways and the number of lanes in a road segment. In this paper, a novel method of hybrid intelligence is initiated to tackle such uncertainty. Here, *the Fuzzy Constrained Boltzmann Machine (FCBM)* is a stochastic graph model that can learn joint probability distribution over its visible units (say n) and hidden feature units (say m). It is evident that there must be a prime driving parameter of the holistic network, which will monitor the interconnection of weights and biases of *the Vehicular Network* for all these features. The highlight of this paper is that the prime driving parameter to control the learning process should be a fuzzy number, as fuzzy logic is used to represent the vague and and uncertain parameters. Therefore, if uncertainty exists due to the random patterns caused by vehicle mobility, the proposed Fuzzy Constrained Boltzmann Machine could remove the noise from the data representation. Thus, the proposed model will be able to predict robustly the mobility in VANET, referring any instance of link failure under *Vehicular Network* paradigm.

7.6.3. Reliable routing architecture

Participants: Mohamed Hadded, Anis Laouiti, Paul Muhlethaler.

Flooding scheme represents one of the fundamental operation in wireless mesh networks. It plays an important role in the design of network and application protocols. Many existing flooding solutions have been studied to address the flooding issues in mesh networks. However, most of them are not able to operate efficiently where there are network equipment failures. In this work, we consider nodes failures and we build the flooding tree the maximum expectation of the throughput (taking into account the potential unavailability of certain nodes). After a formal stochastic definition of the problem, we show how to use a tabu search algorithm, to solve this optimization problem.

FUN Project-Team

7. New Results

7.1. Routing

Participants: Nathalie Mitton, Julien Vandaele.

Wireless sensor and actuator/robot networks need some routing mechanisms to ensure that data travel the network to the sink with some guarantees. The FUN research group has investigated different routing paradigms.

Geographic routing has gained much attention as a basic routing primitive in wireless sensor networks due to its memory-less, scalability, efficiency and low overhead features. Greedy forwarding is the simplest geographic routing scheme, it uses the distance as a forwarding criterion. Nevertheless, it may suffer from communication holes, where no next hop candidate is closer to the destination than the node currently holding the packet. For this purpose, a void handling technique is needed to recover from the void problem and successfully deliver data packets if a path does exist between source and destination nodes. Many approaches have been reported to solve this issue at the expense of extra processing and or overhead. [19] proposes GRACO, an efficient geographic routing protocol with a novel void recovery strategy based on ant colony optimization (ACO). GRACO is able to adaptively adjust the forwarding mechanism to avoid the blocking situation and effectively deliver data packets. Compared to GFG, one of the best performing geographic routing protocols, simulation results demonstrate that GRACO can successfully find shorter routing paths with higher delivery rate, less control packet overhead and shorter end-to-end delay.

Betweenness centrality metrics usually underestimate the importance of nodes that are close to shortest paths but do not exactly fall on them. In [16], [41], we reevaluate the importance of such nodes and propose the ρ-geodesic betweenness centrality, a novel metric that assigns weights to paths (and, consequently, to nodes on these paths) according to how close they are to shortest paths. The paths that are just slightly longer than the shortest one are defined as quasi-shortest paths, and they are able to increase or to decrease the importance of a node according to how often the node falls on them. We compare the proposed metric with the traditional, distance-scaled, and random walk betweenness centralities using four network datasets with distinct characteristics. The results show that the proposed metric, besides better assessing the topological role of a node, is also able to maintain the rank position of nodes overtime compared to the other metrics; this means that network dynamics affect less our metric than others. Such a property could help avoid, for instance, the waste of resources caused when data follow only the shortest paths and reduce associated costs.

To illustrate the data routing over a real demo, in [39], we show a webcam view of the testbed with remotely controlled lighting (ceiling LEDs and a mobile robot carrying a torch). A tight grid of 256 sensors will be used to collect light information. We display live updates of the resulting heatmap, live energy profiles and other performance metrics.

7.2. Security, Safety and Verification

Participants: Nathalie Mitton, Allan Blanchard, Simon Duquennoy.

Current practices of fault-tolerant network design ignore the fact that most network infrastructure faults are localized or spatially correlated (i.e., confined to geo-graphic regions). Network operators require new tools to mitigate the impact of such region-based faults on their infrastructures. Utilizing the support from the U.S. Department of Defense, and by consolidating a wide range of theories and solutions developed in the last few years, [12] designs RAPTOR, an advanced Network Planning and Management Tool that facilitates the design and provisioning of robust and resilient networks.The tool provides multi-faceted network design, evaluation, and simulation capabilities for network planners. Future extensions of the tool currently being worked upon not only expand the tool's capabilities, but also extend these capabilities to heterogeneous interdependent networks such as communication, power, water, and satellite networks.

IoT applications often utilize the cloud to store and provide ubiquitous access to collected data. This naturally facilitates data sharing with third-party services and other users, but bears privacy risks, due to data breaches or unauthorized trades with user data. To address these concerns, we present Pilatus, a data protection platform where the cloud stores only encrypted data, yet is still able to process certain queries (e.g., range, sum). More importantly, Pilatus features a novel encrypted data sharing scheme based on re-encryption, with revocation capabilities and in situ key-update. The solution proposed in [37], [56] includes a suite of novel techniques that enable efficient partially homomorphic encryption, decryption, and sharing. We present performance optimizations that render these cryptographic tools practical for mobile platforms. We implement a prototype of Pilatus and evaluate it thoroughly. Our optimizations achieve a performance gain within one order of magnitude compared to state-of-the-art realizations; mobile devices can decrypt hundreds of data points in a few hundred milliseconds. Moreover, we discuss practical considerations through two example mobile applications (Fitbit and Ava) that run Pilatus on real-world data.

7.3. Alternative communication paradigms

Participants: Antonio Costanzo, Valeria Loscri.

Nowadays, the always growing of connected objects and the strong demand to downsizing the devices in order to make the Internet of Things (IoT) paradigm more pervasive and ubiquitous, has motivated academic and industry people to investigate from one side mechanisms able to adapt quickly to the rapid external changes and to the quality of Services (QoS) parameters defined by the users and imposed by the adoption of new services and from another side, the investigation of portion of spectrum that have not been considered till this moment such as Terahertz band.

Bearing that in mind, we envisaged the possibility to leverage in a synergic way the Software Defined Radio (SDR) paradigm and the controlled mobility of mobiles wireless devices in order to adopt the most suitable modulation scheme and the best position with the objective to improve the network connectivity and coverage area [13].

On the other hand, spectrum scarcity and growing demand of nanocommunication systems have motivated researchers to investigation novel channel models in different portions of spectrum, namely in the THz band.

The fervent research activity in this direction is also motivated by the recent technological advances in new types of materials (e.g. graphene, novel metamaterials) presenting specific features suitable for this frequency spectrum and for the growing demand of downsizing antenna dimension.

In [15], we have investigated the chirality effect and Giant Optical Activity (GOA) and their impact when assuming different power allocation techniques.

On the other hand, when the nature of the matter and the interactions of specific particles and (quasi)particles such as phonons and photons are considered, there is a growing interest to investigate alternative communication paradigms based on these specific phenomena. In [14] we have performed an information theory analysis based on the generation of phonons elements when a source power as a cellphone is applied on biological tissue. The lesson learnt in this works is based on the consideration that where is heat transport it is possible to associate a communication paradigm. Follow this reasoning, in [50] we have revised the most recent advancement in terms of Visible Light Communication (VLC). Specifically, we have investigated Software Defined paradigm for VLC, in order to sketch out the main research directions for this new research domain.

7.4. Self-Organization

Participants: Nathalie Mitton, Valeria Loscri, Farouk Mezghani, Simon Duquennoy, Anjalalaina Jean Cristanel Razafimandimby.

7.4.1. Bayesian communications

In the last few years, Internet has become a very important vector of information sharing. Beyond the interconnection of computers and devices, there is still an important expansion capability, thanks to the capacity to interconnect heterogeneous devices. This extension of Internet known as Internet of Things (IoT) leads to (inter)connection of billions of objects. Nevertheless, IoT paradigm raises many challenges, such as the need to manage a massive amount of data generated by sensing devices. It was observed that, with the increase of sensors density, the redundancy of data increases. Thus, uploading raw data to the cloud can become extremely inefficient.

In order to address this issue, we proposed a Bayesian Inference Approach (BIA), able to remove a great amount of spatio-temporal correlated data [46], [10], [35].

In order to validate these approaches it was considered that experiments in real-world scenarios were needed. More specifically, we considered indoor tests in [46] and agricultural/outdoor experiments in [47]

7.4.2. Alert diffusion

Opportunistic communications present a promising solution as a disaster network recovery in emergency situations such as hurricanes, earthquakes and floods where infrastructure might be damaged. Recent works have proposed opportunistic-based alert diffusion approaches useful for trapped survivors. However, two main features were left behind. On the one hand, these works do not consider the assortment of networks integrated in mobile devices (e.g. WiFi-Direct, WiFi ad-hoc, cellular, bluetooth) and the choice of the network interface is left to the user who has no idea what is best or might be in a physical or psychological distress that impede the efficient selection. On the other hand, most of these works are based on selfish diffusion which might drain quickly the battery power. Moreover, they do not consider various energy level, which obviously influences the alert diffusion scheme. [17], [27], [28], [44] propose COPE and its demo, a cooperative opportunistic alert diffusion approach for disaster scenario that considers mobile devices that come with multiple network interfaces and with various battery power level. In order to maintain mobile devices alive longer, survivors form cliques and zones in which they diffuse alternately and periodically alert messages until reaching a potential rescuers team. Simulation results show that COPE largely outperforms the selfish diffusion scheme in terms of energy consumption while guaranteeing an important alert delivery success.

7.4.3. Consensus-based Leader election

In low-power wireless networking, new applications such as cooperative robots or industrial closed-loop control demand for network-wide consensus at low-latency and high reliability. Distributed consensus protocols is a mature eld of research in a wired context, but has received little attention in low-power wireless settings. In [21], [55], we present A^2: Agreement in the Air, a system that brings distributed consensus to low-power multi-hop networks. A^2 introduces Synchrotron, a synchronous transmissions kernel that builds a robust mesh by exploiting the capture effect, frequency hopping with parallel channels, and link-layer security. A^2 builds on top of this reliable base layer and enables the two-and three-phase commit protocols, as well as network services such as group membership, hopping sequence distribution and re-keying. We evaluate A^2 on four public testbeds with different deployment densities and sizes. A^2 requires only 475 ms to complete a two-phase commit over 180 nodes. The resulting duty cycle is 0.5% for 1-minute intervals. We show that A^2 achieves zero losses end-to-end over long experiments, representing millions of data points. When adding controlled failures, we show that two-phase commit ensures transaction consistency in A^2 while three-phase commit provides liveness at the expense of inconsistency under specific failure scenarios.

7.5. Smart Cities

Participants: Nathalie Mitton, Valeria Loscri, Abdoul Aziz Mbacke.

Smart cities are a key factor in the consumption of materials and resources. As populations grow and resources become scarcer, the efficient usage of these limited goods becomes more important. Building on and integrating with a huge amount of data, the cities of the future are becoming a realization today. There are millions of sensors in place already, monitoring various things in metropolises. In the near future, these sensors will multiply until they can monitor everything from streetlights and trashcans to road conditions and energy consumption. In this context, effective strategies or solutions for refining data sets can play a key role. Based on these premises, we propose in [32] intelligent and adaptive filtering mechanisms as a service (FIlAAS) integrated in the VITAL-OS middleware and will show their feasibility and their effectiveness in the smart city context.

Connecting all these devices to a cloud encompasses the execution of many network tasks at the *edge* and in particular on constrained gateways by low computational resources capabilities. Moreover, these gateways have to deal with the plethora of disparate technologies available in the IoT landscape. To cope with these issues, we introduce a Lightweight Edge Gateway for the Internet of Things (LEGIoT) architecture [18]. It relies on the modular characteristic of microservices and the flexibility of lightweight virtualization technologies to guarantee an extensible and flexible solution. In particular, by combining the implementation of specific frameworks and the benefits of container-based virtualization, our proposal enhances the suitability of edge gateways towards a wide variety of IoT protocols/applications (for both downlink and uplink) enabling an optimized resource management and taking into account requirements such as energy efficiency, multi-tenancy, and interoperability. LEGIoT is designed to be hardware agnostic and its implementation has been tested within a real sensor network. Achieved results demonstrate its scalability and suitability to host different applications meant to provide a wide range of IoT services.

In parallel, we proposed a MOOC in the framework of the IPL CityLab project (See Section 9.2.1), whose working documents are available online [51], [52], [53], [54].

7.6. Smart Grids

Participants: Nathalie Mitton, Jad Nassar.

The Internet of Thing is a on going revolution which promises to interconnect most of our world with billions of connected devices. Hence, data routing and prioritization in IoT is a main challenge in this gigantic network. This is all the more true for the Smart Grids data management where heterogeneous applications and signaling messages have different requirements in terms of reliability, latency and priority. So far, standards on Smart Grid have recommended the use of RPL (Routing Protocol for Low-Power and Lossy networks) protocol for distributing commands over the grid. RPL assures Quality of Service (QoS) at the network layer in wireless sensor networks through the logical subdivision of the network in multiple instances, each one relying on a specific Objective Function. However, RPL is not optimized for Smart Grids, as its main objective function and its associated metric does not allow for QoS differentiation. In order to overcome this, in [31], [45] we propose *OFQS* an objective function with a multi-objective metric that considers the delay and the remaining energy in the battery nodes alongside with the quality of the links. Our function automatically adapts to the number of instances (traffic classes) providing a QoS differentiation based on the different Smart Grid applications requirements. Simulations show that our proposal provides a low packet delivery latency and a higher packet delivery ratio while extending the lifetime of the network compared to literature solutions.

In the same spirit, we have proposed QoSGRACO [36], a routing protocol which takes account of the Quality of Service (QoS) of NAN's traffic by using colored pheromones ant colonies. We show, through simulations, that QoS-GRACO is able to satisfy NAN's requirements, especially in terms of delay and reliability.

7.7. Vehicular networks and smart car platforms

Participants: Nathalie Mitton, Valeria Loscri.

Roadside Units (RSUs) are an important component of vehicular ad hoc networks (VANETs). In a VANET, RSUs are deployed at intersections or some points along a road to help improve network connectivity, data delivery, and thus network services to vehicles. Therefore, RSU deployment has a big impact on the network performance and becomes an important issue in the design of a VANET. In general, RSU deployment is costly. To achieve a good tradeoff between deployment cost and network performance, it is expected to optimize the deployment of RSUs in a VANET. To address this challenge, considerable research work has been conducted on the optimization of RSU deployment for VANET.

To optimize the RSU deployment, the notion of centrality in a social network to RSU deployment has been introduced [40], and use it to measure the importance of an RSU position candidate in RSU deployment. Based on the notion of centrality, we propose a centrality-based RSU deployment approach and formulate the RSU deployment problem as a linear programing problem with the objective to maximize the total centrality of all position candidates selected for RSU deployment under the constraint of a given deployment budget.

Nowadays, many vehicular applications are emerging, arising from entertainment and road safety. The paradigm of Internet of Vehicles (IOV) is proliferating and it is an inevitable convergence of the existing mobile Internet and the concept of Internet-of-Things (IoT). In IoV, Internet of applications go on "wheels", then converting the existing vehicles into "digital cars" equipped with several technologies and sensors. The meeting of mobility with social interactions rises to a particular class of social networks, Vehicular Social Networks [20]. Vehicular social networks are online social networks where the social interactions are built *on-the-fly*, due to the opportunistic links in vehicular networks. In this context, the reliability of message dissemination becomes very important and the adding of social components allows the definition of trust parameter to be directly included in the forwarding technique.

On the other hand, the huge amount of sensors deployed in a car, are arising new types of applications and services, even if data are "confined" in the car.

However, with the increasing number of functionalities, computing and communication resources, that have to be handled by the On Board Unit (OBU), is constantly growing. OBUs are embedded systems with limited hardware resources and a critical software design that makes also a simple update procedure a not trivial operation. These constraints are combined with the typical software lifetime, which is much shorter than the lifetime of mechanical and technological components. Bearing all these points in mind, we have identified in the use of lightweight virtualization technologies a suitable mechanism, which allows to design an OBU that can satisfy several requirements in terms of efficient software and hard- ware resources management [29], [30].

7.8. Robots

Participants: Nathalie Mitton, Valeria Loscri, Anjalalaina Jean Cristanel Razafimandimby.

With the advent of the Internet of Things (IoT) and robotics on one hand and of Cloud Computing on the other hand, people have witnessed a shift in the way they can interact and communicate with their things and their environment. Together with these main concepts, the vision of Robot as a Service can be considered and applied to different contexts and domains. Unfortunately, Remote programming and control of heterogeneous robots are not always possible, which ends up as difficult tasks requiring advanced skills. in order to face these challenges, Open Mobile Cloud Robotics Interface (OMCRI) has been proposed in [26]. It represents an extension of OCCI platform based on the Robot-as-Service paradigm. OMCRI presents interesting features such as modularity and extensibility.

Another interesting concept recently introduced by ABY Research, is the Internet of Robotic Things (IoRT), representing a dynamic actuation. In [48], to realize an efficient deployment and an effective coverage by also keeping a good communication quality, we have proposed an IoT- based and neural network control scheme. The neural network controller, in turn, is completely distributed and mimics perfectly the IoT-based approach. Results show that our approaches are efficient, in terms of convergence time, connectivity, and energy consumption.

7.9. MAC mechanisms

Participants: Nathalie Mitton, Simon Duquennoy, Viktor Toldov.

In the era of the Internet of Things (IoT), the number of connected devices is growing dramatically. Often, connected objects use Industrial, Scientific and Medical (ISM) radio bands for communication. These kinds of bands are available without license, which facilitates development and implementation of new connected objects. However, it also leads to an increased level of interference in these bands. Interference not only negatively affect the Quality of Service, but also causes energy losses, which is especially unfavorable for the energy constrained Wireless Sensor Networks (WSN). In [11], the impact of the interference on the energy consumption of the WSN nodes is studied experimentally. The experimental results were used to estimate the lifetime of WSN nodes under conditions of different levels of interference. Then, a Thompson sampling based Cognitive Radio adaptive solution is proposed and evaluated via both, simulation and hardware implementation. Results show that this approach finds the best channel quicker than other state of the art solutions. Based on a set of experimentations, an adaptive WildMAC MAC layer protocol is proposed and evaluated experimentally.

In parallel, synchronized communication has recently emerged as a prime option for low-power critical applications. Solutions such as Glossy or Time Slotted Channel Hopping (TSCH) have demonstrated end-to-end reliability upwards of 99.9%. In this context, the IETF Working Group 6TiSCH is currently standardizing the mechanisms to use TSCH in low-power IPv6 scenarios [42] identifies a number of challenges when it comes to implementing the 6TiSCH stack [43]. It shows how these challenges can be addressed with practical solutions for locking, queuing, scheduling and other aspects. With this implementation as an enabler, we present an experimental validation and comparison with state-of-the-art MAC protocols. We conduct fine-grained energy profiling, showing the impact of link-layer security on packet transmission. We evaluate distributed time synchronization in a 340-node testbed, and demonstrate that tight synchronization (hundreds of microseconds) can be achieved at very low cost (0.3% duty cycle, 0.008% channel utilization). We finally compare TSCH against traditional MAC layers: low-power listening (LPL) and CSMA, in terms of reliability, latency and energy. We show that with proper scheduling, TSCH achieves by far the highest reliability, and outperforms LPL in both energy and latency.

7.10. RFID

Participants: Nathalie Mitton, Abdoul Aziz Mbacke, Ibrahim Amadou, Gabriele Sabatino.

The advent of RFID (Radio Frequency Identification) has allowed the development of numerous applications. Indeed, solutions such as tracking of goods in large areas or sensing in smart cities are now made possible. However, such solutions encounter two main issues, first is inherent to the technology itself which is readers collisions, the second one being the gathering of read data up to a base station, potentially in a multihop fashion. While the first one has been a main research subject in the late years, the next one has not been investigated for the sole purpose of RFID, but rather for wireless adhoc networks. This multihop tag information collection must be done in regards of the application requirements but it should also care for the deployment strategy of readers to take advantage of their relative positions, coverage, reading activity and deployment density to avoid interfering between tag reading and data forwarding. To the best of our knowledge, the issue for a joint scheduling between tag reading and forwarding has never been investigated so far in the literature, although important. [24] addresses the anti-collision issue in mobile environments. In [23], we propose two new distributed, crosslayer solutions meant for the reduction of collisions and better efficiency of the RFID system but also serve as a routing solution towards a base station. Simulations show high levels of throughput while not lowering on the fairness on medium access staying above 85% in the highest deployment density with up to 500 readers, also providing a 90% data rate. In [25], we propose two distributed and efficient solutions for dense mobile deployments of RFID systems. mDEFAR is an adaptation of a previous work highly performing in terms of collisions reduction, efficiency and fairness in dense static deployments. CORA is more of a locally mutual solution where each reader relies on its neighborhood to enable itself or not. Using a beaconing mechanism, each reader is able to identify potential (non-)colliding neighbors in a running frame and as such chooses to read or not. Performance evaluation shows high performance in terms

of coverage delay for both proposals quickly achieving 100% coverage depending on the considered use case while always maintaining consistent efficiency levels above 70%. Compared to GDRA, our solutions proved to be better suited for highly dense and mobile environments, offering both higher throughput and efficiency. The results reveal that depending on the application considered, choosing either mDEFAR or CORA helps improve efficiency and coverage delay.

GANG Project-Team

6. New Results

6.1. Graph and Combinatorial Algorithms

6.1.1. Induced Matching algorithms

In [21] we study the maximum induced matching problem on a graph G. Induced matchings correspond to independent sets in $L^2(G)$, the square of the line graph of G. The problem is NP-complete on bipartite graphs. In this work, we show that for a number of graph families with forbidden vertex orderings, almost all forbidden patterns on three vertices are preserved when taking the square of the line graph. That is, given a graph class $\mathcal{G}$ characterized by a vertex ordering, and a graph $G = (V, E) \in \mathcal{G}$ with a corresponding vertex ordering σ of V, one can produce (in linear time in the size of G) an ordering on the vertices of $L^2(G)$, that shows that $L^2(G) \in \mathcal{G}$. This result gives alternate closure proofs for the $L^2(\bullet)$ closure operation. Furthermore, these orderings on $L^2(G)$ can be exploited algorithmically to compute a maximum induced matching for graphs belonging to $\mathcal{G}$ faster. We illustrate this latter fact in the second half of the paper where we focus on cocomparability graphs, a large graph class that includes interval, permutation, and trapezoid graphs, and we present the first $O(mn)$ time algorithm to compute a maximum weighted induced matching on G; an improvement from the best known $O(n^4)$ time algorithm for the unweighted case.

6.1.2. The LexBFS cycle on cocomparability graphs

Since its introduction to recognize chordal graphs by Rose, Tarjan, and Lueker, Lexicographic Breadth First Search (LexBFS) has been used to come up with simple, often linear time, algorithms on various classes of graphs. These algorithms, called multi-sweep algorithms, compute a number of LexBFS orderings $\sigma_1, ..., \sigma_k$, where σ_i is used to break ties for σ_{i+1}, we write $LexBFS^+(\sigma_i) = \sigma_{i+1}$. For instance, Corneil et al. gave a linear time multi-sweep algorithm to recognize interval graphs [SODA 1998], Kratsch et al. gave a certifying recognition algorithm for interval and permutation graphs [SODA 2003]. Since the number of LexBFS orderings for a graph is finite, after some fixed number of $^+$ sweeps, we will eventually loop in a sequence of $\sigma_1, ..., \sigma_k$ vertex orderings such that $\sigma_{i+1} = LexBFS^+(\sigma_i)$ modulo k.

In [13] we introduce and study this new graph invariant, *LexCycle*(G), defined as the maximum length of a cycle of vertex orderings obtained via a sequence of LexBFS$^+$. In this work, we focus on graph classes with small LexCycle. We give evidence that a small LexCycle often leads to linear structure that has been exploited algorithmically on a number of graph classes. In particular, we show that for proper interval, interval, co-bipartite, domino-free cocomparability graphs, as well as trees, there exists two orderings σ and τ such that $\sigma = LexBFS^+(\tau)$ and $\tau = LexBFS^+(\sigma)$. One of the consequences of these results is the simplest algorithm to compute a transitive orientation for these graph classes.

It was conjectured by Stacho [2015] that LexCycle is at most the asteroidal number of the graph class, we disprove this conjecture by giving a construction for which the LexCycle(G) grows polynomially in the asteroidal number of G.

6.1.3. Approximation Strategies for Generalized Binary Search in Weighted Trees

In [15], we have considered the following generalization of the binary search problem. A search strategy is required to locate an unknown target node t in a given tree T. Upon querying a node v of the tree, the strategy receives as a reply an indication of the connected component of $T \setminus \{v\}$ containing the target t. The cost of querying each node is given by a known non-negative weight function, and the considered objective is to minimize the total query cost for a worst-case choice of the target.

Designing an optimal strategy for a weighted tree search instance is known to be strongly NP-hard, in contrast to the unweighted variant of the problem which can be solved optimally in linear time. Here, we show that weighted tree search admits a quasi-polynomial time approximation scheme: for any $0 < \varepsilon < 1$, there exists a $(1+\varepsilon)$-approximation strategy with a computation time of $n^{O(\log n/\varepsilon^2)}$. Thus, the problem is not APX-hard, unless $NP \subseteq DTIME(n^{O(\log n)})$. By applying a generic reduction, we obtain as a corollary that the studied problem admits a polynomial-time $O(\sqrt{\log n})$-approximation. This improves previous $\widehat{O}(\log n)$-approximation approaches, where the $\widehat{O}$-notation disregards $O(\mathrm{poly}\log\log n)$-factors.

6.1.4. The Dependent Doors Problem: An Investigation into Sequential Decisions without Feedback

In [24] we introduce the *dependent doors problem* as an abstraction for situations in which one must perform a sequence of possibly dependent decisions, without receiving feedback information on the effectiveness of previously made actions. Informally, the problem considers a set of d doors that are initially closed, and the aim is to open all of them as fast as possible. To open a door, the algorithm knocks on it and it might open or not according to some probability distribution. This distribution may depend on which other doors are currently open, as well as on which other doors were open during each of the previous knocks on that door. The algorithm aims to minimize the expected time until all doors open. Crucially, it must act at any time without knowing whether or which other doors have already opened. In this work, we focus on scenarios where dependencies between doors are both positively correlated and acyclic.

The fundamental distribution of a door describes the probability it opens in the best of conditions (with respect to other doors being open or closed). We show that if in two configurations of d doors corresponding doors share the same fundamental distribution, then these configurations have the same optimal running time up to a universal constant, no matter what are the dependencies between doors and what are the distributions. We also identify algorithms that are optimal up to a universal constant factor. For the case in which all doors share the same fundamental distribution we additionally provide a simpler algorithm, and a formula to calculate its running time. We furthermore analyse the price of lacking feedback for several configurations governed by standard fundamental distributions. In particular, we show that the price is logarithmic in d for memoryless doors, but can potentially grow to be linear in d for other distributions.

We then turn our attention to investigate precise bounds. Even for the case of two doors, identifying the optimal sequence is an intriguing combinatorial question. Here, we study the case of two cascading memoryless doors. That is, the first door opens on each knock independently with probability p_1. The second door can only open if the first door is open, in which case it will open on each knock independently with probability p_2. We solve this problem almost completely by identifying algorithms that are optimal up to an additive term of 1.

6.2. Distributed Computing

6.2.1. Robust Detection in Leak-Prone Population Protocols

In [10], we aim to design population protocols for the problem of detecting a signal in the presence of faults, motivated by scenarios of chemical computation. In contrast to electronic computation, chemical computation is noisy and susceptible to a variety of sources of error, which has prevented the construction of robust complex systems. To be effective, chemical algorithms must be designed with an appropriate error model in mind. Here we consider the model of chemical reaction networks that preserve molecular count (population protocols), and ask whether computation can be made robust to a natural model of unintended "leak" reactions. Our definition of leak is motivated by both the particular spurious behavior seen when implementing chemical reaction networks with DNA strand displacement cascades, as well as the unavoidable side reactions in any implementation due to the basic laws of chemistry. We develop a new "Robust Detection" algorithm for the problem of fast (logarithmic time) single molecule detection, and prove that it is robust to this general model of leaks. Besides potential applications in single molecule detection, the error-correction ideas developed here might enable a new class of robust-by-design chemical algorithms. Our analysis is based on a non-standard hybrid argument, combining ideas from discrete analysis of population protocols with classic Markov chain techniques.

6.2.2. Minimizing Message Size in Stochastic Communication Patterns: Fast Self-Stabilizing Protocols with 3 bits

In [12] we consider the basic PULL model of communication, in which in each round, each agent extracts information from few randomly chosen agents. We seek to identify the smallest amount of information revealed in each interaction (message size) that nevertheless allows for efficient and robust computations of fundamental information dissemination tasks. We focus on the *Majority Bit Dissemination* problem that considers a population of n agents, with a designated subset of *source agents*. Each source agent holds an *input bit* and each agent holds an *output bit*. The goal is to let all agents converge their output bits on the most frequent input bit of the sources (the *majority bit*). Note that the particular case of a single source agent corresponds to the classical problem of *Broadcast* (also termed *Rumor Spreading*). We concentrate on the severe fault-tolerant context of *self-stabilization*, in which a correct configuration must be reached eventually, despite all agents starting the execution with arbitrary initial states. In particular, the specification of who is a source and what is its initial input bit may be set by an adversary.

We first design a general compiler which can essentially transform any self-stabilizing algorithm with a certain property that uses ℓ-bits messages to one that uses only $\log \ell$-bits messages, while paying only a small penalty in the running time. By applying this compiler recursively we then obtain a self-stabilizing *Clock Synchronization* protocol, in which agents synchronize their clocks modulo some given integer T, within $\widetilde{O}(\log n \log T)$ rounds w.h.p., and using messages that contain 3 bits only.

We then employ the new Clock Synchronization tool to obtain a self-stabilizing Majority Bit Dissemination protocol which converges in $\widetilde{O}(\log n)$ time, w.h.p., on every initial configuration, provided that the ratio of sources supporting the minority opinion is bounded away from half. Moreover, this protocol also uses only 3 bits per interaction.

6.2.3. The ANTS Problem

In [6] we introduce the *Ants Nearby Treasure Search (ANTS)* problem, which models natural cooperative foraging behavior such as that performed by ants around their nest. In this problem, k probabilistic agents, initially placed at a central location, collectively search for a treasure on the two-dimensional grid. The treasure is placed at a target location by an adversary and the agents' goal is to find it as fast as possible as a function of both k and D, where D is the (unknown) distance between the central location and the target. We concentrate on the case in which agents cannot communicate while searching. It is straightforward to see that the time until at least one agent finds the target is at least $\Omega(D + D^2/k)$, even for very sophisticated agents, with unrestricted memory. Our algorithmic analysis aims at establishing connections between the time complexity and the initial knowledge held by agents (*e.g.*, regarding their total number k), as they commence the search. We provide a range of both upper and lower bounds for the initial knowledge required for obtaining fast running time. For example, we prove that $\log \log k + \Theta(1)$ bits of initial information are both necessary and sufficient to obtain asymptotically optimal running time, *i.e.,*$O(D + D^2/k)$. We also we prove that for every $0 < \epsilon < 1$, running in time $O(\log^{1-\epsilon} k \cdot (D + D^2/k))$ requires that agents have the capacity for storing $\Omega(\log^{\epsilon} k)$ different states as they leave the nest to start the search. To the best of our knowledge, the lower bounds presented in this paper provide the first non-trivial lower bounds on the memory complexity of probabilistic agents in the context of search problems.

We view this paper as a "proof of concept" for a new type of interdisciplinary methodology. To fully demonstrate this methodology, the theoretical tradeoff presented here (or a similar one) should be combined with measurements of the time performance of searching ants.

6.2.4. Breathe before Speaking: Efficient Information Dissemination despite Noisy, Limited and Anonymous Communication

Distributed computing models typically assume reliable communication between processors. While such assumptions often hold for engineered networks, e.g., due to underlying error correction protocols, their relevance to biological systems, wherein messages are often distorted before reaching their destination, is quite limited. In this study we take a first step towards reducing this gap by rigorously analyzing a model of

communication in large anonymous populations composed of simple agents which interact through short and highly unreliable messages.

In [9] we focus on the broadcast problem and the majority-consensus problem. Both are fundamental information dissemination problems in distributed computing, in which the goal of agents is to converge to some prescribed desired opinion. We initiate the study of these problems in the presence of communication noise. Our model for communication is extremely weak and follows the push gossip communication paradigm: In each round each agent that wishes to send information delivers a message to a random anonymous agent. This communication is further restricted to contain only one bit (essentially representing an opinion). Lastly, the system is assumed to be so noisy that the bit in each message sent is flipped independently with probability $1/2 - \epsilon$, for some small $\epsilon > 0$.

Even in this severely restricted, stochastic and noisy setting we give natural protocols that solve the noisy broadcast and the noisy majority-consensus problems efficiently. Our protocols run in $O(\log n/\epsilon^2)$ rounds and use $O(n \log n/\epsilon^2)$ messages/bits in total, where n is the number of agents. These bounds are asymptotically optimal and, in fact, are as fast and message efficient as if each agent would have been simultaneously informed directly by an agent that knows the prescribed desired opinion. Our efficient, robust, and simple algorithms suggest balancing between silence and transmission, synchronization, and majority-based decisions as important ingredients towards understanding collective communication schemes in anonymous and noisy populations.

6.2.5. Parallel Search with no Coordination

In [23] we consider a parallel version of a classical Bayesian search problem. k agents are looking for a treasure that is placed in one of finately many boxes according to a known distribution p. The aim is to minimize the expected time until the first agent finds it. Searchers run in parallel where at each time step each searcher can "peek" into a box. A basic family of algorithms which are inherently robust is *non-coordinating* algorithms. Such algorithms act independently at each searcher, differing only by their probabilistic choices. We are interested in the price incurred by employing such algorithms when compared with the case of full coordination.

We first show that there exists a non-coordination algorithm, that knowing only the relative likelihood of boxes according to p, has expected running time of at most $10 + 4(1 + \frac{1}{k})^2 T$, where T is the expected running time of the best fully coordinated algorithm. This result is obtained by applying a refined version of the main algorithm suggested by Fraigniaud, Korman and Rodeh in STOC'16, which was designed for the context of linear parallel search.

We then describe an optimal non-coordinating algorithm for the case where the distribution p is known. The running time of this algorithm is difficult to analyse in general, but we calculate it for several examples. In the case where p is uniform over a finite set of boxes, then the algorithm just checks boxes uniformly at random among all non-checked boxes and is essentially 2 times worse than the coordinating algorithm. We also show simple algorithms for Pareto distributions over M boxes. That is, in the case where $p(x) \sim 1/x^b$ for $0 < b < 1$, we suggest the following algorithm: at step t choose uniformly from the boxes unchecked in $\{1, ..., \min(M, \lfloor t/\sigma \rfloor)\}$, where $\sigma = b/(b + k - 1)$. It turns out this algorithm is asymptotically optimal, and runs about $2 + b$ times worse than the case of full coordination.

6.2.6. Wait-free local algorithms

When considering distributed computing, reliable message-passing synchronous systems on the one side, and asynchronous failure-prone shared-memory systems on the other side, remain two quite independently studied ends of the reliability/asynchrony spectrum. The concept of locality of a computation is central to the first one, while the concept of wait-freedom is central to the second one. In [2] we propose a new DECOUPLED model in an attempt to reconcile these two worlds. It consists of a synchronous and reliable communication graph of n nodes, and on top a set of asynchronous crash-prone processes, each attached to a communication node. To illustrate the DECOUPLED model, the paper presents an asynchronous 3-coloring algorithm for the processes of a ring. From the processes point of view, the algorithm is wait-free. From a locality point of view, each

process uses information only from processes at distance $O(\log * n)$ from it. This local wait-free algorithm is based on an extension of the classical Cole and Vishkin's vertex coloring algorithm in which the processes are not required to start simultaneously.

6.2.7. Immediate t-resilient Snapshot

An immediate snapshot object is a high level communication object, built on top of a read/write distributed system in which all except one processes may crash. It allows each process to write a value and obtains a set of pairs (process id, value) such that, despite process crashes and asynchrony, the sets obtained by the processes satisfy noteworthy inclusion properties. Considering an n-process model in which up to t processes are allowed to crash, [14] is on the construction of t-resilient immediate snapshot objects.

6.2.8. Decidability classes for mobile agents computing

In [7], we establish a classification of decision problems that are to be solved by mobile agents operating in unlabeled graphs, using a deterministic protocol. The classification is with respect to the ability of a team of agents to solve decision problems, possibly with the aid of additional information. In particular, our focus is on studying differences between the decidability of a decision problem by agents and its verifiability when a certificate for a positive answer is provided to the agents (the latter is to the former what NP is to P in the framework of sequential computing). We show that the class MAV of mobile agents verifiable problems is much wider than the class MAD of mobile agents decidable problems. Our main result shows that there exist natural MAV-complete problems: the most difficult problems in this class, to which all problems in MAV are reducible via a natural mobile computing reduction. Beyond the class MAV we show that, for a single agent, three natural oracles yield a strictly increasing chain of relative decidability classes.

6.2.9. Distributed Detection of Cycles

Distributed property testing in networks has been introduced by Brakerski and Patt-Shamir (2011), with the objective of detecting the presence of large dense sub-networks in a distributed manner. Recently, Censor-Hillel et al. (2016) have shown how to detect 3-cycles in a constant number of rounds by a distributed algorithm. In a follow up work, Fraigniaud et al. (2016) have shown how to detect 4-cycles in a constant number of rounds as well. However, the techniques in these latter works were shown not to generalize to larger cycles C_k with $k \geq 5$. In [19], we completely settle the problem of cycle detection, by establishing the following result. For every $k \geq 3$, there exists a distributed property testing algorithm for C_k-freeness, performing in a constant number of rounds. All these results hold in the classical CONGEST model for distributed network computing. Our algorithm is 1-sided error. Its round-complexity is $O(1/\epsilon)$ where $\epsilon \in (0, 1)$ is the property testing parameter measuring the gap between legal and illegal instances.

6.2.10. What Can Be Verified Locally?

In [18], we are considering *distributed network computing*, in which computing entities are connected by a network modeled as a connected graph. These entities are located at the nodes of the graph, and they exchange information by message-passing along its edges. In this context, we are adopting the classical framework for *local distributed decision*, in which nodes must collectively decide whether their network configuration satisfies some given boolean predicate, by having each node interacting with the nodes in its vicinity only. A network configuration is accepted if and only if every node individually accepts. It is folklore that not every Turing-decidable network property (e.g., whether the network is planar) can be decided locally whenever the computing entities are Turing machines (TM). On the other hand, it is known that every Turing-decidable network property can be decided locally if nodes are running *non-deterministic* Turing machines (NTM). However, this holds only if the nodes have the ability to guess the identities of the nodes currently in the network. That is, for different sets of identities assigned to the nodes, the correct guesses of the nodes might be different. If one asks the nodes to use the same guess in the same network configuration even with different identity assignments, i.e., to perform *identity-oblivious* guesses, then it is known that not every Turing-decidable network property can be decided locally.

We show that every Turing-decidable network property can be decided locally if nodes are running *alternating* Turing machines (ATM), and this holds even if nodes are bounded to perform identity-oblivious guesses. More specifically, we show that, for every network property, there is a local algorithm for ATMs, with at most 2 alternations, that decides that property. To this aim, we define a hierarchy of classes of decision tasks where the lowest level contains tasks solvable with TMs, the first level those solvable with NTMs, and level k contains those tasks solvable with ATMs with k alternations. We characterize the entire hierarchy, and show that it collapses in the second level. In addition, we show separation results between the classes of network properties that are locally decidable with TMs, NTMs, and ATMs, and we establish the existence of completeness results for each of these classes, using novel notions of *local reduction*.

6.2.11. Certification of Compact Low-Stretch Routing Schemes

On the one hand, the correctness of routing protocols in networks is an issue of utmost importance for guaranteeing the delivery of messages from any source to any target. On the other hand, a large collection of *routing schemes* have been proposed during the last two decades, with the objective of transmitting messages along short routes, while keeping the routing tables small. Regrettably, all these schemes share the property that an adversary may modify the content of the routing tables with the objective of, e.g., blocking the delivery of messages between some pairs of nodes, without being detected by any node.

In [17], we present a simple *certification* mechanism which enables the nodes to locally detect any alteration of their routing tables. In particular, we show how to locally verify the stretch-3 routing scheme by Thorup and Zwick [SPAA 2001] by adding certificates of $\widetilde{O}(\sqrt{n})$ bits at each node in n-node networks, that is, by keeping the memory size of the same order of magnitude as the original routing tables. We also propose a new *name-independent* routing scheme using routing tables of size $\widetilde{O}(\sqrt{n})$ bits. This new routing scheme can be locally verified using certificates on $\widetilde{O}(\sqrt{n})$ bits. Its stretch is 3 if using handshaking, and 5 otherwise.

6.2.12. Error-Sensitive Proof-Labeling Schemes

Proof-labeling schemes are known mechanisms providing nodes of networks with *certificates* that can be *verified* locally by distributed algorithms. Given a boolean predicate on network states, such schemes enable to check whether the predicate is satisfied by the actual state of the network, by having nodes interacting with their neighbors only. Proof-labeling schemes are typically designed for enforcing fault-tolerance, by making sure that if the current state of the network is illegal with respect to some given predicate, then at least one node will detect it. Such a node can raise an alarm, or launch a recovery procedure enabling the system to return to a legal state. We introduce *error-sensitive* proof-labeling schemes. These are proof-labeling schemes which guarantee that the number of nodes detecting illegal states is linearly proportional to the edit-distance between the current state and the set of legal states. By using error-sensitive proof-labeling schemes, states which are far from satisfying the predicate will be detected by many nodes, enabling fast return to legality. In [20], we provide a structural characterization of the set of boolean predicates on network states for which there exist error-sensitive proof-labeling schemes. This characterization allows us to show that classical predicates such as, e.g., acyclicity, and leader admit error-sensitive proof-labeling schemes, while others like regular subgraphs don't. We also focus on *compact* error-sensitive proof-labeling schemes. In particular, we show that the known proof-labeling schemes for spanning tree and MST, using certificates on $O(\log n)$ bits, and on $O(\log^2 n)$ bits, respectively, are error-sensitive, as long as the trees are locally represented by adjacency lists, and not by a pointer to the parent.

6.2.13. Distributed Property Testing

In [16], we designed distributed testing algorithms of graph properties in the CONGEST model [Censor-Hillel et al. 2016], especially for testing subgraph-freeness. Testing a given property means that we have to distinguish between graphs having the property, and graphs that are ϵ-far from having it, meaning that one must remove an ϵ-fraction of the edges to obtain it. We established a series of results, among which:

- Testing H-freeness in a constant number of rounds, for any graph H that can be transformed into a tree by removing a single edge. This includes, e.g., cycle-freeness for any constant cycle, and K_4-freeness. As a byproduct, we give a deterministic CONGEST protocol determining whether a graph contains a fixed tree as a subgraph.

- For cliques K_k with $k \geq 5$, we show that K_k-freeness can be tested in $O(\left(\frac{m}{\epsilon}\right)^{\frac{1}{2}+\frac{1}{k-2}})$ rounds, where m is the number of edges in the network graph.
- We describe a general procedure for converting ϵ-testers with $f(D)$ rounds, where D denotes the diameter of the graph, to work in $O((\log n)/\epsilon) + f((\log n)/\epsilon)$ rounds, where n is the number of processors of the network. We then apply this procedure to obtain an ϵ-tester for testing whether a graph is bipartite.

These protocols extend and improve previous results of [Censor-Hillel et al. 2016] and [Fraigniaud et al. 2016].

6.3. Models and Algorithms for Networks

6.3.1. Analysis of Multiple Random Walks on Paths and Grids

In [22], we derive several new results on multiple random walks on "low-dimensional" graphs. First, inspired by an example of a weighted random walk on a path of three vertices given by Efremenko and Reingold, we prove the following dichotomy: as the path length n tends to infinity, we have a super-linear speed-up w.r.t. the cover time if and only if the number of walks k is equal to 2. An important ingredient of our proofs is the use of a continuous-time analogue of multiple random walks, which might be of independent interest. Finally, we also present the first tight bounds on the speed-up of the cover time for any d-dimensional grid with $d \geq 2$ being an arbitrary constant, and reveal a sharp transition between linear and logarithmic speed-up.

6.3.2. Decomposing a Graph into Shortest Paths with Bounded Eccentricity

In [11], we introduce the problem of hub-laminar decomposition which generalizes that of computing a shortest path with minimum eccentricity (MESP). Intuitively, it consists in decomposing a graph into several paths that collectively have small eccentricity and meet only near their extremities. The problem is related to computing an isometric cycle with minimum eccentricity (MEIC). It is also linked to DNA reconstitution in the context of metagenomics in biology. We show that a graph having such a decomposition with long enough paths can be decomposed in polynomial time with approximated guaranties on the parameters of the decomposition. Moreover, such a decomposition with few paths allows to compute a compact representation of distances with additive distortion. We also show that having an isometric cycle with small eccentricity is related to the possibility of embedding the graph in a cycle with low distortion.

6.3.3. Individual versus collective cognition in social insects

The concerted responses of eusocial insects to environmental stimuli are often referred to as collective cognition at the level of the colony. To achieve collective cognition, a group can draw on two different sources: individual cognition and the connectivity between individuals. Computation in neural networks, for example, is attributed more to sophisticated communication schemes than to the complexity of individual neurons. The case of social insects, however, can be expected to differ. This is because individual insects are cognitively capable units that are often able to process information that is directly relevant at the level of the colony. Furthermore, involved communication patterns seem difficult to implement in a group of insects as they lack a clear network structure. In [5] we discusses links between the cognition of an individual insect and that of the colony. We provide examples for collective cognition whose sources span the full spectrum between amplification of individual insect cognition and emergent group-level processes.

INFINE Project-Team

6. New Results

6.1. Online Social Networks (OSN)

6.1.1. Capacity of Information Processing Systems

- Participants: Laurent Massoulié and Kuang Xu

We propose and analyze a family of information processing systems, where a finite set of experts or servers are employed to extract information about a stream of incoming jobs. Each job is associated with a hidden label drawn from some prior distribution. An inspection by an expert produces a noisy outcome that depends both on the job's hidden label and the type of the expert, and occupies the expert for a finite time duration. A decision maker's task is to dynamically assign inspections so that the resulting outcomes can be used to accurately recover the labels of all jobs, while keeping the system stable. Among our chief motivations are applications in crowd-sourcing, diagnostics, and experiment designs, where one wishes to efficiently learn the nature of a large number of items, using a finite pool of computational resources or human agents. We focus on the capacity of such an information processing system. Given a level of accuracy guarantee, we ask how many experts are needed in order to stabilize the system, and through what inspection architecture. Our main result provides an adaptive inspection policy that is asymptotically optimal in the following sense: the ratio between the required number of experts under our policy and the theoretical optimal converges to one, as the probability of error in label recovery tends to zero. This work was firstly accepted and presented at the COLT conference.

6.2. Spontaneous Wireless Networks (SWN)

6.2.1. Spatio-Temporal Prediction of Cellular Data Traffic

- Participants: Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, Carlos Sarraute

The understanding of human behaviors is a central question in multi-disciplinary research and has contributed to a wide range of applications. The ability to foresee human activities has essential implications in many aspects of cellular networks. In particular, the high availability of mobility prediction can enable various application scenarios such as location-based recommendation, home automation, and location-related data dissemination; the better understanding of future mobile data traffic demand can help to improve the design of solutions for network load balancing, aiming at improving the quality of Internet-based mobile services. Although a large and growing body of literature has investigated the topic of predicting human mobility, there has been little discussion in anticipating mobile data traffic in cellular networks, especially in spatiotemporal view of individuals. We address the problem of understanding spatiotemporal mobile data traffic demand for individuals and perform an theoretical and empirical analysis of jointly predicting human whereabouts and mobile data traffic, by collaboratively mining human mobility dataset and mobile data traffic dataset. Our contributions are summarized as follows:

- We investigate the limits of predictability by measuring the maximum predictability that any algorithm has potential to achieve based on tools of information theory. Our theoretical analysis shows that it is theoretically possible to anticipate the individual demand with a typical accuracy of 75% despite the heterogeneity of users and with an improved accuracy of 80% using joint prediction with mobility information. This work was published at the IEEE LCN 2017 international conference and the Technical report RT-0483 brings a full description of the work, which is being prepared for a journal submission.

- We evaluate the state-of-the-art predictors and propose novel solutions for predicting mobile data traffic via machine learning algorithms. Our data-driven test on the performance of these predictors show that the 2nd order Markov predictor outperforms all the legacy time series predictors. It can achieve a mean accuracy of 62% but can hardly have an enhancement from knowing human mobility information. Besides, based on machine learning techniques, our proposed solutions can achieve a typical accuracy of 70% and have a 1% 5% degree of improvement by learning individual whereabouts (what confirms the predictability theoretical results). Finally, our analysis show that knowing mobile data traffic of a user can significantly help the prediction of his whereabouts for 50% of the users, leading to an improvement up to 10% regarding accuracy. The Technical Report hal-01675573 brings more details on this work. A conference paper is also in preparation.

All those works were performed in the context of the Guangshuo Chen's PhD thesis, who will defend in March 2018.

6.2.2. Human Mobility completion of Sparse Call Detail Records for Mobility Analysis

- Participants: Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, Sahar Hoteit

Call Detail Records (CDR) are an important source of information in the study of diverse aspects of human mobility. The accuracy of mobility information granted by CDR strongly depends on the radio access infrastructure deployment and the frequency of interactions between mobile users and the network. As cellular network deployment is highly irregular and interaction frequencies are typically low, CDR are often characterized by spatial and temporal sparsity, which, in turn, can bias mobility analyses based on such data. In this paper, we precisely address this subject. First, we evaluate the spatial error in CDR, caused by approximating user positions with cell tower locations. Second, we assess the impact of the limited spatial and temporal granularity of CDR on the estimation of standard mobility metrics. Third, we propose novel and effective techniques to reduce temporal sparsity in CDR, by leveraging regularity in human movement patterns.

These works have been published as invited papers at the ACM CHANTS 2016 workshop (in conjunction with ACM MobiCom 2016) and at the IEEE DAWM workshop (in conjunction with IEEE Percom 2017). A journal version (also registered as TR: hal-01646608) is in revision at the Computer Communication Elsevier Journal, and got the first notification asking for minor revisions. Finally, a new completion methodology improving the previously described that leverages tensor factorization was designed and will be submitted to a journal: the technical report hal-01675570 describes this work.

6.2.3. Sampling frequency of human mobility

- Participants: Panagiota Katsikouli, Aline Carneiro Viana, Marco Fiore, Alberto Tarable

Recent studies have leveraged tracking techniques based on positioning technologies to discover new knowledge about human mobility. These investigations have revealed, among others, a high spatiotemporal regularity of individual movement patterns. Building on these findings, we aim at answering the question *"at what frequency should one sample individual human movements so that they can be reconstructed from the collected samples with minimum loss of information?"*. Our quest for a response leads to the discovery of *(i)* seemingly universal spectral properties of human mobility, and *(ii)* a linear scaling law of the localization error with respect to the sampling interval. Our findings are based on the analysis of fine-grained GPS trajectories of 119 users worldwide. The applications of our findings are related to a number of fields relevant to ubiquitous computing, such as energy-efficient mobile computing, location-based service operations, active probing of subscribers' positions in mobile networks and trajectory data compression. to an international conference in the next months. This work was published at the IEEE Globecom 2017 international conference.

We are improving the currently published sampling approache by incorporating human behavioral features at the sampling decisions to make it more adaptive. This is an on-going work with Panagiota Katsikouli, who spent 5 months in our team working as an internship and is currently doing a Post-Doc at the AGORA Inria team.

6.2.4. Inference of human personality from mobile phones datasets

- Participants: Adriano di Luzio, Aline Carneiro Viana, Julinda Stefa, Katia Jaffres

Personality research has enjoyed a strong resurgence over the past decade. Trait-based personality theories define personality as the traits that predict a person's behavior through learning and habits. Personality traits are relatively stable over time, differ across individuals, and most importantly, influence behavior. In psychology, the human personality has been modeled into a set of independent factors that, together, accurately describe any individual: The Five Factors Personality Model. This personality model presents the Big Five personality traits, often represented by the OCEAN acronym: Openness: appreciation for a variety of experiences; Conscientiousness: planning ahead rather than being spontaneous; Extraversion: being sociable, energetic and talkative; Agreeableness: being kind, sympathetic and happy to help; Neuroticism: inclined to worry or be vulnerable or temperamental.
This is a very recently started work, where we are firstly analysing the relationship between smartphone usages (i.e., social interactions, content interest, mobility, and communication) and personality traits in the Big Five Model. Most of the studies on personality traits were performed by social scientists and in particular, by psychologists. Studies reveal that one of the most distal influences shaping personality lie in the environment where development occurs. Nevertheless, the identification of precise environmental sources impacting personality is still an open research. More recently, computer science researchers have tried to extract personality from datasets collected through smartphones. Although laying the ground work to understand human personality from smartphones usage, much still remain to be investigated. Thus, we are performing analysis to study the correlation between traits and technological features. We plan then to establish a methodology to infer traints from features and consequently, to investigate how different traits influence different features.

This is an on-going work with Adriano di Luzio, who spent 4 months in our team working as an internship, and Julinda Stefa, an invited research visitor at Infine.

6.2.5. Predicting new places to visit in human mobility decision

- Participants: Maria Astefanoaei, Aline Carneiro Viana, Rik Sarkar

Most location prediction methods need a large user mobility history to accurately predict the next location of a user (markov chains, rnn). These methods are particularly good for predicting locations that are frequently visited by users, but not as good for predicting new places or how a user?s trajectories change in case of random events. We amend this by using contextual information to manage new places and random events and the movement patterns of users who exhibit similar behaviours. In this context, we plan to use the user?s profile and social ties to identify the most probable next category of locations (type of actions: entertainment, social, food etc.). Then, use subtrajectory similarity to predict the route taken to the identified area. This is an on-going work with the intern Maria Astefanoaei and her advisor, who spent 5 months in our team.

6.2.6. Data offloading decision via mobile crowdsensing

- Participants: Emanuel Lima, Aline Carneiro Viana, Ana Aguiar

With the steady growth of smart-phones sales [1], the demand for services that generate mobile data traffic has grown tremendously. WiFi offloading has been considered as a promising solution to the recent boost up of mobile data consumption that is making excessive demands on cellular networks in metropolitan areas. The idea consists in shifting the traffic off of cellular networks to WiFi networks. Characterizing the capacity and availability of a chaotic deployed dense WiFi network is crucial to understand and decide where and when to o oad data. This is the first goal of this work, where the MACACO dataset was considered in the characterization. Our final goal is the design of a decision strategy allowing a mobile phone of a user to decide if offload or not her traffic, i.e., when, where (using what Access Point in her usual mobility) and how (if the traffic will be offloadied to one or more Access Points). This is an on-going work with the intern emanuel Lima and his advisor, who spent 4 months in our team.

6.2.7. Infering friends in the crowd in Device-to-Device communication

- Participants: Rafael Costa, Aline Carneiro Viana, Leobino Sampaio, Artur Ziviani

The next generation of mobile phone networks (5G) will have to deal with spectrum bottleneck and other major challenges to serve more users with high-demanding requirements. Among those are higher scalability and data rates, lower latencies and energy consumption plus reliable ubiquitous connectivity. Thus, there is a need for a better spectrum reuse and data offloading in cellular networks while meeting user expectations. According to literature, one of the 10 key enabling technologies for 5G is device-to-device (D2D) communications, an approach based on direct user involvement. Nowadays, mobile devices are attached to human daily life activities, and therefore communication architectures using context and human behavior information are promising for the future. User-centric communication arose as an alternative to increase capillarity and to offload data traffic in cellular networks through opportunistic connections among users. Although having the user as main concern, solutions in the user-centric communication/networking area still do not see the user as an individual, but as a network active element. Hence, these solutions tend to only consider user features that can be measured from the network point of view, ignoring the ones that are intrinsic from human activity (e.g., daily routines, personality traits, etc). In this work, we plan to investigate how human-aspects and behavior can be useful to leverage future device-to-device communication. This is a recently started PhD thesis subject, aiming the design of a methodology to select next-hops in a D2D communication that will be human-aware: i.e., that will consider not only available physical resoures at the mobile device of a wireless neighbor, her mobility features and restrictions but also any information allowing to infer how much sharing willing she is.

6.3. Internet of Things (IoT) and Information Centric Networking (ICN)

6.3.1. Low-power Internet of Things with NDN and Cooperative Caching

- Participants: Oliver Hahm, Emmanuel Baccelli, Thomas C. Schmidt, Matthias Wählisch, Cédric Adjih, and Laurent Massoulié

Energy efficiency is a major driving factor in the Internet of Things (IoT). In this context, an IoT approach based on Information-Centric Networking (ICN) offers prospects for low energy consumption. Indeed, ICN can provide local in-network content caching so that relevant IoT content remains available at any time while devices are in deep-sleep mode most of the time. In our paper on the subject, we evaluated NDN enhanced with CoCa, a simple side protocol we designed to exploit content names together with smart interplay between cooperative caching and power-save sleep capabilities on IoT devices. We performed extensive, large scale experiments on real hardware with IoT networks comprising of up to 240 nodes, and on an emulator with up to 1000 nodes. We have shown in practice that, with NDN+CoCa, devices can reduce energy consumption by an order of magnitude while maintaining recent IoT content availability above 90 %. We furthermore provided auto-configuration mechanisms enabling practical ICN deployments on IoT networks of arbitrary size with NDN+CoCa. With such mechanisms, each device could autonomously configure names and auto-tune parameters to reduce energy consumption as demonstrated in our paper.

6.3.2. Information Centric Networking for the IoT Robotics

- Participants: Loic Dauphin, Emmanuel Baccelli, Cédric Adjih

In the near-future, humans will interact with swarms of low-cost, interconnected robots. Such robots will hence integrate the Internet of Things, and coin the term IoT robotics. Using ROS (Robot Operating System) is currently the dominant approach to implement distributed robotic software modules communicating with one another. ROS nodes can publish or subscribe to topics, which are named and typed data streams sent over the network. In our work on the subject, we presented preliminary work exploring the potential of using NDN as network primitive for ROS2 nodes (the newest version of ROS).

6.3.3. Data Synchronization through Information Centric Networking

- Participants: Ayat Zaki Hindi, Cédric Adjih, Michel Kieffer, Claudio Weidmann

The use of Named Data Networking (NDN) for distributed multiuser applications, e.g. group messaging and file sharing, requires NDN synchronization protocols to maintain the same shared dataset (and its updates) among all nodes. ChronoSync, RoundSync, and PartialSync are some proposals to address this issue.

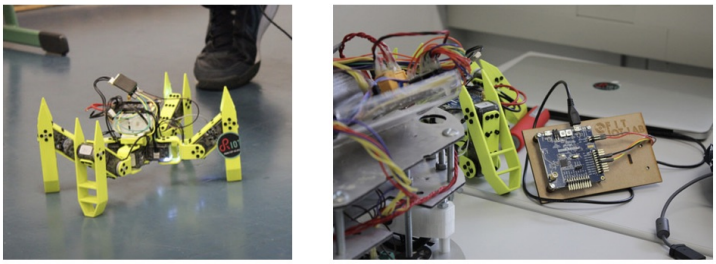

Figure 1. Demonstration of the use of NDN in IoT Robotics

In our work on the subject, we focused on the state-of-the-art protocol RoundSync: we study its core features, that permit participating nodes to detect, propagate, and reconcile all changes. Particular attention is given to the case of multiple changes per round. We then proposed an improved variant, iRoundSync, that exchanges fewer messages in the multiple-change case and is more resilient to packet losses. We have quantified the performance gain of iRoundSync on a simple topology.

6.4. Internet of Things (IoT) and 5G

6.4.1. Efficient Random Access for 5G Systems: Coded Slotted Aloha

- Participants: Ehsan Ebrahimi Khaleghi, Cédric Adjih, Amira Alloum, Paul Mï¿œhlethaler, Vinod Kumar

Motivated by scenario requirements for 5G cellular networks, we have studied one of the candidate protocols for massive random access: the family of random access methods known asCoded Slotted ALOHA (CSA). A recent trend in research hasexplored aspects of such methods in various contexts, but oneaspect has not been fully taken into account: the impact of pathloss, which is a major design constraint in long-range wireless networks. In one article, we explored the behavior of CSA, by focusing on the path loss component correlated to the distanceto the base station. Path loss provides opportunities for capture,improving the performance of CSA. We revised methods for estimating CSA behavior, provide bounds of performance, and then,focusing on the achievable throughput, we extensively explored the key parameters, and their associated gain (experimentally). Our results shed light on the behavior of the optimal distribution of repetitions in actual wireless networks.

6.4.2. Real Implementation of Coded Slotted Aloha

- Participants: Cédric Adjih, Vinod Kumar

In 2017, we implemented Coded Slotted Aloha (CSA) as a proof of concept on our FIT IoT-LAB testbed (with 20+ nodes), with 802.15.4 transmissions and using a real SDR software.

This was presented in the seminar of the GT task 2 Future Access Networks of Digicosme. It was also presented as part of the tutorial "IoT in practice" in the ANTS 2017 conference.

6.5. Resource and Traffic Management

6.5.1. Utility Optimization Approach to Network Cache Design

Participants: Mostafa Dehghan, Laurent Massoulié, Don Towsley, Daniel Menasche, Y.c. Tay.

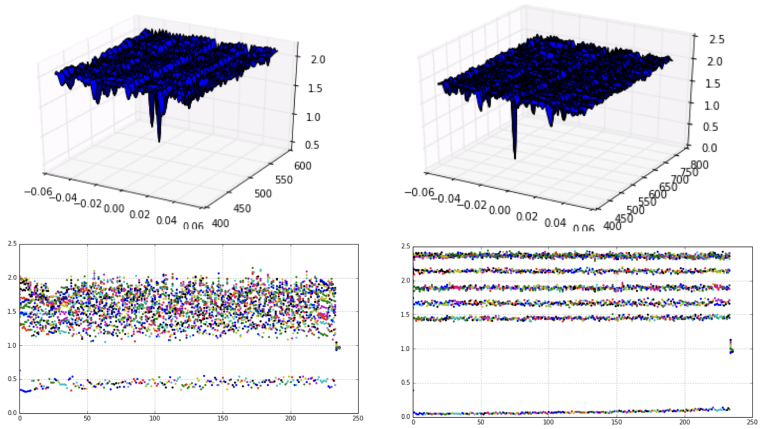

Figure 2. Example of demodulation steps in Coded Slotted Aloha

In any caching system, the admission and eviction policies determine which contents are added and removed from a cache when a miss occurs. Usually, these policies are devised so as to mitigate staleness and increase the hit probability. Nonetheless, the utility of having a high hit probability can vary across contents. This occurs, for instance, when service level agreements must be met, or if certain contents are more difficult to obtain than others. In this paper, we propose utility-driven caching, where we associate with each content a utility, which is a function of the corresponding content hit probability. We formulate optimization problems where the objectives are to maximize the sum of utilities over all contents. These problems differ according to the stringency of the cache capacity constraint. Our framework enables us to reverse engineer classical replacement policies such as LRU and FIFO, by computing the utility functions that they maximize. We also develop online algorithms that can be used by service providers to implement various caching policies based on arbitrary utility functions.

This work was published and presented at the IEEE Infocom 2016 conference as ”A Utility Optimization Approach to Network Cache Design”.

MADYNES Team

7. New Results

7.1. Monitoring

7.1.1. Quality of Experience Monitoring

Participants: Isabelle Chrisment [contact], Thibault Cholez, Vassili Rivron, Lakhdar Meftah [University of Lille].

We have pursued our work on smartphone usage monitoring with the SPIRALS team (Inria/Université de Lille) and more specifically on proposing new methods to help measure the QoE and to protect the user's privacy when collecting such data.

In parallel, to evaluate our methods, we need a testing framework to automate testing of WiFi P2P mobile apps at scale. In [20] we proposed AndroFleet, a large-scale WiFi P2P testing framework. Androfleet can perform User Acceptance Testing for a fleet of emulators, by emulating the hardware behavior of the peer discovery, it gives the developers the ability to control P2P specific behaviors (peers joining and leaving).

7.1.2. Active Monitoring

Participants: Abdelkader Lahmadi [contact], Jérôme François, Frédéric Beck [LHS], Loic Rouch [LHS].

Following the work done in 2016, we pursued our collaboration with the regional PME TracIP (http://www.tracip.fr) on the development of attack assessment and forensics platform dedicated to industrial control systems. The platform involves multiple PLC from different manufacturers and real devices of factory automation systems (see 6.7.1).

During the year 2017, we have demonstrated that off-the-shelf hardware is sufficient to take over any Z-Wave network without knowing its topology or compromising any original devices and remaining unnoticeable for the primary controller. Our attack consists in building an adversary Z-Wave universal controller by reprogramming a mainstream USB stick controller. The technique exploits two features provided by the USB stick which allow (1) to set the network identifier (HomeID) and (2) to learn many devices identifiers even if they are not physically available. The attack has been demonstrated in Blackhat Europe 2017 by Loic Rouch (https://www.blackhat.com/eu-17/briefings/schedule/#a-universal-controller-to-take-over-a-z-wave-network-8459).

7.1.3. Service-level Monitoring of HTTPS traffic

Participants: Thibault Cholez [contact], Wazen Shbair, Jérôme François, Isabelle Chrisment.

We previously proposed an alternative technique to investigate HTTPS traffic which aims to be robust, privacy-preserving and practical with a service-level identification of HTTPS connections, i.e. to name the services, without relying on specific header fields that can be easily altered. We have defined dedicated features for HTTPS traffic that are used as input for a multi-level identification framework based on machine learning algorithms processing full TLS sessions. Our evaluation based on real traffic shows that we can identify encrypted web services with a high accuracy. In 2017, we finished to develop our solution to make it fully usable in real-time [1]. We now provide our prototype implementation (https://gitlab.inria.fr/swazen/HTTPSFirewall) in open-source. It operates by extending the iptables/netfilter architecture. It receives and demultiplexes the arriving HTTPS packets to a related flow. As soon as the number of packets in a given flow reaches a threshold, the identification engine extracts the features and runs the C4.5 algorithm to predict the HTTPS service of the flow.

7.1.4. Monitoring Programmable Networks

Participants: Jérôme François [contact], Olivier Festor, Paul Chaignon [Orange Labs], Kahina Lazri [Orange Labs], Thibault Delmas [Orange Labs].

Software-Defined Networking brings new capabilities in operating networks including monitoring. In the state-of-the art many proposals have been made to enhance monitoring of networks using OpenFlow or other proposed programmable frameworks. In a preliminary work [11], we reviewed them in order to highlight what are the remaining challenges to be addressed in that area. The main issue is the trade-off to be made between the strong expressibility (especially stateful operations) and capability of monitoring techniques that are necessary for advanced operation purposes and the complexity it induces if we want to keep the pace with line-rate packet processing. Another important aspect is the security as adding programmable monitoring functions may lead to introduce security threats. Our current work is thus focused on adding monitoring capacity while guaranteeing line-rate operations and safety requirements even when programs are deployed on running network switches.

7.1.5. Smart Contracts Monitoring

Participants: Jérôme François [contact], Sofiane Lagraa, Radu State [University of Luxembourg], Jérémy Charlier [University of Luxembourg].

Blockchain technologies are skyrocketing and the team is interested in assessing the impact of such technologies on networking, and if necessary managing the coupling between them. Indeed, blockchain efficiency resides in an overlay network built on top of a real infrastructure which needs to properly support it. Orchestrating network ressources, *i.e.* adding some network capacity, might be helpful but supposes first an in-depth monitoring of blockchain interactions. In a first work, we thus evaluated the relation among smart contracts. We defined methods to discover smart contracts interactions and the different group properties. This approach relies on graph modelling and mining techniques as well as tensor modelling combined with stochastic processes. It underlines actual exchanges between smart contracts and targets the predictions of future interactions among the communities. Comparative study between graph analysis and tensor analysis is provided for predictions of smart contract interactions. Finally, virtual reality visualization based on Unity 3D game engine has been applied [12].

7.1.6. Sensor networks monitoring

Participants: Rémi Badonnel, Isabelle Chrisment, Olivier Festor, Abdelkader Lahmadi [contact], Anthea Mayzaud.

Our work on IoT security monitoring has been published in IEEE Transactions on Network and Service Management [4]. This concerns more specifically our distributed monitoring architecture for detecting attacks against RPL networks. The RPL routing protocol has been standardized by IETF to enable a lightweight and robust routing in lower-power and lossy networks. After having compared existing IoT monitoring solutions, we have proposed a detection strategy for RPL version number attacks. This one relies on our monitoring architecture to preserve constrained node resources, in the context of AMI infrastructures. A versioning mechanism is incorporated into RPL in order to maintain an optimized topology. However, an attacker can exploit this mechanism to significantly damage the network and reduce its lifetime. We have exploited monitoring node collaboration to identify the attacker, the localization process being performed by the root after gathering detection information from all monitoring nodes. We have evaluated our solution through experiments and have analyzed the performance according to defined metrics. We have shown that the false positive rate of our solution can be reduced by a strategic monitoring node placement. We have also considered the scalability issue, by modeling this placement as an optimization problem and quantifying the number of required monitoring nodes to ensure acceptable false positive rates.

7.2. Security

7.2.1. Security analytics

Participants: Jérôme François [contact], Abdelkader Lahmadi, Sofiane Lagraa, Soline Blanc, Giulia de Santis, Olivier Festor, Radu State [University of Luxembourg], Christian Hammerschmidt [University of Luxembourg].

In 2017, we have continued our active cooperation with the High Security Lab (HSL) in Nancy. The latter provides the infrastructure to support two main projects in security analytics, namely the FUI HuMa project and the ATT AMICS. Thanks to darknet data of the HSL, we developped two methods based on graph-mining to extract knowledge. The first one focuses on port scanning analysis in order to profile the behaviours and patterns of attackers. By representing consecutive targeted ports in an aggregated graph format, we assess then the centrality of port number using different metrics and highlights valuable correlation among some of them. We are particularly able to identify patterns of scanning related to a specific setup (e.g. medical environment) [17]. We then extended this method to security events analysis by constructing multiple graphs to be analyzed with an outlier technique. The rationale is to represent individual behaviors and detect those which deviate from the majority. The method has been successfully applied to botnet detection in [16]. We are currently leveraging our graph analysis in order to provide to the community a new metric or distance to be applied when comparing port numbers. Indeed, numerical comparison is meaningless in that context and we could leverage either a semantic database (such as Wikipedia) or attacker database (darknet) to derive a meaningful metric, *i.e.* representing a real correlation between port numbers (TCP or UDP).

Furthermore, we continue our work on using Hidden Markov Models for analysing TCP scanning activities. We are now in a stage where individual models from different scanner tools or configurations (e.g. targeted ports) are used in order to automatically learn unique signatures then applied on non-labelled data.

7.2.2. NDN Security

Participants: Thibault Cholez [contact], Xavier Marchal, Olivier Festor, Jérôme François, Salvatore Signorello [University of Luxembourg], Radu State [University of Luxembourg], Samuel Marchal [Aalto University].

Information Centric Networking (ICN) is seen as a promising solution to re-conciliate the Internet usage with its core architecture. However, to be considered as a realistic alternative to IP, ICN must evolve from a pure academic proposition deployed in test environments to an operational solution in which security is assessed from the protocol design to its running implementation. Among ICN solutions, Named Data Networking (NDN), together with its reference implementation NDN Forwarding Daemon (NFD), acts as the most mature proposal but its vulnerability against the Content Poisoning Attack (CPA) is considered as a critical threat that can jeopardize this architecture. So far, existing works in that area have fallen into the pit of coupling a biased and partial phenomenon analysis with a proposed solution, hence lacking a comprehensive understanding of the attack's feasibility and impact in a real network. In a joint work with our colleagues from UTT and in the context of the ANR DOCTOR projet, we demonstrated through an experimental measurement campaign that CPA can easily and widely affect NDN. Our contribution is threefold: (1) we propose three realistic attack scenarios relying on both protocol design and implementation weaknesses; (2) we present their implementation and evaluation in a testbed based on the latest NFD version; and (3) we analyze their impact on the different ICN nodes (clients, access and core routers, content provider) composing a realistic topology. This work was published in IM 2017 conference [21].

Also, still in the context of the DOCTOR project, we refined our architecture to securely deploy NDN over NFV. Indeed, combining NFV fast service deployment and SDN fine grained control of data flows allows comprehensive network security monitoring. The DOCTOR architecture allows detecting, assessing and remediating attacks. NDN is an example of application made possible by SDN and NFV coexistence, since hardware implementation would be too expensive. We showed how NDN routers can be implemented and managed as VNFs. Security monitoring of the DOCTOR architecture is performed at two levels. First, host-level monitoring, provided by CyberCAPTOR, uses an attack graph approach based on network topology knowledge. It then suggests remediations to cut attack paths. We show how our monitoring tool integrates SDN and NFV specificities and how SDN and NFV make security monitoring more efficient. Then, application-level monitoring relies on the MMT probe. It monitors NDN-specific metrics from inside the VNFs and a central component can detect attack patterns corresponding to known flaws of the NDN protocol. These attacks are fed to the CyberCAPTOR module to integrate NDN attacks in attack graphs. This work was published in a book chapter "Guide to Security in SDN and NFV" from Springer's Computer Communications and Networks collection [35].

Finally, in cooperation with the University of Luxembourg, we have investigated interest flooding attacks in NDN. By nature, NDN communication assumes that requesting a content leads to emit an interest and forwarding it in the network until it reaches an appropriate content provider which then sends back data through the reverse path. Interest flooding attacks forge interests (requests) which cannot be satisfied by any data to be sent back to the emitter. As such, both the network and nodes are overloaded as the interests are flooded into the network and intermediate nodes have to store them locally in the pending interest table. We observed that most of literature mechanisms have been evaluated with very simple attack models. Actually, we had a great expertise in phishing attacks and social engineering that can be used to generate realistic phishing names for the NDN naming scheme. We thus create a new stealthy attack relying on natural language processing techniques to forge interests very similar to legitimate ones making inefficient all proposed counter-measures from the state-of-the-art [25].

7.2.3. Configuration security automation

Participants: Rémi Badonnel [contact], Abdelkader Lahmadi, Olivier Festor, Nicolas Schnepf, Maxime Compastié.

The main research challenge addressed in this work is focused on enabling configuration security automation in dynamic networks and services. In particular our objective is to support the efficient configuration and orchestration of security management operations.

The continuous growth and variety of networking significantly increases the complexity of management. It requires novel autonomic methods and techniques contributing to detection and prevention performances with respect to vulnerabilities and attacks.

We have pursued during Year 2017 the efforts on the orchestration of security functions in the context of mobile smart environments, with our joint work with Stephan Merz of the VeriDis project-team at Inria Nancy. We had already defined an automated verification technique, based on an extension of an SDN language, for checking both the control and the data planes related to security chains [24]. Complementarily, we proposed a strategy for generating SDN policies for protecting Android environments based on automata learning. Our solution collects traces of flow interactions of their applications, aggregates them in order to build finite-state models, and then infer SDN policy rules. We have designed and implemented aggregation and automata learning algorithms that allow precise and generic models of applications to be built. These models will be then used for configuring chains of security functions specified in the Pyretic language and verified with our Synaptic checker. We have developed a prototype of our solution implementing these algorithms, and evaluated its performances through a series of experiments based on the backend process miners Synoptic and Invarimint, in addition to our own algorithm. The experiments showed the benefits and limits of these methods in terms of simplicity, precision, genericity and expressivity, while varying the level of aggregation of the input flow traces.

In addition, we have worked on our software-defined security framework, for enabling the enforcement of security policies in distributed cloud environments. This framework relies on the autonomic paradigm to dynamically configure and adjust these mechanisms to distributed cloud constraints, and exploit the software-defined logic to express and propagate security policies to the considered cloud resources [13]. In particular, we have investigated during Year 2017 the exploitability of unikernels to support our framework. Unikernels permit to build highly-constrained configurations limited to the strict necessary with a time-limited validity. We take benefits of their properties to reduce the attack exposure of cloud resources. We have formalized and integrated into our software-defined security framework, on-the-fly generation mechanisms of unikernel images that cope with security policy requirements. In that context, security mechanisms are directly integrated to the unikernel images at building time. A proof of concept prototype based on MirageOS was developed and the performance of such a software-based security strategy was evaluated through extensive series of experiments. We have also compared them to other regular virtualization solutions. Our results show that the costs induced by security mechanisms integration are relatively limited, and unikernels are well suited to minimize risk exposure.

7.3. Experimentation, Emulation, Reproducible Research

This section covers our work on experimentation on testbeds (mainly Grid'5000), on emulation (mainly around the Distem emulator), and on Reproducible Research.

7.3.1. Grid'5000 design and evolutions

Participants: Florent Didier, Arthur Garnier, Imed Maamria, Lucas Nussbaum [contact], Olivier Demengeon [SED], Teddy Valette [SED].

The team was again heavily involved in the evolutions and the governance of the Grid'5000 testbed.

7.3.1.1. Technical team management

Since the beginning of 2017, Lucas Nussbaum serves as the Grid'5000 *directeur technique* (CTO), managing the global technical team (9 FTE).

7.3.1.2. SILECS project

We are also heavily involved in the ongoing SILECS project, that aims at creating a new infrastructure on top of the foundations of Grid'5000 and FIT in order to meet the experimental research needs of the distributed computing and networking communities.

7.3.1.3. Promoting the testbed

In order to promote the testbed to the french devops and sysadmin community, we presented in [27] an overview of the testbed's capabilities.

7.3.1.4. Disk reservation

We contributed a new feature that will greatly help Big Data experimenters: the ability to reserve disks on nodes, in order to leave large datasets stored on nodes between nodes reservations.

7.3.1.5. Automated testing of the testbed

In order to ensure that all services remain functional, and that experimental results remain trustworthy and reproducible, we designed an infrastructure to automatically test the testbed and detect misconfigurations, regressions, uncontrolled hardware heterogeneity, etc. This work was described in [23] and later presented in [34].

7.3.1.6. Support for SDN experiments

We started the development of a tool to orchestrate SDN experiments on Grid'5000, combining KaVLAN and OpenVSwitch.

7.3.2. Emulation with Distem

Participants: Alexandre Merlin, Lucas Nussbaum [contact].

The ADT SDT project started in March. Initial work focused on improving the software developing infrastructure by adding automated regression tests on both correctness and performance. This should allow a new release in early 2018.

7.3.3. I/O access patterns analysis with eBPF

Participants: Abdulqawi Saif, Lucas Nussbaum [contact], Ye-Qiong Song.

In the context of Abdulqawi Saif's CIFRE PhD (with Xilopix), we explored the relevance of an emerging instrumentation technology for the Linux kernel, eBPF, and used it to analyze I/O access patterns of two popular NoSQL databases. A publication on this topic is expected in early 2018.

7.3.4. Performance study of public clouds

Participants: Souha Bel Haj Hassine, Lucas Nussbaum [contact].

We worked on clouds performance in the context of an ongoing collaboration with *CloudScreener*, a French startup founded in 2012 that has developed tools for cloud price and performance benchmarks and automated cloud recommendation to optimize the decision making process in the context of cloud computing. We designed methods and tools to do performance evaluation of public clouds focusing on (1) outlining performance variability over time; (2) identifying adverse strategies that might be deployed by cloud providers in order to vary the performance level over time.

7.3.4.1. Testbeds federation and collaborations in the testbeds community

The Fed4FIRE+ H2020 project started in January 2017 and will run until the end of September 2021. This project aims at consolidating the federation of testbeds in Europe of which Grid'500 is a member.

We are also active in the GEFI initiative that aims at building links between the US testbeds community (GENI) and their european (FIRE), japanese and brazilian counterparts. We participated in the annual GEFI meeting where gave two talks [33][34] and chaired the session on reproducibility.

7.3.4.2. Experimentation and reproducible research

In addition to the work already mentioned on testbed testing [23], [34], we worked on a survey of testbeds and their features for reproducible research [22]. We also gave several talks on reproducible research and testbeds at *École ARCHI* [5], *École RESCOM* [6], and Inria webinars on Reproducible Research [7].

7.4. Routing

7.4.1. NDN routing

Participants: Isabelle Chrisment [contact], Thomas Silverston, Elian Aubry.

As NDN relies on content names instead of host address it cannot rely on traditional Internet routing. Therefore it is essential to propose a routing scheme adapted for NDN. In [8] we have presented SRSC, our SDN-based Routing Scheme for CCN/NDN and its implementation. SRSC relies on the SDN paradigm. A controller is responsible to forward decisions and to set up rules into NDN nodes. So we have implemented SRSC into NDNx. We have deployed an NDN testbed within a virtual environment emulating a real ISP topology in order to evaluate the performances of our proposal with real-world experiments. We have demonstrated the feasibility of SRSC and its ability to forward Interest messages in a fully deployed NDN environment while keeping low overhead and computation time and high caching performances.

7.4.2. Energy-Aware and QoS Routing for Wireless Sensor Networks

Participants: Evangelia Tsiontsiou, Bernardetta Addis, Ye-Qiong Song [contact].

The main research problems in the domain of routing data packets in a multi-hop wireless sensor network are the optimisation of the energy and the routing under multi-criteria QoS constraints (e.g., energy, reliability, delay, ...). To address these problems, we proposed, in the PhD thesis of E. Tiontsiou, two contributions. The first contribution is an optimal probabilistic energy-aware routing protocol, allowing to energy usage balancing. Comparing to the existing probabilistic routing protocols, our solution is based on the computation of the optimal probabilities by solving a linear programming problem. Our second contribution is an operator calculus algebra based multi-constrained routing protocol. It is fundamentally different from the existing solutions since it can simultaneously consider several constraints, instead of their combination.

7.5. Smart*: design, multi-modeling and co-simulation and supervision of mobile CPS/IoT

Participants: Laurent Ciarletta [contact], Ye-Qiong Song, Yannick Presse, Julien Vaubourg, Emmanuel Nataf, Petro Aksonenko, Virgile Dauge, Louis Viard, Florian Greff, Virginie Galtier, Thomas Paris.

Vincent Chevrier (former Maia team, Dep 5, LORIA) is a collaborator and the correspondant for the MS4SG/MECSYCO project, as well as Christine Bourjot (former MAIA team, Dep 5, LORIA).

Sylvain Contassot-Vivier (Dep 3, Loria) is a collaborator on the Grone project and is directing Virgile Daugé with Laurent Ciarletta.

Pierre-Etienne Moreau is a collaborator on the CEOS project and is directing Louis Viard with Laurent Ciarletta.

Virginie Galtier from CentraleSupélec is now a member of the Loria laboratory and will integrate the future Simbiot team (Systems of Interactive aMBient Intelligent ObjecTs).

In Pervasive or Ubiquitous Computing, a growing number of communicating/computing devices are collaborating to provide users with enhanced and ubiquitous services in a seamless way.

These systems, embedded in the fabric of our daily lives, are complex: numerous interconnected and heterogeneous entities are exhibiting a global behavior impossible to forecast by merely observing individual properties. Firstly, users physical interactions and behaviors have to be considered. They are influenced and influence the environment. Secondly, the potential multiplicity and heterogeneity of devices, services, communication protocols, and the constant mobility and reorganization also need to be addressed. Our research in this field is going towards both closing the loop between humans and systems, physical and computing systems, and taming the complexity, using multi-modeling (to combine the best of each domain specific model) and co-simulation (to design, develop and evaluate) as part of a global conceptual and practical toolbox.

We proposed the AA4MM meta-model [45] that solves the core challenges of multimodeling and simulation coupling in an homogeneous perspective. In AA4MM, we chose a multi-agent point of view: a multi-model is a society of models; each model corresponds to an agent and coupling relationships correspond to interaction between agents. In the MECSYCO-NG (formerly MS4SG, Multi Simulation for Smart Grids) project which involves some members of the former MAIA team, Madynes and EDF R&D on smart-grid simulation, we developed a proof of concepts for a smart-appartment case that serves as a basis for building up use cases, and we have worked on some specific cases provided by our industrial partner. We also collaborated with researchers from the Green UL laboratory.

In 2017 we worked on the following research topics:

- Overall assessment and evaluation of complex systems.
- Cyber Physical Systems and Smart *.

  We have continued the design and implementation of the Aetournos platform at Loria which will be part of the Creativ'Lab. The collective movements of a flock of flying communicating robots / UAVs, evolving in potentially perturbed environment constitutes a good example of a Cyber Physical System. Several projects have started during the last part of 2017. One of the emerging topic in this area is the safety of Mobile IoT/CPS with regards to their environment and users.

  – The Grone (Interreg) project involves partners from the 4 countries of the Grande Région (Centrale Supélec, LIST, Univ Luxembourg, Univ Liège, Fraunhofer IZFP to name a few). The main goal is to develop UAV based solution for the surveillance of industrial and agricultural sites and the exploration of GPS denied and underground environments. A PhD has been started in March 2017 (Systèmes cyber-physiques autonomes et communicants en milieux hostiles. Application à l'exploration par robots mobiles. Virgile Daugé).

  – and the CEOS (FUI22) project involving high profile companies (Thales C&S, EDF R&D, ENEDIS and Aéroport de Lyon) as well as academic partners (AOSTE2 Inria, ESIEE) and collaborating SMEs (RT@W, ADCIS, Alerion). This project focuses on the safety of UAV based monitoring solutions for OIV (Opérateurs d'Intéret Vital) infrastructures. A PhD has been started in November 2017 (Environnement de développement et d'analyse de propriétés pour des systèmes cyber-physiques mobiles. Louis Viard).

  The work on Software Defined Real-Time Mesh networks (Florian Greff's PhD CIFRE with Thales R&T) has given many results as he plans to defend his work in march 2018 [15], [39], [14].

On more specific subject of innovative sensors for mobile and interactive IoT, a collaborative project with the KPI (Ukraine) university has been started with a projected PhD (Méthodes optimisées de calibration, d'alignement et algorithmes d'attitude avancés pour les systèmes de navigation inertiels fixés, Petro Aksonenko). Several papers have been published [9] [44] in 2017.

- (Very Serious) Gaming: Starburst Gaming. During some exploratory work, we have seen the potential of these Pervasive Computing ressources in the (Very Serious) Gaming area which led us to the Starburst Computing SATT projects in 2016 and 2017. A spin off has been founded in 2017 that is getting the licences for the resulting IP (the software is under the APP process at the time of this writing). Starburst is already involved in a AMI project with the Globlinz game studio and the lab and has officially been accepted in novembre 2017 and will be operational in 2018.
- Smart *: MS4SG / MECSYCO-NG has given us the opportunity to link simulations tools with a strong focus on FMI (Functional Mockup Interface) and network simulators (NS3/Omnet++). We have so far successfully applied our solution to the simulation of smart apartment complex and to combine the electrical and networking part of a Smart Grid. The AA4MM software is now MECSYCO and has seen constant improvements in 2017 thanks to the ressources provided by the MECSYCO-NG project in collaboration with EDF R&D (http://www.mecsyco.com), and the work of Thomas Paris and Julien Vaubourg.

  Starting from domain specific and heterogenous models and simulators, the MECSYCO suite allows for multi systems integration at several levels: conceptual, formal and software. A couple of visualization tools have been developed as proof of concepts both at run-time and post-mortem.

The technical report [43] has been extended into a journal paper under revision for a publication in 2018.

7.6. Quality-of-Service

7.6.1. Self-adaptive MAC protocol for both QoS and energy efficiency in IoT

Participants: Shuguo Zhuo, François Despaux, Ye-Qiong Song [contact].

The diversity of IoT applications implies the requirement of reliable yet efficient MAC solutions for supporting transmissions for various traffic patterns. We have mainly contributed to enhance the implementation of the high efficient traffic self-adaptive MAC protocols. As part of RIOT ADT project, our main achievements are the development of two MAC protocols lw-MAC and GoMacH [26]. lw-MAC is similar to X-MAC and ContikiMAC. It allows to introduce a first duty-cycled MAC into RIOT IoT protocol stack. GoMacH is a nearly optimal protocol that provides high reliability and throughput for handling various traffic loads in IoT. GoMacH seamlessly integrates several outstanding techniques. It adopts the phase-lock scheme to achieve low-power duty-cycled communication. It also utilizes a dynamic slots allocation scheme for providing accurate and instantaneous throughput boost. Furthermore, like in TSCH, GoMacH spreads its communications onto IEEE 802.15.4's 16 channels, leading to high reliability. GoMacH has been implemented in open source on RIOT OS, and has also been seamlessly integrated into IETF's 6LoWPAN/RPL/UDP stack as well as CCN-light. Experimental results on SAMR21-xpro test-beds and IoT-LAB verify the practicality of GoMacH and its capabilities for consistently providing high throughput, high delivery ratio, and low radio duty-cycle. They are both publically available on the RIOT open source github.

7.6.2. QoS and fault-tolerance in distributed real-time systems

Participants: Florian Greff, Laurent Ciarletta, Arnauld Samama [Thales TRT], Dorin Maxim, Ye-Qiong Song [contact].

The QoS must be guaranteed when dealing with real-time distributed systems interconnected by a network. Not only task schedulability in processors, but also message schedulability in networks should be analyzed for validating the system design. Fault-tolerance is another critical issue that one must take into account.

In collaboration with Thales TRT industrial partner as part of a CIFRE PhD work, we have developed a Software-Defined Real-time Network (SDRN) framework [14]. SDRN deals with the real-time flow allocation problem in mesh networks. The objective is to find a suitable path under delay constraint while allowing load balancing. For this purpose, combined online flow admission control and pathfinding algorithms have been developed on an SDN-like controller. At switch level, each output port is ruled by a credit-based weighted round robin, allowing isolation of flows. As a consequence, a freshly admitted flow will not influence existing flows, allowing incremental online admission of new flows. This approach has been applied to a RapidIO mesh network example and compared with the compositional performance analysis method. Numerical results clearly show the benefit of our proposal in terms of complexity and delay bound pessimism. In [15], Fault-tolerance issue in mesh networks has been addressed. In fact, one of the major advantages of a mesh topology is its ability to leverage the path redundancy in order to recover from link or node failures, through a flow reconfiguration process. However, one needs to ensure that hard real-time packets will keep being delivered on time during this transient reconfiguration period. Anticipating each possible fault is very complex and can result in a waste of network resource. Our contribution is the combination of an optimized content-centric source routing in nominal mode and a destination-tag flexible and scalable routing in transient recovery mode. We show the benefit of this approach in terms of flexibility and network resource utilization. Our method can ensure real-time properties enforcement even during the transient reconfiguration period. Algorithms have been developed to extend the SDRN flow allocation and routing methods in order to implement this hybrid fault-tolerant extension.

As part of Eurostars RETINA project, in the in-vehicle networking domain, we have focused on the evaluation of the worst-case response time of AVB traffic under time-aware shaper of TSN (time-sensitive networking). It is a hierarchical real-time scheduling problem, where a packet is scheduled by the credit-based shaper, priority and time-aware shaper (TDMA). We have proved that the eligible interval approach, developed for AVB, is still hold for TSN case. The worst case delay expression, as well as the feasibility condition are deduced. Our methods (analysis and simulation) are applied to an automotive use case, which is defined within the Eurostars RETINA project, and where both control data traffic and AVB traffic must be guaranteed. It has been shown that our delay bound is tight in single switch case [19].

NEO Project-Team

7. New Results

7.1. Stochastic Modeling

Participants: Alain Jean-Marie, Hlib Mykhailenko, Eleni Vatamidou.

7.1.1. Semi-Markov Accumulation Processes

E. Vatamidou and A. Jean-Marie have introduced in [37] a new accumulation process, the Semi-Markov Accumulation Process (SMAP). This class of processes extends the framework of continuous-time Markov Additive Processes (MAPs) by allowing the underlying environmental component to be a semi-Markov process instead of a Markov process. They follow an analytic approach to derive a Master Equation formula of the Renewal type that describes the evolution of SMAPs in time. They show that under exponential holding times, a matrix exponential form analogous to the matrix exponent of a MAP is attained. Finally, they consider an application of these results where closed-form solutions are rather easy to achieve.

7.1.2. The marmoteCore platform

The development of `marmoteCore` (see Section 6.1) has been pursued. The software library is now mature enough to develop complex models, such as in [33]. Its architecture and main capabilities have been presented in [26]. `marmoteCore` provides the classes necessary to represent the state space of Markov models, from the elementary bricks that are interval or rectangular domains, simplices, or binary sequences. From there, the user easily programs the construction of probability transition matrices or infinitesimal generators. Structural analysis methods allow to identify recurrent and transient classes, and to compute the period of the model. Numerous methods allow the Monte Carlo simulation of the chain, the computation of transient and stationary distributions, as well as hitting times. `marmoteCore` is organized in a hierarchy of Markov models, from the simplest ones (Poisson process, two-state chains, ...) to the most general ones, including classes of models with a particular interest, such as QBDs. It is therefore possible to program solution methods specifically optimized and adapted to the level of structure of the model.

7.2. Random Graph and Matrix Models

Participants: Arun Kadavankandy, Konstantin Avrachenkov.

In [27] A. Kadavankandy and K. Avrachenkov in collaboration with L. Cottatellucci (EURECOM, France) and R. Sundaresan (IIS Bangalore, India) propose a local message passing algorithm based on Belief Propagation (BP) to detect a small hidden Erdös-Rényi (ER) subgraph embedded in a larger sparse ER random graph in the presence of side-information. The side-information considered is in the form of revealed subgraph nodes called cues, some of which may be erroneous. Namely, the revealed nodes may not all belong to the subgraph, and it is not known to the algorithm a priori which cues are correct and which are incorrect. The authors show that asymptotically as the graph size tends to infinity, the expected fraction of misclassified nodes approaches zero for any positive value of a parameter λ, which represents the effective Signal-to-Noise Ratio of the detection problem. Previous works on subgraph detection using BP without side-information showed that BP fails to recover the subgraph when $\lambda < 1/e$. These new results thus demonstrate the substantial gains in having even a small amount of side-information.

PageRank has numerous applications in information retrieval, reputation systems, machine learning, and graph partitioning. In [8] K. Avrachenkov and A. Kadavankandy in collaboration with L. Ostroumova and A. Raigorodskii (Yandex, Russia) study PageRank in undirected random graphs with an expansion property. The Chung-Lu random graph is an example of such a graph. They show that in the limit, as the size of the graph goes to infinity, PageRank can be approximated by a mixture of the restart distribution and the vertex degree distribution. They also extend the result to Stochastic Block Model (SBM) graphs, where they show that there is a correction term that depends on the community partitioning.

7.3. Data Analysis and Learning

Participants: Konstantin Avrachenkov, Hlib Mykhailenko, Giovanni Neglia, Dmytro Rubanov.

7.3.1. Unsupervised learning

In [21] K. Avrachenkov in collaboration with A. Kondratev and V. Mazalov (both from Petrozavodsk State Univ., Russia) apply cooperative game-theoretic methods for community detection in networks. The traditional methods for detecting community structure are based on selecting denser subgraphs inside the network. Their new approach is to use the methods of cooperative game theory that highlight not only the link density but also the mechanisms of cluster formation. Specifically, they suggest two approaches from cooperative game theory: the first approach is based on the Myerson value, whereas the second approach is based on hedonic games. Both approaches allow to detect clusters with various resolution. However, the tuning of the resolution parameter in the hedonic games approach is particularly intuitive. Furthermore, the modularity based approach and its generalizations can be viewed as particular cases of the hedonic games.

Kernels and, broadly speaking, similarity measures on graphs are extensively used in graph-based unsupervised and semi-supervised learning algorithms as well as in the link prediction problem. In [19] K. Avrachenkov and D. Rubanov in collaboration with P. Chebotarev (Trapeznikov Institute of Control Sciences, Russia) analytically study proximity and distance properties of various kernels and similarity measures on graphs. This can potentially be useful for recommending the adoption of one or another similarity measure in a machine learning method. Also, they numerically compare various similarity measures in the context of spectral clustering and observe that normalized heat-type similarity measures with log modification generally perform the best.

7.3.2. Semi-supervised learning

Graph-based semi-supervised learning for classification endorses a nice interpretation in terms of diffusive random walks, where the regularisation factor in the original optimisation formulation plays the role of a restarting probability. Recently, a new type of biased random walks for characterising certain dynamics on networks have been defined and rely on the γ-th power of the standard Laplacian matrix L. In particular, these processes embed long range transitions, the Levy flights, that are capable of one-step jumps between far-distant states (nodes) of the graph. In [24] K. Avrachenkov in collaboration with E. Bautista, S. De Nigris, P. Abry and P. Gonçalves (from DANTE Inria team and ENS Lyon) build upon these volatile random walks to propose a new version of graph based semi-supervised learning algorithms whose classification outcome could benefit from the dynamics induced by the fractional transition matrix. In [22] using the framework of Levy flights, they further improve the classification outcome, even in settings traditionally poorly performing such as unbalanced classes, and they derive a theoretical rule for classification decision.

In [6] K. Avrachenkov in collaboration with P. Chebotarev (Trapeznikov Institute of Control Sciences, Russia) and A. Mishenin (Saint Petersburg Univ., Russia) study a semi-supervised learning method based on the similarity graph and regularized Laplacian. They give convenient a optimization formulation of the regularized Laplacian method and establish its various properties. In particular, they show that the kernel of the method can be interpreted in terms of discrete and continuous-time random walks and possesses several important properties of proximity measures. Both optimization and linear algebra methods can be used for efficient computation of the classification functions. The authors demonstrate on numerical examples that the regularized Laplacian method is robust with respect to the choice of the regularization parameter and outperforms the Laplacian-based heat kernel methods.

7.3.3. Distributed computing

In distributed graph computation, graph partitioning is an important preliminary step, because the computation time can significantly depend on how the graph has been split among the different executors. In [30] H. Mykhailenko and G. Neglia, in collaboration with F. Huet (I3S) propose a framework for distributed edge partitioning based on simulated annealing. The framework can be used to optimize a large family of partitioning metrics. They provide sufficient conditions for convergence to the optimum as well as discuss

which metrics can be efficiently optimized in a distributed way. They implemented these partitioners in Apache GraphX and performed a preliminary comparison with JA-BE-JA-VC, a state-of-the-art partitioner that inspired the new approach. They show that this approach can provide improvements, but further research is required to identify suitable metrics to optimize as well as to design a more efficient exploration phase for the algorithm without sacrificing convergence properties.

Because of the significant increase in size and complexity of the networks, the distributed computation of eigenvalues and eigenvectors of graph matrices has become very challenging and yet it remains as important as before. In [20] K. Avrachenkov in collaboration with P. Jacquet (Nokia Bell Labs) and J. Sreedharan (Purdue Univ., USA) develop efficient distributed algorithms to detect, with higher resolution, closely situated eigenvalues and corresponding eigenvectors of symmetric graph matrices. We model the system of graph spectral computation as physical systems with Lagrangian and Hamiltonian dynamics. The spectrum of Laplacian matrix, in particular, is framed as a classical spring-mass system with Lagrangian dynamics. The spectrum of any general symmetric graph matrix turns out to have a simple connection with quantum systems and it can be thus formulated as a solution to a Schrödinger-type differential equation. Taking into account the higher resolution requirement in the spectrum computation and the related stability issues in the numerical solution of the underlying differential equation, we propose the application of symplectic integrators to the calculation of eigenspectrum. The effectiveness of the proposed techniques is demonstrated with numerical simulations on real-world networks of different sizes and complexities.

7.4. Game Theory

Participants: Eitan Altman, Konstantin Avrachenkov.

7.4.1. Dynamic potential games

In [11] K. Avrachenkov in collaboration with V. Mazalov and A. Rettieva (both from Petrozavodsk State Univ., Russia) treat discrete-time game-theoretic models of resource exploitation as dynamic potential games. The players (countries or firms) exploit a common stock on the infinite time horizon. The main aim is to obtain a potential for the linear-quadratic games of this type. The class of games where a potential can be constructed as a quadratic form is identified. As an example, the dynamic game of bioresource management is considered and the potentials are constructed in the case of symmetric and asymmetric players.

7.4.2. A Hawk and Dove game with infinite state space

In [16], E. Altman, in collaboration with A. Aradhye and R. El-Azouzi (UAPV) consider the Hawk-Dove game in which each of infinitely many individuals, involved with pairwise encounters with other individuals, can decide whether to act aggressively (Hawk) or peacefully (Dove). Each individual is characterized by its strength. The strength distribution among the population is assumed to be fixed and not to vary in time. If both individuals involved in an interaction are Hawks, there will be a fight, the result of which will be determined by the strength of each of the individuals involved. The larger the difference between the strength of the individuals is, the larger is the cost for the weaker player involved in the fight. The goal is to study the influence of the parameters (such as the strength level distribution) on the equilibrium of the game. The authors show that for some parameters there exists a threshold equilibrium policy while for other parameters there is no equilibrium policy at all.

7.5. Applications in Telecommunications

Participants: Zaid Allybokus, Eitan Altman, Konstantin Avrachenkov, Giovanni Neglia, Sarath Pattathil, Berksan Serbetci, Alina Tuholukova.

7.5.1. Caching

As cellular network operators are struggling to keep up with the rapidly increasing traffic demand, two key directions are deemed necessary for beyond 4G networks: (i) extensive cell densification to improve spatial reuse of wavelengths, and (ii) storage of content as close to the user as possible to cope with the backhaul constraints and increased interference. However, caching has mostly been studied with an exclusive focus either on the backhaul network (e.g. the "femto-caching" line of work) or on the radio access (e.g. through coded caching or cache-aided Coordinated MultiPoint, CoMP). As a result, an understanding of the impact of edge caching on network-wide and end-to-end performance is lacking. In [32] A. Tuholukova and G. Neglia in collaboration with T. Spyropoulos (EURECOM) investigate the problem of optimal caching in a context where nearby small cells ("femto-helpers") can coordinate not just in terms of what to cache but also to perform Joint Transmission (a type of CoMP). They show that interesting tradeoffs arise between caching policies that improve radio access and ones that improve backhaul, and propose an algorithm that provably achieves an 1/2-approximation ratio to the optimal one (which is NP-hard), and performs well in simulated scenarios.

Cache policies to minimize the content retrieval cost have been studied through competitive analysis when the miss costs are additive and the sequence of content requests is arbitrary. More recently, a cache utility maximization problem has been introduced, where contents have stationary popularities and utilities are strictly concave in the hit rates. In [29] G. Neglia in collaboration with D. Carra (Univ. of Verona) and P. Michiardi (EURECOM) bridges the two formulations, considering linear costs and content popularities. They show that minimizing the retrieval cost corresponds to solving an online knapsack problem, and we propose new dynamic policies inspired by simulated annealing, including DYNQLRU, a variant of QLRU. For such policies they prove asymptotic convergence to the optimum under the characteristic time approximation. In a real scenario, popularities vary over time and their estimation is very difficult. DYNQLRU does not require popularity estimation, and realistic, trace-driven evaluation shows that it significantly outperforms state-of-the-art policies, with up to 45% cost reduction.

Still following the idea that in large communication systems it is beneficial both for the users and for the network as a whole to store content closer to users, one particular implementation of such an approach is to co-locate caches with wireless base stations. In [5] K. Avrachenkov in collaboration with X. Bai and J. Goseling (both from Univ. of Twente, the Netherlands) study geographically distributed caching of a fixed collection of files. They model cache placement with the help of stochastic geometry and optimize the allocation of storage capacity among files in order to minimize the cache miss probability. They consider both per cache capacity constraints as well as an average capacity constraint over all caches. The case of per cache capacity constraints can be efficiently solved using dynamic programming, whereas the case of the average constraint leads to a convex optimization problem. The authors demonstrate that the average constraint leads to significantly smaller cache miss probability. Finally, they suggest a simple LRU-based policy for geographically distributed caching and show that its performance is close to the optimal.

In [7] K. Avrachenkov in collaboration with J. Goseling and B. Serbetci (both from Univ. of Twente, the Netherlands) consider caching in cellular networks in which each base station is equipped with a cache that can store a limited number of files. The popularity of the files is known and the goal is to place files in the caches such that the probability that a user at an arbitrary location in the plane will find the file that she requires in one of the covering caches is maximized. They develop distributed asynchronous algorithms for deciding which contents to store in which cache. Such cooperative algorithms require communication only between caches with overlapping coverage areas and can operate in asynchronous manner. The development of the algorithms is principally based on an observation that the problem can be viewed as a potential game. Their basic algorithm is derived from the best response dynamics. The authors demonstrate that the complexity of each best response step is independent of the number of files, linear in the cache capacity and linear in the maximum number of base stations that cover a certain area. Then, they show that the overall algorithm complexity for a discrete cache placement is polynomial in both network size and catalog size. In practical examples, the algorithm converges in just a few iterations. Also, in most cases of interest, the basic algorithm finds the best Nash equilibrium corresponding to the global optimum. Two extensions of the basic algorithm areprovided, based on stochastic and deterministic simulated annealing which find the global optimum. Finally,

the authors demonstrate the hit probability evolution on real and synthetic networks numerically and show that their distributed caching algorithm performs significantly better than storing the most popular content, probabilistic content placement policy and Multi-LRU caching policies.

7.5.2. Software Defined Networks (SDN)

The performance of computer networks relies on how bandwidth is shared among different flows. Fair resource allocation is a challenging problem particularly when the flows evolve over time. To address this issue, bandwidth sharing techniques that quickly react to the traffic fluctuations are of interest, especially in large scale settings with hundreds of nodes and thousands of flows. In [17] Z. Allybokus and K. Avrachenkov in collaboration with J. Leguay and L. Maggi (both from Huawei Research, Paris) propose a distributed algorithm that tackles the fair resource allocation problem in a distributed SDN control architecture. Their algorithm continuously generates a sequence of resource allocation solutions converging to the fair allocation while always remaining feasible, a property that standard primal-dual decomposition methods often lack. Thanks to the distribution of all computer intensive operations, they demonstrate that they can handle large instances in real-time.

In [18] K. Avrachenkov in collaboration with V. Borkar and S. Pattathil (both from IIT Bombay, India) consider the Generalized Additive Increase Multiplicative Decrease (G-AIMD) dynamics for resource allocation with alpha fairness utility function. This dynamics has a number of important applications such as internet congestion control, charging electric vehicles, and smart grids. They prove indexability for the special case of MIMD model and provide an efficient scheme to compute the index. The use of index policy allows to avoid the curse of dimensionality. They also demonstrate through simulations for another special case, AIMD, that the index policy is close to optimal and significantly outperforms a natural heuristic which penalizes the strongest user.

7.5.3. Network formation games

The paper [15] deals with a network formation game while balancing multiple, possibly conflicting objectives like cost, performance, and resiliency to viruses. It is part of a collaboration between Inria (E. Altman), Delft Univ. (S. Trajanovski, F. Kuipers, P. van Mieghem) and UAPV (Y. Hayel) which started within the CONGAS European project. Each player (node) aims to minimize its cost in installing links, the probability of being infected by a virus and the sum of hop counts on its shortest paths to all other nodes. In this article the authors (1) determine the Nash Equilibria and the Price of Anarchy for the network formation game, (2) demonstrate that the Price of Anarchy (PoA) is usually low, which suggests that (near-)optimal topologies can be formed in a decentralized way, and (3) give suggestions for practitioners for those cases where the PoA is high and some centralized control/incentives are advisable.

7.5.4. User association in LTE

Within the Inria-Nokia joint labs, C.S. Chen and L. Roullet (Nokia) N. Trabelsi (former member of MAESTRO) and E. Altman, and R. El-Azouzi (UAPV) have proposed a distributed algorithm for optimizing user Association and resource allocation in LTE networks. The solution is based on a game theoretic approach, which permits to compute Cell Individual Offset (CIO) and a pattern of power transmission over frequency and time domain for each cell. Simulation results show significant benefits in the average throughput and also cell edge user throughput of 40% and 55% gains respectively. Furthermore, we also obtain a meaningful improvement in energy efficiency.

7.5.5. Matching games for solving the association problem in WIFI

Matching games are a powerful framework for formulating and for solving user association problems. In [13], M. Touati (Orange Labs) and M. Coupechoux (Telecom ParisTech), R. El-Azouzi (UAPV), E. Altman and J. M. Kelif (Orange Labs) have considered the problem of association in a particular complex context of matching games with externalities in which the ranking of various associations by a player depends on association decisions of other player. This situation occurs in multi-rate IEEE 802.11 WLANs. traditional user association based on the strongest received signal and the well known anomaly of the MAC protocol can lead

to overloaded Access Points (APs), and poor or heterogeneous performance. They show that theit proposed association scheme can greatly improve the efficiency of 802.11 with heterogeneous nodes. The mechanism can be implemented as a virtual connectivity management layer to achieve efficient APs-user associations without modification of the MAC layer.

7.5.6. A stochastic game for competition over relay opportunities in DTN networks

In [12], K. P. Naveen (Indian Institute of Technology, Madras) and E. Altman in collaboration with A. Kumar (IISc Bangalore) consider an opportunistic wireless communication setting, in which two nodes (referred to as forwarders) compete to choose a relay node from a set of relays, as they ephemerally become available (e.g., wake up from a sleep state). Each relay, when it becomes available (or arrives), offers a (possibly different) " reward " to each forwarder. Each forwarder's objective is to minimize a combination of the delay incurred in choosing a relay and the reward offered by the chosen relay. As an example, the authors develop the reward structure for the specific problem of geographical forwarding over a common set of sleep-wake cycling relays. They formulate the model as a stochastic game theoretic variant of the asset selling problem studied in the Operations Research literature. They study two variants of the generic relay selection problem, namely, the completely observable and the partially observable cases. These cases are based on whether a forwarder (in addition to observing its reward) can also observe the reward offered to the other forwarder. The structure of Nash Equilibrium Policy Pairs is studied and characterized.

7.5.7. Aid for visually inpaired persons

S. Boularouk, D. Josselin (UAPV) and E. Altman pursue in [34], [35] the design of a geographic recommendation and alarm system for visually impaired persons. In [34] they propose a vocal ontology of Open-StreetMap (OSM) data for the apprehension of space by visually impaired people. They propose a simple but usable method to extract data from OSM databases in order to send them using Text To Speech technology. They focus on how to help people suffering from visual disability to plan their itinerary, to comprehend a map by querying computer and getting information about surrounding environment in a mono-modal human-computer dialogue. In [35] they further study the benefit of IoT for people with disabilities, particularly for visually impaired and blind people mobility. They propose a simple prototype using OpenStreetMap data combined to physical environment data measured from sensors connected to a Arduino board through Speech recognition.

7.5.8. Routing games over the line

In [28] A. Karouit, M. Haddad, A. El Matar (UAPV) and E. Altman study a sequential routing game where several users send traffic to a destination on a line. Each user arrives at some time epoch with a given capacity. Then, he ships its demand over time on a shared resource. The state of a player evolves according to whether he decides to transmit or not. The decision of each user is thus spatio-temporal control. The authors provide an explicit expression for the equilibrium of such systems and compare it to the global optimum case. In particular, they compute the price of anarchy of such schemes and identify a Braess-type paradox in the context of sequential routing games.

7.5.9. Multicriteria Games of congestion

In [23], A. Boukoftane and M. Haddad (UAPV) in collaboration with E. Altman and N. Oukid (Univ. de Saad Dahlab) consider a routing game in a network that contains lossy links. They consider a multi-objective problem where the players have each a weighted sum of a delay cost and a cost for losses. They compute the equilibrium and optimal solution (which are unique). They discover here in addition to the classical Kameda type paradox another paradoxical behavior in which higher loss rates have a positive impact on delay and therefore higher quality links may cause a worse performance even in the case of a single player.

7.5.10. Speed estimation in cellular networks

The paper [25], is part of a joint ongoing work within the Inria-Nokia joint lab on the SelfNet ADR which focused on speed estimation. It involved E. Altman, M. Haddad (UAPV), D.G. Herculea, C.S. Chen and V. Capdevielle (Nokia). The authors provide a new online algorithm for mobile user speed estimation in 3GPP Long Term Evolution (LTE)/LTE-Advanced networks. The proposed method leverages on uplink sounding

reference signal power measurements performed at the base station, also known as eNodeB, and remains effective even under large sampling period. Extensive performance evaluation of the proposed algorithm is carried out using field traces from realistic environment. The on-line solution is proven highly efficient in terms of computational requirement, estimation delay, and accuracy.

7.6. Applications in Social Networks

Participant: Eitan Altman.

7.6.1. Posting behavior

In [10], Eitan Altman, together with A. Masson (SAFRAN Group, formerly with MAESTRO) and Y. Hayel (UAPV), pursue two objectives. First they model the posting behaviour of publishers in Social Networks which have externalities. Secondly, they propose content active filtering in order to increase content diversity from different publishers. By externalities, is meant that when the quantity of posted contents from a specific publisher impacts the popularity of other posted contents. The authors introduce a dynamical model to describe the posting behaviour of publishers taking into account these externalities. This model is based on stochastic approximations and sufficient conditions are provided to ensure its convergence to a unique rest point. A closed form of this rest point is provided, and it is shown that it can be obtained as the unique equilibrium of a non-cooperative game. Content Active Filtering (CAF) are actions taken by the administrator of the Social Network in order to promote some objectives related to the quantity of contents posted in various contents. An objective of the CAF can be maximizing the diversity of posted contents. Finally, the authors illustrate their results through numerical simulations and they validate them with real data extracted from social networks.

7.7. Applications to Renewable Resources and Energy

Participants: Sara Alouf, Alain Jean-Marie, Dimitra Politaki.

7.7.1. Stochastic models for solar power

In [31], D. Politaki and S. Alouf develop a stochastic model for the solar power at the surface of the earth. They combine a deterministic model of the clear sky irradiance with a stochastic model for the so-called clear sky index to obtain a stochastic model for the actual irradiance hitting the surface of the earth. Their clear sky index model is a 4-state semi-Markov process where state durations and clear sky index values in each state have phase-type distributions. They use per-minute solar irradiance data to tune the model, hence they are able to capture small time scales fluctuations. They compare this model with the on-off power source model developed by Miozzo et al. (2014) for the power generated by photovoltaic panels, and to a modified version that they propose. Computing the autocorrelation functions for all proposed models, they find that the irradiance model surpasses the on-off models and it is able to capture the multiscale correlations that are inherently present in the solar irradiance. The power spectrum density of generated trajectories matches closely that of measurements. This new irradiance model can be used not only in the mathematical analysis of energy harvesting systems but also in their simulation.

In [45], D. Politaki, S. Alouf and A. Jean-Marie in collaboration with F. Hermenier (Nutanix) aim at the performance analysis of a data center feb by renewable energy recources. They describe the data center system, proposing a new queuing model BMAP/PH/c which represents the queue length in a system having c servers, where arrivals are determined by a Batch Markov Arrival process and service times have a phase-type distribution. They validate this model using real traces. Next, they characterize the data center google workload traces which are available in the web and they validate that the jobs arrive to the system in groups (batches) and wait at the queue. The waiting time is diverse according to the available resources, job size etc. The authors then compute the empirical CDF of the service time and try to fit it with well-known distributions like exponential, Pareto etc. However, the Kolmogorov-Smirnov test rejects the null hypothesis at the 1% significance level which shows that service time doesn't fit with any well-known distribution.

7.7.2. Sustainable management of water consumption

Alain Jean-Marie, Mabel Tidball (INRA, Montpellier, France), Fernando Ordóñez and Victor Bucarey López (Univ. de Chile, Chile), consider in [36] a discrete time, infinite horizon dynamic game of groundwater extraction. A Water Agency charges an extraction cost to water users, and controls the marginal extraction cost so that it depends linearly on total water extraction (through a parameter n) and on rainfall (through parameter m). The water users are selfish and myopic, and the goal of the agency is to give them incentives them so as to, at the same time, improve their total welfare and improve the long-term level of the resource.

This problem is studied in two situations for a linear-quadratic model. In the first situation, the parameters n and m are considered to be fixed over time, and the Agency selects the value that maximizes the total discounted welfare of agents. A first result shows that when the Water Agency is patient (discount rate close to one), the optimal marginal extraction cost asks for strategic interactions between agents.

In the second situation, the authors look at the dynamic Stackelberg game where the Agency decides at each time what cost parameter they must announce in order to maximize the welfare function. This becomes a highly non-linear optimal control problem. Some preliminary results are presented.

RAP2 Team

4. New Results

4.1. Resource Allocation in Large Data Centres

Participants: Christine Fricker, Philippe Robert, Guilherme Thompson, Veronica Quintuna Rodriguez.

With the emergence of new networking paradigms such as Cloud Computing and related technologies (Fog Computing, VNF, etc.) new challenges in understanding, modelling and improving systems relying on these technologies arise. Our research goal is to understand how the stochastic nature of the access to these systems affects their performance, and to design algorithms which can improve global performance using local information. This research is made in collaboration with Fabrice Guillemin, from Orange Labs.

Building up from the results previously obtained by this team, we have extend our research towards more complex systems, investigating the behaviour of multi-resource systems, which are globally stable but local congested, a problem that naturally arises from the decentralization of resources. We investigate a cooperation scheme between processing facilities, where congestion-maker clients, the one with the largest demand the locally congested resource are systematically forwarded to the another data centre when some threshold on the occupation level is reached. These thresholds are chosen to anticipate sufficiently in advance potential shortages of any resource in any data centre. After providing some convergence results, we are able to express the performance of the system in terms of the invariant distribution of an inhomogeneous random walk on the plane. We derive optimal threshold parameters, improving the performance of the distributed Cloud Computing system in such a way that it approaches the efficiency of a centralised system. Currently, a document is being prepared for publication, but the main results are presented in G. Thompson's PhD Document [2].

4.2. Ressource allocation in vehicle sharing systems

Participants: Christine Fricker, Yousra Chabchoub.

Vehicle sharing systems are becoming an urban mode of transportation, and launched in many cities, as Velib' and Autolib' in Paris. Managing such systems is quite difficult. One of the major issues is the availability of the resources: vehicles or free slots to return them. These systems became a hot topic in Operation Research and the importance of stochasticity on the system behavior leads us to propose mathematical stochastic models. The problem is to understand the system behavior and how to manage these systems in order to improve the allocation of both resources to users. This work is in collaboration with El Sibai Rayane (ISEP), Plinio Santini Dester (École Polytechnique), Hanène Mohamed (Université Paris-Ouest), and Danielle Tibi (Université Paris Diderot).

4.2.1. Stochastic modelling of bike-sharing systems

The goal is to derive the stationary behavior of the state process in a quite general model: number of bikes in the stations and in routes between two stations. Our stochastic model is the first one taking into account the finite number of spots at the stations. The basic model for bike-sharing systems comes within the framework of closed networks with two types of nodes: single server/finite capacity nodes and infinite servers/infinite capacity nodes. The effect of local saturation is modeled by generalized blocking and rerouting procedures, under which, as a key argument, the stationary state is proved to have product-form. For a class of large closed Jackson networks submitted to capacity constraints, asymptotic independence of the nodes in normal traffic phase is proved at stationarity under mild assumptions, using a Local Limit Theorem. The limiting distributions of the queues are explicit. In the Statistical Mechanics terminology, the equivalence of ensembles - canonical and grand canonical - is proved for specific marginals. This widely extends the existing results on heterogeneous bike-sharing systems. The grand canonical approximation can then be used for adjusting the total number of bikes and the capacities of the stations to the expected demand. [12]

4.2.2. Local load balancing policies.

Recently we investigated some load balancing algorithms for stochastic networks to improve the bike sharing system behavior. We focus on the choice of the least loaded station among two to return the bike, the so called Power of choice. Nevertheless, in real systems, this choice is local. Thus the main challenge is to deal with the choice between two neighboring stations.

For that, a set of N queues, with a local choice policy, is studied. When a customer arrives at queue i, he joins the least loaded queue between queues i and $i+1$. When the load tends to zero, we obtain an asymptotic for the stationary distribution of the number of customers at a queue. The main result is that, in equilibrium, queue lengths decay geometrically when ρ tends to 0, N fixed. It allows to compare local choice, no choice and *Power of choice*. The local policy changes the exponential decay with respect to no choice but does not lead to an improvement (double exponential tail decay) comparable to the random choice model. [19].

For a bike-sharing homogeneous model, we study a deterministic cooperation between the stations, two by two. Analytic results are achieved in an homogeneous bike-sharing model. They concern the mean-field limit as the system is large, and its equilibrium point. Results on performance mainly involve an original closed form expression of the stationary blocking probability and new tight bounds for the mean of the total number of customers in the classical join-the-shortest-queue model. These results are compared by simulations with the policy where the users choose the least loaded between two neighboring stations. It turns out that, because of randomness, the choice between two neighbours gives better performance than grouping stations two by two.

It relies on new results for the classical system of two queues under the join-the-shortest-queue policy. We revisited the study of the stationary distribution. A simple analytical solution is proposed. Using standard generating function arguments, a simple expression of the blocking probability is derived, which as far as we know is original. Furthermore, from the balance equations, all stationary probabilities are obtained as explicit combinations of those of states $(0,k)$ for $0 \leq k \leq K$. The blocking probability is also obtained for a variant with two queues under JSQ, where the constraint is on the total capacity of the system.

This extends to the infinite capacity and asymmetric cases, i.e., when the queues have different service rates. For the initial symmetric finite capacity model, the stationary probabilities of states $(0,k)$ can be obtained recursively from the blocking probability. In the other cases, they are implicitly determined through some functional equation that characterizes their generating function. For the infinite capacity symmetric model, we provide an elementary proof of a result by Cohen which gives the solution of the functional equation in terms of an infinite product with explicit zeroes and poles. See [9].

We use data, trip data (trips collected in a month) obtained from JCDecaux and reports on station status collected as open data, to test local choice policy. Indeed we designed and tested a new method that globally improves the distribution of the resources (bikes and docks) among the stations. It relies on a local small change in user behaviors, by adapting their trips to resource availability around their departure and arrival stations. Results show that, even with a partial user collaboration, the proposed method increases significantly the global balance of the bike sharing system and therefore the user satisfaction. This is done using trip data sets. The key of our study is to detect spatial outliers, objects having a behavior significantly different from their spatial neighbors, in a context where neighbors are heavily correlated. Moran scatterplot is a well-known method that exploits similarity between neighbors in order to detect spatial outliers. We proposed an improved version of Moran scatterplot, using a robust distance metric called Gower similarity. Using this new version of Moran scatterplot, we identified many spatial outliers stations (often with much more available bikes, or with much more empty docks during the day) in Velib. For the occupancy data set obtained by modifiying trips, the number of spatial outliers drastically decreases. See [18].

4.3. Scaling Methods

Participants: Davit Martirosyan, Philippe Robert, Wen Sun.

4.3.1. Large Unreliable Stochastic Networks

The reliability of a large distributed system is studied. The framework is a system where files are stored on servers. When one of these servers breaks down, all files on it are lost. We assume that these files could be retrieved immediately and re-allocated among other servers while the failed server restarts but empty. It is a reasonable assumption since the failure rate is quit small comparing to an effective recovery mechanism. It is also assumed that each server is connected with a subset of servers in the system. When it breaks down, files on it are re-allocated on the servers that in this subset, following a given policy. Our main interest is the influence on the loads due to two allocation algorithms: the " Random Choice" (RC) policy and the "Power of d Choices" (PoC) policy.

- (RC) Each copy join a server in the subset at random.
- (PoC) Each copy chooses d servers in the subset at random, and joins the least loaded one.

The asymptotic behaviors of these two policies are investigated through mean field models. We have show that when the number of servers getting large, the load of each server can be approached by a linear (resp. non-linear) Markov process for RC (resp. PoC) policy. The equilibrium distributions of these asymptotic processes are also given.

For the case $d = 2$ and all the servers are connected, see the paper [15]. This is a joint work with Inria/UPMC Team Regal. For a generalized case, there is a paper in preparation.

4.3.2. Bandwidth Allocation in Large Data Center

We are investigating a problem of efficient resource allocation in a large data center. In our model, the following is assumed. Each job that should be treated arrives to an M/M/C queue and is placed in it if the latter is not exhausted. Otherwise, it is sent to another queue for the possible implementation with the help of a certain canal, whose size is finite. A mean-field or the so called chaoticity result is established. Informally speaking, we show that the stochastic process that describes the evolution of our system converges to a non-random limit. We then study the stability properties of this limiting process and prove that it has a unique equilibrium that attracts exponentially all solutions that are issued from its small neighborhood. Moreover, we also show that if the size of the canal is infinite (i.e., the jobs go freely to another queue when not served), the uniqueness for the fixed point problem is not guaranteed and, depending on some physical parameters, one can have no solution, a unique solution or two solutions. This phenomenon is quite surprising and it seems that is was not observed before. We also investigate the stability of equilibrium points. Some techniques used in our proofs come from theories developed in the context of PDEs.

4.4. Stochastic Models of Biological Networks

Participants: Renaud Dessalles, Philippe Robert, Wen Sun.

4.4.1. Stochastic Modelling of self-regulation in the protein production system of bacteria.

This is a collaboration with Vincent Fromion from INRA Jouy-en-Josas, which started in December 2013.

In prokaryotic cells (e.g. E. Coli. or B. Subtilis) the protein production system has to produce in a cell cycle (i.e. less than one hour) more than 10^6 molecules of more than 2500 kinds, each having different level of expression. The bacteria uses more than 67% of its resources to the protein production. Gene expression is a highly stochastic process: bacteria sharing the same genome, in a same environment will not produce exactly the same amount of a given protein. Some of this stochasticity can be due to the system of production itself: molecules, that take part in the production process, move freely into the cytoplasm and therefore reach any target in the cell after some random time; some of them are present in so much limited amount that none of them can be available for a certain time; the gene can be deactivated by repressors for a certain time, etc. We study the integration of several mechanisms of regulation and their performances in terms of variance and distribution. As all molecules tends to move freely into the cytoplasm, it is assumed that the encounter time between a given entity and its target is exponentially distributed.

4.4.1.1. Models with Cell Cycle

Usually, classical models of protein production do not explicitly represent several aspects of the cell cycle: the volume variations, the division and the gene replication. Yet these aspects have been proposed in literature to impact the protein production. We have therefore proposed a series of "gene-centered" models (that concentrates on the production of only one type of protein) that integrates successively all the aspects of the cell cycle. The goal is to obtain a realistic representation of the expression of one particular gene during the cell cycle. When it was possible, we analytically determined the mean and the variance of the protein concentration using Marked Poisson Point Process framework.

We based our analysis on a simple model where the volume changes across the cell cycle, and where only the mechanisms of protein production (transcription and translation) are represented. The variability predicted by this model is usually assimilated to the "intrinsic noise" (i.e. directly due to the protein production mechanism itself). We then add the random segregation of compounds at division to see its effect on protein variability: at division, every mRNA and every protein has an equal chance to go to either of the two daughter cells. It appears that this division sampling of compounds can add a significant variability to protein concentration. This effect directly depends on the relative variance (Fano factor) of the protein concentration: this effect is stronger as the relative variance is low. The dependence on the relative variance can be explained by considering a simplified model. With parameters deduced from real experimental measures, we estimate that the random segregation of compounds can double the variability of the genes with the lowest relative variance.

Finally, we integrate the gene replication to the model: at some point in the cell cycle, the gene is replicated, hence doubling the transcription rate. We are able to give analytical expressions for the mean and the variance of protein concentration at any moment of the cell cycle; it allows to directly compare the variance with the previous model with division. We show that gene replication has little impact on the protein variability: an environmental state decomposition shows that the part of the variance due to gene replication represents only at most 2% of the total variability predicted by the model.

Finally, we have investigated other possible sources of variability by presenting other simulations that integrate some specific aspects: variability in the production of RNA-polymerases and ribosomes, uncertainty in the division and DNA replication decisions, etc. None of the considered aspects seems to have a significant impact on the protein variability.

In the end, these results are compared to the real experimental measure of protein variability. It appears that the models with cell cycle presented above tend to underestimate the protein variability especially for highly expressed proteins. See Dessalles [1] and Dessalles et al. [17]

4.4.2. Stochastic Modelling of Protein Polymerization

This is a collaboration with Marie Doumic, Inria MAMBA team. The first part of our work focuses on the study of the polymerization of protein. This phenomenon is involved in many neurodegenerative diseases such as Alzheimer's and Prion diseases, e.g. mad cow. In this context, it consists in the abnormal aggregation of proteins. Curves obtained by measuring the quantity of polymers formed in in vitro experiments are sigmoids: a long lag phase with almost no polymers followed by a fast consumption of all monomers. Furthermore, repeating the experiment under the same initial conditions leads to somewhat identical curves up to translation. After having proposed a simple model to explain this fluctuations, we studied a more sophisticated model, closer to the reality. We added a conformation step: before being able to polymerize, proteins have to misfold. This step is very quick and remains at equilibrium during the whole process. Nevertheless, this equilibrium depends on the polymerization which is happening on a slower time scale. The analysis of these models involves stochastic averaging principles.

We have also investigated a more detailed model of polymerisation by considering the the evolution of the number of polymers with different sizes $(X_i(t))$ where $X_i(t)$ is the number of polymers of size i at time t. By assuming that the transitions rates are scaled by a large parameter N, it has been shown that, in the limit, the process $(X_i^N(t))$ is converging to the solution of Becker-Döring equations as N goes to infinity. For another model including nucleation, we have given an asymptotic description of the lag time at the first and second order. These results are obtained in particular by proving stochastic averaging theorems.

4.4.3. Central Limit Theorems

We have investigate the fluctuations of the stochastic Becker-Döring model of polymerization when the initial size of the system converges to infinity. A functional central limit problem is proved for the vector of the number of polymers of a given size. It is shown that the stochastic process associated to fluctuations is converging to the strong solution of an infinite dimensional stochastic differential equation (SDE) in a Hilbert space. We haveproved that, at equilibrium, the solution of this SDE is a Gaussian process. The proofs are based on a specific representation of the evolution equations, the introduction of a convenient Hilbert space and several technical estimates to control the fluctuations, especially of the first coordinate which interacts with all components of the infinite dimensional vector representing the state of the process. See Sun [21]

4.4.4. Study of the Nucleation Phenomenon

We have investigated a new stochastic model describing the time evolution of a polymerization process. The initial state of the system consists only of isolated monomers. We study the *lag time* of the polymerization process, that is, the first instant when a fraction of the initial monomers is polymerized, i.e. the fraction of monomers used in the polymers. The mathematical model includes *a nucleation property*: polymers with a size below some threshold n_c, the size of the nucleus, are quickly fragmented into smaller polymers. For a size greater than n_c, the fragmentation still occurs but at a smaller rate. A scaling approach is used, by taking the volume N of the system as a scaling parameter. If $n_c \geq 3$, under quite general assumptions on the way polymers are fragmented, we prove a limit theorem for the instant T^N of creation of the first "stable" polymer, i.e. a polymer of size n_c. It is proved that the distribution of T^N/N^{n_c-3} converges to an exponential distribution. We also show that, if $n_c \geq 4$, then the lag time has the same order of magnitude as T^N and, if $n_c = 3$, it is of the order of $\log N$. An original feature of our model is the significant variability (asymptotic exponential distribution) proved for the instants associated to polymerization. This is a well known phenomenon observed in the experiments in biology but it has not been really proved in appropriate mathematical models up to now. The results are proved via a series of (quite) delicate technical estimates for occupations measures on fast time scales associated to the first n_c coordinates of the corresponding Markov process. Extensive Stochastic calculus with Poisson processes, several coupling arguments and classical results from continuous branching processes theory are the main ingredients of the proofs.

SOCRATE Project-Team

6. New Results

6.1. Flexible Radio Front-End

6.1.1. RFID tag-to-tag communication

RFID is a well-known technique for wireless authentication. Usually, such a system only consists on a reader communicating with one or several tags. The concept of passive RFID tag-to-tag communications has been recently introduced and opens new promising perspectives, especially in the field of Internet-of-Things. In this work, a simulation framework was proposed as a new tool allowing the performance evaluation of tag-to-tag radio links. The modeling takes into consideration the external source supplying the communication between tags, radiating characteristics of tag antennas, and reception system aspects. Performance results are expressed in terms of Bit Error Rate (BER) with respect to the distance between the tags and the position of the energy source relative to the position of the two tags [36], [35].

6.1.2. Optimization of waveforms for energy harvesting

We have studied the incidence of the modulation scheme as well as the input power on the RF to DC rectifier conversion efficiency for an energy harvesting system based on radiowaves. A commercial energy harvesting (EH) P21XXCSR-EVB evaluation board from Powercast Corporation is used as measurement target and several waveforms are employed to evaluate the rectification efficiency. With a continuous wave as reference, QPSK, QAM and OFDM waveforms usage demonstrates that digital modulated signals can lead to a better efficiency. Thus, by selecting a high peak to average power ratio (PAPR), and under certain conditions, the performance of the energy harvesting circuit is enhanced [30].

6.2. Multi-User Communications

6.2.1. Fundamental limits : contributions in Multi-User Information Theory (MU-IT)

6.2.1.1. Interference channel with feedback

In this work [29], [44], [43], the $\eta-$Nash equilibrium ($\eta-$NE) region of the two-user linear deterministic interference channel (IC) with noisy channel-output feedback is characterized for all $\eta > 0$. The $\eta-$NE region, a subset of the capacity region, contains the set of all achievable information rate pairs that are stable in the sense of an $\eta-$NE. More specifically, given an $\eta-$NE coding scheme, there does not exist an alternative coding scheme for either transmitter-receiver pair that increases the individual rate by more than η bits per channel use. Existing results such as the η-NE region of the linear deterministic IC without feedback and with perfect output feedback are obtained as particular cases of the result. We also characterized in [15] the price of anarchy (PoA) and the price of stability (PoS) of this $\eta-$NE. The price of anarchy is the ratio between the sum-rate capacity and the smallest sum-rate at an $\eta-$NE. The price of stability is the ratio between the sum-rate capacity and the biggest sum-rate at an $\eta-$NE. Some of the main conclusions of this work are the following: (a) When both transmitter-receiver pairs are in low interference regime, the PoA can be made arbitrarily close to one as η approaches zero, subject to a particular condition. More specifically, there are scenarios in which even the worst $\eta-$NE (in terms of sum-rate) is arbitrarily close to the Pareto boundary of the capacity region. (b) The use of feedback plays a fundamental role on increasing the PoA, in some interference regimes. This is basically because in these regimes, the use of feedback increases the sum-capacity, whereas the smallest sum-rate at an $\eta-$NE remains the same. (c) The PoS is equal to one in all interference regimes. This implies that there always exists an $\eta-$NE in the Pareto boundary of the capacity region. The ensemble of conclusions of this work reveal the relevance of jointly using equilibrium selection methods and channel-output feedback for reducing the effect of anarchical behavior of the network components in the $\eta-$NE sum-rate of the interference channel.

6.2.1.2. Simultaneous information and energy transmission

In this work [42], [25], [48], the fundamental limits of simultaneous information and energy transmission in the two-user Gaussian interference channel (G-IC) with and without feedback are fully characterized. More specifically, an achievable and converse region in terms of information and energy transmission rates (in bits per channel use and energy-units per channel use, respectively) are identified. In both cases, with and without feedback, an achievability scheme based on power-splitting, common randomness, rate splitting, block-Markov superposition coding, and backward decoding is presented. Finally, converse regions for both cases are obtained using some of the existing outer bounds for information transmission rates, as well as a new outer bound for the energy transmission rate.

6.2.1.3. Ultra-dense wireless networks

Ultra-dense networks represent an interesting model for future IoT networks. The analysis of these metworks relies on the association of stochastic geometry models with information theory in the finite blocklength regime. Considering an isolated wireless cell containing a high density of nodes, the fundamental limit can be defined as the maximal number of nodes the associate base station can serve under some system level constraints including maximal rate, reliability, latency and transmission power. This limit can be investigated in the downlink, modeled as a spatial continuum broadast channel (SCBC) as well as in the uplink modeled as a spatial continuum multiple access channel (SCMAC). In this work, we define the different steps towards the characterization of this fundamental limit, considering four figures of merit: energy efficiency, spectral efficiency, latency, reliabilty [13]. To address this question in the uplink scenario [10], we uses a large scale Multiple Access Channel (MAC) to model IoT nodes randomly distributed over the coverage area of a unique base station. The traffic is represented by an information rate spatial density $\rho(x)$. This model, referred to as the Spatial Continuum Multiple Access Channel, is defined as the asymptotic limit of a sequence of discrete MACs. The access capacity region of this channel is defined as the set of achievable information rate spatial densities achievable with vanishing transmission errors and under a sum-power constraint. Simulation results validate the model and show that this fundamental limit theoretically achievable when all nodes transmit simultaneously over an infinite time, may be reached even with a relatively small number of simultaneous transmitters (typically around 20 nodes) which gives credibility to the model. The results also highlight the potential interest of non-orthogonal transmissions for IoT uplink transmissions when compared to an ideal time sharing strategy. We then developed a powerful analytical model of wireless network with Superposition Coding (SC), also referred to as Non Orthogonal Multiple Access (NOMA), taking into consideration a multi cell interference limited network. This model allows to establish a closed form expression of the minimum power a base station (BS) needs to transmit to its users and to achieve a given SINR (signal to interference plus noise ratio) whatever its location in the area covered by the base station. It moreover allows to establish a closed form expression of the minimum total transmit power of a base station. These closed form expressions allow to establish performance of wireless networks, by minimizing the base stations transmit powers. As an application, we show that these closed form expressions allow to quantify the energetic performance, spectral efficiency, total throughput and the coverage of a BS, in a simple and quick way.

6.2.1.4. Broadcast channel in the Finite Blocklength regime

In order to analyse wireless networks with short packets, theoretical results in information theory for the finite blocklength regime in multi-user scenarios were missin. In [34], [33], we analyzed the performance of superposition coding for Gaussian broadcast channels with finite blocklength. To this end, we adapted two different achievability bounds, the dependence testing and the $\kappa - \beta-$bounds introduced by Polyanskiy et al. in 2010 to the broadcast setting. The distinction between these bounds lies in fixing either the input or the output distributions of the channel. For the first case of the dependence testing bound, an upper bound on the average error probability of the system is derived whereas for the latter, lower bounds on the maximal code sizes of each user are presented.

6.2.1.5. Capacity sensitivity

In this work [22], [40], a new framework based on the notion of *capacity sensitivity* is introduced to study the capacity of continuous memoryless point-to-point channels. The capacity sensitivity reflects how the capacity changes with small perturbations in any of the parameters describing the channel, even when the capacity is

not available in closed-form. This includes perturbations of the cost constraints on the input distribution as well as on the channel distribution. The framework is based on continuity of the capacity, which is shown for a class of perturbations in the cost constraint and the channel distribution. The continuity then forms the foundation for obtaining bounds on the capacity sensitivity. As an illustration, the capacity sensitivity bound is applied to obtain scaling laws when the support of additive $\alpha-$stable noise is truncated.

6.2.2. Performance evaluation of large scale systems

6.2.2.1. UNB networks performance evaluation

UNB (Ultra Narrow Band) stands out as one promising PHY solution for low-power, low-throughput and long-range IoT. The dedicated MAC scheme is RFTMA (Random Frequency and Time Multiple Access), where nodes access the channel randomly both in frequency and in time domain, without prior channel sensing. This blind randomness sometimes introduces interference and packet losses. In order to quantify the system performance, we have derived and exploited a theoretical expression of the outage probability in a UNB based IoT network, when taking into account both interference due to the spectral randomness and path loss due to the propagation [14], [5]. Besides, we also proposed to use the well-known SIC (Successive Interference Cancellation) to cancel the interference in a recursive way. We provided a theoretical analysis of network performance, when considering jointly SIC and the specific spectral randomness of UNB. We analytically and numerically highlighted the SIC efficiency in enhancing UNB system performance [26].

6.2.2.2. Wireless networks on FIT/CorteXlab

In this work we study the FIT/CorteXlab platform where all radio nodes are confined to an electromagnetically (EM) shielded environment and have flexible radio-frequency (RF) front-end for experimenting on software defined radio (SDR) and cognitive radio (CR). A unique feature of this testbed is that it offers roughly 40 SDR nodes that can be accessed from anywhere in the world in a reproducible manner: the electromagnetic shield prevents from external interference and channel variability. In this work [16] we show why it is important to have such a reproducible radio experiment testbed and we highlight the reproducibility by the channel characteristics between the nodes of the platform. We back our claims with a large set of measurements done in the testbed, that also refines our knowledge on the propagation characteristics of the testbed.

One of the major goals of the 5G technology roadmap is to create disruptive innovation for the efficient use of the radio spectrum to enable rapid access to bandwidth-intensive multimedia services over wireless networks. The biggest challenge toward this goal lies in the difficulty in exploiting the multicast nature of the wireless channel in the presence of wireless users that rarely access the same content at the same time. Recently, the combined use of wireless edge caching and coded multicasting has been shown to be a promising approach to simultaneously serve multiple unicast demands via coded multicast transmissions, leading to order-of-magnitude bandwidth efficiency gains. However, a crucial open question is how these theoretically proven throughput gains translate in the context of a practical implementation that accounts for all the required coding and protocol overheads. In [3], in collaboration with Nokia Bell Labs, New Jersey, we first provide an overview of the emerging caching- aided coded multicast technique, including state-of-the-art schemes and their theoretical performance. We then focus on the most competitive scheme proposed to date and describe a fully working prototype implementation in CorteXlab, one of the few experimental facilities where wireless multiuser communication scenarios can be evaluated in a reproducible environment. We use our prototype implementation to evaluate the experimental performance of state-of-the-art caching-aided coded multicast schemes compared to state-of-the-art uncoded schemes, with special focus on the impact of coding computation and communication overhead on the overall bandwidth efficiency performance. Our experimental results show that coding overhead does not significantly affect the promising performance gains of coded multicasting in small-scale realworld scenarios, practically validating its potential to become a key next generation 5G technology.

6.2.3. Cognitive networks

6.2.3.1. Game theory based approaches

In [4], a generalization of the satisfaction equilibrium (SE) for games in satisfaction form (SF) is presented. This new solution concept is referred to as the generalized satisfaction equilibrium (GSE). In games in SF, players choose their actions to satisfy an individual constraint that depends on the actions of all the others. At a GSE, players that are unsatisfied are unable to unilaterally deviate to be satisfied. The concept of GSE generalizes the SE in the sense that it allows mixed-strategy equilibria in which there exist players who are unable to satisfy their individual constraints. The pure-strategy GSE problem is closely related to the constraint satisfaction problem and finding a pure-strategy GSE is proven to be NP-hard. The existence of at least one GSE in mixed strategies is proven for the class of games in which the constraints are defined by a lower limit on the expected utility. A dynamics referred to as the satisfaction response is shown to converge to a GSE in certain classes of games. Finally, Bayesian games in SF and the corresponding Bayesian GSE are introduced. These results provide a theoretical framework for studying service-level provisioning problems in communications networks as shown by several examples.

Device-to-device (D2D) communications can enhance spectrum and energy efficiency due to direct proximity communication and frequency reuse. However, such performance enhancement is limited by mutual interference and energy availability, especially when the deployment of D2D links is ultra-dense. In this contribution [9], we present a distributed power control method for ultra-dense D2D communications underlaying cellular communications. In this power control method, in addition to the remaining battery energy of the D2D transmitter, we consider the effects of both the interference caused by the generic D2D transmitter to others and interference from all others' caused to the generic D2D receiver. We formulate a mean-field game (MFG) theoretic framework with the interference mean-field approximation. We design the cost function combining both the performance of the D2D communication and cost for transmit power at the D2D transmitter. Within the MFG framework, we derive the related Hamilton-Jacobi-Bellman (HJB) and Fokker-Planck-Kolmogorov (FPK) equations. Then, a novel energy and interference aware power control policy is proposed, which is based on the Lax-Friedrichs scheme and the Lagrange relaxation. The numerical results are presented to demonstrate the spectrum and energy efficiency performances of our proposed approach. Index Terms—Device-to-device communication, mean field game, spectrum efficiency, energy efficiency.

6.2.3.2. Learning approaches

Fast initialization of cognitive radio systems is a key problem in a variety of wireless communication systems, particularly for public safety organizations in emergency crises. In the initialization problem, the goal is to rapidly identify an unoccupied frequency band. In this contribution [21], we formalize the initialization problem within the framework of active hypothesis testing. We characterize the optimal scanning policy in the case of at most one free band and show that the policy is computationally challenging. Motivated by this challenge for the implementation of the optimal policy and the need to cope with an unknown number of interferers larger than one, we propose the constrained DGF algorithm. We show that for strict constraints on the maximum number of observations, the constrained DGF algorithm can outperform the error probability of the state-of-the-art C-SPRT algorithm by an order of magnitude, for comparable average delays.

6.2.3.3. Asynchronous transmissions in VLC

In a visible light communications system (VLC), light sources are responsible for both illumination, communications and positioning. These light sources inevitably interfere each others at the receiver. To retain the appealing advantage that VLC systems can reuse existing lighting infrastructure, using an extra network to control or synchronize the light sources should be avoided. This work [31] proposes an uncoordinated multiple access scheme for VLC systems with positioning capability. The proposed scheme does not require a central unit to coordinate the transmission of the transmitters. Transmitters can be asynchronous with one another and with the receiver. Each transmitter is allocated a unique codeword with L chips for a system with up to $(L-1)/2$ transmitters where L is prime. Due to the linear growth in complexity with respect to number of transmitters, our proposed scheme is feasible for systems with large numbers of transmitters. Our novel decoder can minimize the effect of additive Gaussian noise at the receiver side. Simulation results show that the proposed decoder outperforms zero-forcing decoder.

6.2.4. Contributions in other application fields

6.2.4.1. Smart Grids

The advanced operation of future electricity distribution systems is likely to require significant observability of the different parameters of interest (e.g., demand, voltages, currents, etc.). Ensuring completeness of data is, therefore, paramount. In this context, an algorithm for recovering missing state variable observations in electricity distribution systems is presented in [47]. The proposed method exploits the low rank structure of the state variables via a matrix completion approach while incorporating prior knowledge in the form of second order statistics. Specifically , the recovery method combines nuclear norm minimization with Bayesian estimation. The performance of the new algorithm is compared to the information-theoretic limits and tested trough simulations using real data of an urban low voltage distribution system. The impact of the prior knowledge is analyzed when a mismatched covariance is used and for a Markovian sampling that introduces structure in the observation pattern. Numerical results demonstrate that the proposed algorithm is robust and outperforms existing state of the art algorithms.

In addition, Gaussian random attacks that jointly minimize the amount of information obtained by the operator from the grid and the probability of attack detection are presented in [38]. The construction of the attack is posed as an optimization problem with a utility function that captures two effects: firstly, minimizing the mutual information between the measurements and the state variables; secondly, minimizing the probability of attack detection via the Kullback-Leibler (KL) divergence between the distribution of the measurements with an attack and the distribution of the measurements without an attack. Additionally, a lower bound on the utility function achieved by the attacks constructed with imperfect knowledge of the second order statistics of the state variables is obtained. The performance of the attack construction using the sample covariance matrix of the state variables is numerically evaluated. The above results are tested in the IEEE $30-$Bus test system.

6.2.4.2. Molecular Communications

Molecular communications is emerging as a technique to support coordination in nanonetworking, particularly in biochemical systems. In complex biochemical systems such as in the human body, it is not always possible to view the molecular communication link in isolation as chemicals in the system may react with chemicals used for the purpose of communication. There are two consequences: either the performance of the molecular communication link is reduced; or the molecular link disrupts the function of the biochemical system. As such, it is important to establish conditions when the molecular communication link can coexist with a biochemical system. In this work [45], we develop a framework to establish coexistence conditions based on the theory of chemical reaction networks. We then specialize our framework in two settings: an enzyme-aided molecular communication system; and a low-rate molecular communication system near a general biochemical system. In each case, we prove sufficient conditions to ensure coexistence.

6.3. Software Radio Programming Model

6.3.1. Dataflow programming models

Parallel computers have become ubiquitous and current processors contain several execution cores. A variety of low-level tools exist to program these chips efficiently, but they are considered hard to program, to maintain, and to debug, because they may exhibit non-deterministic behaviors. A solution is to use the higher-level formalism of dataflow programming to specify only the operations to perform and their dependencies. This paradigm may then be combined with the Polyhedral Model, which allows automatic parallelization and optimization of loop nests. This makes programming easier by delegating the low-level work to compilers and static analyzers [41].

Existing dataflow runtime systems either focus on the efficient execution of a single data-flow application, or on scenarii where applications are known a priori. CalMAR is a Multi-Application Dataflow Runtime built on top of the RVC-Cal environment that addresses the problem of executing an a priori unknown number of dataflow applications concurrently on the same multi-core system. Its efficiency has been validated compared to the RVC-CAL traditional approach [27].

6.3.2. Environments for transiently powered devices

An important research initiative has been started in Socrate recently: the study of the new NVRAM technology and its use in ultra-low power context. NVRAM stands for Non-Volatile Radom Access Memory. Non-Volatile mémory has been extising for a while (Nand Flash for instance) but was not sufficiently fast to be used as main memory. Many emerging technologies are forseen for Non-Volatile RAM to replace current RAM [58].

Socrate has started a work on the applicability of NVRAM for *transiantly powered systems*, i.e. systems which may undergo power outage at any time. This study resulted in the Sytare software presented in a research report and at the IoENT conference [39], [37], [17] and also to the starting of an Inria Project Lab: ZEP.

The Sytare software introduces a checkpointing system that takes into account peripherals (ADC, leds, timer, radio communication, etc.) present on all embedded system. Checkpointing is the natural solution to power outage: regularly save the state of the system in NVRAM so as to restore it when power is on again. However, no work on checkpointing took into account the restoration of the states of peripherals, Sytare provides this possibility

6.3.3. Filter synthesis

[46] presents an open-source tool for the automatic design of reliable finite impulse response (FIR) filters, targeting FPGAs. It shows that user intervention can be limited to a very small number of relevant input parameters: a high-level frequency-domain specification, and input/output formats. All the other design parameters are computed automatically, using novel approaches to filter coefficient quantization and direct-form architecture implementation. Our tool guarantees a priori that the resulting architecture respects the specification, while attempting to minimize its cost. Our approach is evaluated on a range of examples and shown to produce designs that are very competitive with the state of the art, with very little design effort.

Linear Time Invariant (LTI) filters are often specified and simulated using high-precision software, before being implemented in low-precision fixed-point hardware. A problem is that the hardware does not behave exactly as the simulation due to quantization and rounding issues. The article [53] advocates the construction of LTI architectures that behave as if the computation was performed with infinite accuracy, then rounded only once to the low-precision output format. From this minimalist specification, it is possible to deduce the optimal values of many architectural parameters, including all the internal data formats. This requires a detailed error analysis that captures not only the rounding errors but also their infinite accumulation in recursive filters. This error analysis then guides the design of hardware satisfying the accuracy specification at the minimal hardware cost. This generic methodology is detailed for the case of low-precision LTI filters in the Direct Form I implemented in FPGA logic. The approach is demonstrated by a fully automated and open-source architecture generator tool, and validated on a range of Infinite Impulse Response filters.

6.3.4. Hardware computer arithmetic

In collaboration with researchers from Istanbul, Turkey, operators have been developed for division by a small positive constant [8]. The first problem studied is the Euclidean division of an unsigned integer by a constant, computing a quotient and a remainder. Several new solutions are proposed and compared against the state of the art. As the proposed solutions use small look-up tables, they match well the hardware resources of an FPGA. The article then studies whether the division by the product of two constants is better implemented as two successive dividers or as one atomic divider. It also considers the case when only a quotient or only a remainder are needed. Finally, it addresses the correct rounding of the division of a floating-point number by a small integer constant. All these solutions, and the previous state of the art, are compared in terms of timing, area, and area-timing product. In general, the relevance domains of the various techniques are very different on FPGA and on ASIC.

[23] presents the new framework for semi-automatic circuit pipelining that will be used in future releases of the FloPoCo generator. From a single description of an operator or datapath, optimized implementations are obtained automatically for a wide range of FPGA targets and a wide range of frequency/latency trade-offs. Compared to previous versions of FloPoCo, the level of abstraction has been raised, enabling easier development, shorter generator code, and better pipeline optimization. The proposed approach is also more flexible

than fully automatic pipelining approaches based on retiming: in the proposed technique, the incremental construction of the pipeline along with the circuit graph enables architectural design decisions that depend on the pipeline.

FPGAs are well known for their ability to perform non-standard computations not supported by classical microprocessors. Many libraries of highly customizable application-specific IPs have exploited this capablity. However, using such IPs usually requires handcrafted HDL, hence significant design efforts. High Level Synthesis (HLS) lowers the design effort thanks to the use of C/C++ dialects for programming FPGAs. However, high-level C language becomes a hindrance when one wants to express non-standard computations: this languages was designed for programming microprocessors and carries with it many restrictions due to this paradigm. This is especially true when computing with floating-point, whose data-types and evaluation semantics are defined by the IEEE-754 and C11 standards. If the high-level specification was a computation on the reals, then HLS imposes a very restricted implementation space. [32]attempts to bridge FPGA application-specific efficiency and HLS ease of use. It specifically targets the ubiquitous floating-point summation-reduction pattern. A source-to-source compiler transforms selected floating-point additions into sequences of simpler operators using non-standard arithmetic formats. This improves performance and accuracy for several benchmarks, while keeping the ease of use of a high-level C description.

The previous uses a variation of Kulisch' proposal to use an internal accumulator large enough to cover the full exponent range of floating-point. With it, sums and dot products become exact operations with one single rounding at the end. This idea failed to materialize in general purpose processors, as it was considered too slow and/or too expensive in terms of resources. It may however be an interesting option in reconfigurable computing, where a designer may use use smaller, more resource-efficient floating-point formats, knowing that sums and dot products will be exact. Another motivation of this work is that these exact operations, contrary to classical floating point ones, are associative, which enables better compiler optimizations in a High-Level Synthesis context. Kulisch proposed several architectures for the large accumulator, all using a sign/magnitude representation: the internal accumulator always represents a positive significand. [52] introduces an architecture using a 2's complement representation instead, and demonstrates improvements over Kulisch' proposal in both area and speed.

Another alternative to floating point is the UNUM, a variable length floating-point format conceived to replace the formats defined in the IEEE 754 standard. [18] discusses the implementation of UNUM arithmetic and reports hardware implementation results for the main UNUM operators.