RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2016

Section New Results

Edition: 2017-08-25

ASAP Project-Team

6. New Results

6.1. Theory of Distributed Systems

6.1.1. t-Resilient Immediate Snapshot is Impossible

Participant: Michel Raynal.

Immediate snapshot is the basic communication object on which relies the read/write distributed computing model made up of n crash-prone asynchronous processes, called iterated distributed model. Each iteration step (usually called a round) uses a new immediate snapshot object, which allows the processes to communicate and cooperate. More precisely, the x-th immediate snapshot object can be used by a process only when it executes the x-th round. An immediate snapshot object can be implemented by an $(n-1)$-resilient algorithm, i.e. an algorithm that tolerates up to $(n-1)$ process crashes (also called wait-free algorithm). Considering a t-crash system model (i.e. a model in which up to t processes are allowed to crash), this work [46] is on the construction of an extension of immediate snapshot objects to t-resiliency. In the t-crash system model, at each round each process may be ensured to get values from at least $n-t$ processes, and t-immediate snapshot has the properties of classical immediate snapshot (1-immediate snapshot) but ensures that each process will get values form at least $n-t$ processes. Its main result is the following. While there is a (deterministic) t-resilient read/write-based algorithm implementing t-immediate snapshot in a t-crash system when $t = n-1$, there is no t-resilient algorithm in a t-crash model when $t \in [1..(n-2)]$. This means that the notion of t-resilience is inoperative when one has to implement immediate snapshot for these values of t: the model assumption "at most $t < n-1$ processes may crash" does not provide us with additional computational power allowing for the design of genuine t-resilient algorithms (genuine meaning that such a t-resilient algorithm would work in the t-crash model, but not in the $(t+1)$-crash model). To show these results, the paper relies on well-known distributed computing agreement problems such as consensus and k-set agreement.

This work was done in collaboration with Carole Delporte, Hugues Fauconnier, and Sergio Rajsbaum, and appeared at SIROCCO 2016.

6.1.2. Two-Bit Messages are Sufficient to Implement Atomic Read/Write Registers in Crash-Prone Systems

Participant: Michel Raynal.

Atomic registers are certainly the most basic objects of computing science. Their implementation on top of an n-process asynchronous message-passing system has received a lot of attention. It has been shown that $t < n/2$ (where t is the maximal number of processes that may crash) is a necessary and sufficient requirement to build an atomic register on top of a crash-prone asynchronous message-passing system. Considering such a context, this work [49] presents an algorithm which implements a single-writer multi-reader atomic register with four message types only, and where no message needs to carry control information in addition to its type. Hence, two bits are sufficient to capture all the control information carried by all the implementation messages. Moreover, the messages of two types need to carry a data value while the messages of the two other types carry no value at all. As far as we know, this algorithm is the first with such an optimality property on the size of control information carried by messages. It is also particularly efficient from a time complexity point of view.

This work was done in collaboration with Achour Mostefaoui, and appeared at PODC 2016.

6.2. Network and Graph Algorithms

6.2.1. Vertex Coloring with Communication and Local Memory Constraints in Synchronous Broadcast Networks

Participants: Hicham Lakhlef, Michel Raynal, Francois Taiani.

This work [41] considers the broadcast/receive communication model in which message collisions and message conflicts can occur because processes share frequency bands. (A collision occurs when, during the same round, messages are sent to the same process by too many neighbors. A conflict occurs when a process and one of its neighbors broadcast during the same round.) More precisely, this work considers the case where, during a round, a process may either broadcast a message to its neighbors or receive a message from at most m of them. This captures communication-related constraints or a local memory constraint stating that, whatever the number of neighbors of a process, its local memory allows it to receive and store at most m messages during each round. This work defines first the corresponding generic vertex multi-coloring problem (a vertex can have several colors). It focuses then on tree networks, for which it presents a lower bound on the number of colors K that are necessary (namely, $K = \lceil \frac{\Delta}{m} \rceil + 1$, where Δ is the maximal degree of the communication graph), and an associated coloring algorithm, which is optimal with respect to K.

6.2.2. Optimal Collision/Conflict-Free Distance-2 Coloring in Wireless Synchronous Broadcast/Receive Tree Networks

Participants: Davide Frey, Hicham Lakhlef, Michel Raynal.

We studied the problem of decentralized distance-2 coloring in message-passing systems where communication is (a) synchronous and (b) based on the "broadcast/receive" pair of communication operations. "Synchronous" means that time is discrete and appears as a sequence of time slots (or rounds) such that each message is received in the very same round in which it is sent. "Broadcast/receive" means that during a round a process can either broadcast a message to its neighbors or receive a message from one of them. In such a communication model, no two neighbors of the same process, nor a process and any of its neighbors, must be allowed to broadcast during the same time slot (thereby preventing message collisions in the first case, and message conflicts in the second case). From a graph theory point of view, the allocation of slots to processes is know as the distance-2 coloring problem: a color must be associated with each process (defining the time slots in which it will be allowed to broadcast) in such a way that any two processes at distance at most 2 obtain different colors, while the total number of colors is "as small as possible". In this context, we proposed a parallel message-passing distance-2 coloring algorithm suited to trees, whose roots are dynamically defined. This algorithm, which is itself collision-free and conflict-free, uses $\Delta + 1$ colors where Δ is the maximal degree of the graph (hence the algorithm is color-optimal). It does not require all processes to have different initial identities, and its time complexity is $O(d\Delta)$, where d is the depth of the tree. As far as we know, this is the first distributed distance-2 coloring algorithm designed for the broadcast/receive round-based communication model, which owns all the previous properties. We published these results in [39].

6.2.3. Efficient Plurality Consensus, or: The Benefits of Cleaning Up from Time to Time

Participant: George Giakkoupis.

Plurality consensus considers a network of n nodes, each having one of k opinions. Nodes execute a (randomized) distributed protocol with the goal that all nodes adopt the *plurality* (the opinion initially supported by the most nodes). Communication is realized via the random phone call model. A major open question has been whether there is a protocol for the complete graph that converges (w.h.p.) in polylogarithmic time and uses only polylogarithmic memory per node (local memory). We answered this question affirmatively.

In [22], we propose two protocols that need only mild assumptions on the bias in favor of the plurality. As an example of our results, consider the complete graph and an arbitrarily small constant multiplicative bias in favor of the plurality. Our first protocol achieves plurality consensus in $O(\log k \cdot \log\log n)$ rounds using $\log k + O(\log\log k)$ bits of local memory. Our second protocol achieves plurality consensus in $O(\log n \cdot \log\log n)$ rounds using only $\log k + 4$ bits of local memory. This disproves a conjecture by Becchetti et al. (SODA'15) implying that any protocol with local memory $\log k + O(1)$ has worst-case runtime $\Omega(k)$. We provide similar bounds for much weaker bias assumptions. At the heart of our protocols lies an *undecided state*, an idea introduced by Angluin et al. (Distributed Computing'08).

This work was done in collaboration with Petra Berenbrink (SFU), Tom Friedetzky (Durham University), and Peter Kling (SFU).

6.2.4. Bounds on the Voter Model in Dynamic Networks

Participants: George Giakkoupis, Anne-Marie Kermarrec.

In the *voter model*, each node of a graph has an opinion, and in every round each node chooses independently a random neighbour and adopts its opinion. We are interested in the *consensus time*, which is the first point in time where all nodes have the same opinion. In [23], we consider dynamic graphs in which the edges are rewired in every round (by an adversary) giving rise to the graph sequence $G_1, G_2, \cdots$, where we assume that G_i has conductance at least ϕ_i. We assume that the degrees of nodes don't change over time as one can show that the consensus time can become super-exponential otherwise. In the case of a sequence of d-regular graphs, we obtain asymptotically tight results. Even for some static graphs, such as the cycle, our results improve the state of the art. Here we show that the expected number of rounds until all nodes have the same opinion is bounded by $O(m/(\delta \cdot \phi))$, for any graph with m edges, conductance ϕ, and degrees at least δ. In addition, we consider a *biased* dynamic voter model, where each opinion i is associated with a probability P_i, and when a node chooses a neighbour with that opinion, it adopts opinion i with probability P_i (otherwise the node keeps its current opinion). We show for any regular dynamic graph, that if there is an $\epsilon > 0$ difference between the highest and second highest opinion probabilities, and at least $\Omega(\log n)$ nodes have initially the opinion with the highest probability, then all nodes adopt w.h.p. that opinion. We obtain a bound on the convergence time, which becomes $O(\log n/\phi)$ for static graphs.

This work was done in collaboration with Petra Berenbrink (SFU), and Frederik Mallmann-Trenn (SFU).

6.2.5. How Asynchrony Affects Rumor Spreading Time

Participant: George Giakkoupis.

In standard randomized (push-pull) rumor spreading, nodes communicate in synchronized rounds. In each round every node contacts a random neighbor in order to exchange the rumor (i.e., either push the rumor to its neighbor or pull it from the neighbor). A natural asynchronous variant of this algorithm is one where each node has an independent Poisson clock with rate 1, and every node contacts a random neighbor whenever its clock ticks. This asynchronous variant is arguably a more realistic model in various settings, including message broadcasting in communication networks, and information dissemination in social networks.

In [35] we study how asynchrony affects the rumor spreading time, that is, the time before a rumor originated at a single node spreads to all nodes in the graph. Our first result states that the asynchronous push-pull rumor spreading time is asymptotically bounded by the standard synchronous time. Precisely, we show that for any graph G on n-nodes, where the synchronous push-pull protocol informs all nodes within $T(G)$ rounds with high probability, the asynchronous protocol needs at most time $O(T(G) + \log n)$ to inform all nodes with high probability. On the other hand, we show that the expected synchronous push-pull rumor spreading time is bounded by $O(\sqrt{n})$ times the expected asynchronous time.

These results improve upon the bounds for both directions shown recently by Acan et al. (PODC 2015). An interesting implication of our first result is that in regular graphs, the weaker push-only variant of synchronous rumor spreading has the same asymptotic performance as the synchronous push-pull algorithm.

This work was done in collaboration with Yasamin Nazari and Philipp Woelfel from the University of Calgary.

6.2.6. Amplifiers and Suppressors of Selection for the Moran Process on Undirected Graphs

Participant: George Giakkoupis.

In [47] we consider the classic Moran process modeling the spread of genetic mutations, as extended to structured populations by Lieberman et al. (Nature, 2005). In this process, individuals are the vertices of a connected graph G. Initially, there is a single mutant vertex, chosen uniformly at random. In each step, a random vertex is selected for reproduction with a probability proportional to its fitness: mutants have fitness $r > 1$, while non-mutants have fitness 1. The vertex chosen to reproduce places a copy of itself to a uniformly random neighbor in G, replacing the individual that was there. The process ends when the mutation either reaches fixation (i.e., all vertices are mutants), or gets extinct. The principal quantity of interest is the probability with which each of the two outcomes occurs.

A problem that has received significant attention recently concerns the existence of families of graphs, called strong amplifiers of selection, for which the fixation probability tends to 1 as the order n of the graph increases, and the existence of strong suppressors of selection, for which this probability tends to 0. For the case of directed graphs, it is known that both strong amplifiers and suppressors exist. For the case of undirected graphs, however, the problem has remained open, and the general belief has been that neither strong amplifiers nor suppressors exist. In this work we disprove this belief, by providing the first examples of such graphs. The strong amplifier we present has fixation probability $1 - \widetilde{O}(n^{-1/3})$, and the strong suppressor has fixation probability $\widetilde{O}(n^{-1/4})$. Both graph constructions are surprisingly simple. We also prove a general upper bound of $1 - \widetilde{\Omega}(n^{-1/3})$ on the fixation probability of any undirected graph. Hence, our strong amplifier is existentially optimal.

6.3. Scalable Systems

6.3.1. Cache locality is not enough: High-Performance Nearest Neighbor Search with Product Quantization Fast Scan

Participants: Fabien Andre, Anne-Marie Kermarrec.

Nearest Neighbor (NN) search in high dimension is an important feature in many applications (e.g., image retrieval, multimedia databases). Product Quantization (PQ) is a widely used solution which offers high performance, i.e., low response time while preserving a high accuracy. PQ represents high-dimensional vectors (e.g., image descriptors) by compact codes. Hence, very large databases can be stored in memory, allowing NN queries without resorting to slow I/O operations. PQ computes distances to neighbors using cache-resident lookup tables, thus its performance remains limited by (i) the many cache accesses that the algorithm requires, and (ii) its inability to leverage SIMD instructions available on modern CPUs. In this paper, we advocate that cache locality is not sufficient for efficiency. To address these limitations, in [19] we design a novel algorithm, PQ Fast Scan, that transforms the cache-resident lookup tables into small tables, sized to fit SIMD registers. This transformation allows (i) in-register lookups in place of cache accesses and (ii) an efficient SIMD implementation. PQ Fast Scan has the exact same accuracy as PQ, while having 4 to 6 times lower response time (e.g., for 25 million vectors, scan time is reduced from 74ms to 13ms).

6.3.2. Toward an Holistic Approach of Systems-of-Systems

Participants: Simon Bouget, David Bromberg, Francois Taiani.

Large scale distributed systems have become ubiquitous, from on-line social networks to the Internet-of-Things. To meet rising expectations (scalability, robustness, flexibility,...) these systems increasingly espouse complex distributed architectures, that are hard to design, deploy and maintain. To grasp this complexity, developers should be allowed to assemble large distributed systems from smaller parts using a seamless, high-level programming paradigm. We present in [24] such an assembly-based programming framework, enabling developers to easily define and realize complex distributed topologies as a construction of simpler blocks (e.g. rings, grids). It does so by harnessing the power of self-organizing overlays, that is made accessible to developers through a high-level Domain Specific Language and self-stabilizing run-time. Our evaluation further shows that our approach is generic, expressive, low-overhead and robust.

6.3.3. Speed for the Elite, Consistency for the Masses: Differentiating Eventual Consistency in Large-Scale Distributed Systems

Participants: Davide Frey, Pierre-Louis Roman, Francois Taiani.

Eventual consistency is a consistency model that emphasizes liveness over safety; it is often used for its ability to scale as distributed systems grow larger. Eventual consistency tends to be uniformly applied to an entire system, but we argue that there is a growing demand for differentiated eventual consistency requirements.

We address this demand with UPS [34], a novel consistency mechanism that offers differentiated eventual consistency and delivery speed by working in pair with a two-phase epidemic broadcast protocol. We propose a closed-form analysis of our approach's delivery speed, and we evaluate our complete mechanism experimentally on a simulated network of one million nodes. To measure the consistency trade-off, we formally define a novel and scalable consistency metric that operates at runtime. In our simulations, UPS divides by more than 4 the inconsistencies experienced by a majority of the nodes, while reducing the average latency incurred by a small fraction of the nodes from 6 rounds down to 3 rounds.

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

6.3.4. *Bringing Secure Bitcoin Transactions to your Smartphone*

Participants: Davide Frey, Pierre-Louis Roman, Francois Taiani.

To preserve the Bitcoin ledger's integrity, a node that joins the system must download a full copy of the entire Bitcoin blockchain if it wants to verify newly created blocks. At the time of writing, the blockchain weights 79 GiB and takes hours of processing on high-end machines. Owners of low-resource devices (known as thin nodes), such as smartphones, avoid that cost by either opting for minimum verification or by depending on full nodes, which weakens their security model.

In this work [33], we propose to harden the security model of thin nodes by enabling them to verify blocks in an adaptive manner, with regards to the level of targeted confidence, with low storage requirements and a short bootstrap time. Our approach exploits sharing within a distributed hash table (DHT) to distribute the storage load, and a few additional hashes to prevent attacks on this new system.

This work was done in collaboration with Marc X. Makkes and Spyros Voulgaris from Vrije Universiteit Amsterdam (The Netherlands).

6.3.5. *Multithreading Approach to Process Real-Time Updates in KNN Algorithms*

Participants: Anne-Marie Kermarrec, Nupur Mittal, Javier Olivares.

K-Nearest Neighbors algorithm is the core of a considerable amount of online services and applications, like recommendation engines, content-classifiers, information retrieval systems, etc. The users of these services change their preferences and evolve with time, aggravating the computational challenges of KNN more with the ever evolving data to process. In this work [48], we present *UpKNN*: an efficient thread-based approach to take the updates of users preferences into account while it computes the KNN efficiently, keeping a check on the wall-time.

By using an efficient thread-based approach, *UpKNN* processes millions of updates online, on a single commodity PC. Our experiments confirm the scalability of *UpKNN*, both in terms of the number of updates processed and the threads used. *UpKNN* achieves speedups ranging from 13.64X to 49.5X in the processing of millions of updates, with respect to the performance of a non-partitioned baseline. These results are a direct consequence of reducing the number of disk operations, roughly speaking, only 1% disk operations are performed as compared to the baseline.

6.3.6. *The Out-of-Core KNN Awakens: The Light Side of Computation Force on Large Datasets*

Participants: Anne-Marie Kermarrec, Javier Olivares.

K-Nearest Neighbors (KNN) is a crucial tool for many applications, e.g. recommender systems, image classification and web-related applications. However, KNN is a resource greedy operation particularly for large datasets. We focus on the challenge of KNN computation over large datasets on a single commodity PC with limited memory. We propose a novel approach [27] to compute KNN on large datasets by leveraging both disk and main memory efficiently. The main rationale of our approach is to minimize random accesses to disk, maximize sequential accesses to data and efficient usage of only the available memory.

We evaluate our approach on large datasets, in terms of performance and memory consumption. The evaluation shows that our approach requires only 7% of the time needed by an in-memory baseline to compute a KNN graph.

6.3.7. Partial Replication Policies for Dynamic Distributed Transactional Memory in Edge Clouds

Participant: Francois Taiani.

Distributed Transactional Memory (DTM) can play a fundamental role in the coordination of participants in edge clouds as a support for mobile distributed applications. DTM emerges as a concurrency mechanism aimed at simplifying distributed programming by allowing groups of operations to execute atomically, mirroring the well-known transaction model of relational databases. In spite of recent studies showing that partial replication approaches can present gains in the scalability of DTMs by reducing the amount of data stored at each node, most DTM solutions follow a full replication scheme. The few partial replicated DTM frameworks either follow a random or round-robin algorithm for distributing data onto partial replication groups. In order to overcome the poor performance of these schemes, this work [36] investigates policies to extend the DTM to efficiently and dynamically map resources on partial replication groups. The goal is to understand if a dynamic service that constantly evaluates the data mapped into partial replicated groups can contribute to improve DTM based systems performance.

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

6.3.8. Being Prepared in a Sparse World: The Case of KNN Graph Construction

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

Work [25] presents KIFF, a generic, fast and scalable KNN graph construction algorithm. KIFF directly exploits the bipartite nature of most datasets to which KNN algorithms are applied. This simple but powerful strategy drastically limits the computational cost required to rapidly converge to an accurate KNN solution, especially for sparse datasets. Our evaluation on a representative range of datasets show that KIFF provides, on average, a speed-up factor of 14 against recent state-of-the art solutions while improving the quality of the KNN approximation by 18

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

6.3.9. Exploring the Use of Tags for Georeplicated Content Placement

Participants: Stephane Delbruel, Davide Frey, Francois Taiani.

A large portion of today's Internet traffic originates from streaming and video services. Such services rely on a combination of distributed datacenters, powerful content delivery networks (CDN), and multi-level caching . In spite of this infrastructure, storing, indexing, and serving these videos remains a daily engineering challenge that requires increasing efforts on the part of providers and ISPs. In this work [30], we explore how the tags attached to videos by users could help improve this infrastructure, and lead to better performance on a global scale. Our analysis shows that tags can be interpreted as markers of a video's geographic diffusion, with some tags strongly linked to well identified geographic areas. Based on our findings, we demonstrate the potential of tags to help predict distribution of a video's views, and present results suggesting that tags can help place videos in globally distributed datacenters. We show in particular that even a simplistic approach based on tags can help predict a minimum of 65.9% of a video's views for a majority of videos, and that a simple tag-based placement strategy is able to improve the hit rate of a distributed on-line video service by up to 6.8% globally over a naive random allocation.

6.3.10. Mignon: A Fast Decentralized Content Consumption Estimation in Large-Scale Distributed Systems

Participants: Stephane Delbruel, Davide Frey, Francois Taiani.

Although many fully decentralized content distribution systems have been proposed, they often lack key capabilities that make them difficult to deploy and use in practice. In this work [31], we look at the particular problem of content consumption prediction, a crucial mechanism in many such systems. We propose a novel, fully decentralized protocol that uses the tags attached by users to on-line content, and exploits the properties of self-organizing kNN overlays to rapidly estimate the potential of a particular content without explicit aggregation.

6.4. Privacy in User Centric Applications

6.4.1. Hybrid Recommendations with Dynamic Similarity Measure

Participants: Anne-Marie Kermarrec, Nupur Mittal.

This project aims to combine the classical methods of content based and collaborative filtering recommendations, in addition to dynamic similarity computations. The objective is to exploit the varied item-data available from the world wide web, to overcome trivial problems like that of cold-start. In this work, we have designed a new similarity metric inspired from the existing DICE similarity that takes into account changing item/user behavior to compute updated similarity values for the purpose of recommendations. The work leverages the idea of content based recommendations as a first step to create vivid user and item profiles that are iteratively updated.

This work was done in collaboration with Rachid Guerraoui (EPFL, Switzerland), Rhicheek Patra (EPFL, Switzerland).

6.4.2. Lightweight Privacy-Preserving Averaging for the Internet of Things

Participants: Davide Frey, George Giakkoupis, Julien Lepiller.

The number of connected devices is growing continuously, and so is their presence into our everyday lives. From GPS-enabled fitness trackers, to smart fridges that tell us what we need to buy at the grocery store, connected devices—things—have the potential to collect and make available significant amounts of information. On the one hand, this information may provide useful services to users, and constitute a statistical gold mine. On the other, its availability poses serious privacy threats for users. In this work, we designed two new protocols that make it possible to aggregate personal information collected by smart devices in the form of an average, while preventing attackers from learning the details of the non-aggregated data. The first protocol exploits randomness and decomposition into shares as techniques to obfuscate the value associated with each node and lightweight encryption techniques to withstand eavesdropping attacks. The second exploits only randomness and encryption. We carried out a preliminary evaluation and published the results related to the first protocol in [18].

This work was done in collaboration with Tristan Allard from the DRUID Team at IRISA, Rennes.

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

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

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

percent. A preliminary version of this work was published at EuroSec 2015 [57]. This year, we significantly extended the work during the summer internship of Florestan De Moor. Specifically, we considered new attacks that specifically target our novel similarity metric and showed that regardless of the attack configuration, our metric can preserve the privacy of users without hampering recommendation quality. A new paper with these new results was submitted to PETS 2017.

6.4.4. Privacy-Preserving Distributed Collaborative Filtering

Participants: Davide Frey, Anne-Marie Kermarrec.

In this work, we propose a new mechanism to preserve privacy while leveraging user profiles in distributed recommender systems. Our mechanism relies on (*i*) an original obfuscation scheme to hide the exact profiles of users without significantly decreasing their utility, as well as on (*ii*) a randomized dissemination protocol ensuring differential privacy during the dissemination process.

We compare our mechanism with a non-private as well as with a fully private alternative. We consider a real dataset from a user survey and report on simulations as well as planetlab experiments. We dissect our results in terms of accuracy and privacy trade-offs, bandwith consumption, as well as resilience to a censorship attack. In short, our extensive evaluation shows that our twofold mechanism provides a good trade-off between privacy and accuracy, with little overhead and high resilience.

This work was done with Antoine Boutet and Arnaud Jegou when they were part of the team, and in collaboration with Rachid Guerraoui from EPFL. But the complete results were published this year in [15].

CIDRE Project-Team

7. New Results

7.1. Intrusion Detection

7.1.1. Intrusion Detection in Distributed Systems

Alert Correlation: In large systems, multiple (host and network) Intrusion Detection Systems (IDS) and many sensors are usually deployed. They continuously and independently generate notifications (event's observations, warnings and alerts). To cope with this amount of collected data, alert correlation systems have to be designed. An alert correlation system aims at exploiting the known relationships between some elements that appear in the flow of low level notifications to generate high semantic meta-alerts. The main goal is to reduce the number of alerts returned to the security administrator and to allow a higher level analysis of the situation. However, producing correlation rules is a highly difficult operation, as it requires both the knowledge of an attacker, and the knowledge of the functionalities of all IDSes involved in the detection process. In the context of the PhD of Erwan Godefroy [1], we focus on the transformation process that allows to translate the description of a complex attack scenario into correlation rules and its assessment. We show that, once a human expert has provided an action tree derived from an attack tree, a fully automated transformation process can generate exhaustive correlation rules that would be tedious and error prone to enumerate by hand.

Long lived attack campaigns known as Advanced Persistent Threats (APTs) have emerged as a serious security risk. These attack campaigns are customised for their target and performed step by step during months on end. The major difficulty in detecting an APT is keeping track of the different steps logged over months of monitoring and linking them. In [11], we describe TerminAPTor, an APT detector which highlights links between the traces left by attackers in the monitored system during the different stages of an attack campaign. TerminAPTor tackles this challenge by resorting to Information Flow Tracking (IFT). Our main contribution is showing that IFT can be used to highlight APTs. Additionally, we describe a generic representation of APTs and validate our IFT-based APT detector.

Inferring the normal behavior of an application: In [29], [6], [41], we propose an approach to detect intrusions that affect the behavior of distributed applications. To determine whether an observed behavior is normal or not (occurence of an attack), we rely on a model of normal behavior. This model has been built during an initial training phase (machine learning approach). During this preliminary phase, the application is executed several times in a safe environment. The gathered traces (sequences of actions) are used to generate an automaton that characterizes all these acceptable behaviors. To reduce the size of the automaton and to be able to accept more general behaviors that are close to the observed traces, the automaton is transformed. These transformations may lead to introduce unacceptable behaviors. Our current work aims at identifying the possible errors tolerated by the compacted automaton.

This approach is particularly appealing to detect intrusions in industrial control systems since these systems exhibit well-defined behaviors at different levels: network level (network communication patterns, protocol specifications, etc.), control level (continue and discrete process control laws), or even the state of the local resources (memory or CPU). Industrial control systems (ICS) can be subject to highly sophisticated attacks which may lead the process towards critical states. Due to the particular context of ICS, protection mechanisms are not always practical, nor sufficient. On the other hand, developing a process-aware intrusion detection solution with satisfactory alert characterization remains an open problem. In [20], we focus on process-aware attacks detection in sequential control systems. We build on results from runtime verification and specification mining to automatically infer and monitor process specifications. Such specifications are represented by sets of temporal safety properties over states and events corresponding to sensors and actuators. The properties are then synthesized as monitors which report violations on execution traces. We develop an efficient specification mining algorithm and use filtering rules to handle the large number of mined properties. Furthermore, we introduce the notion of activity and discuss its relevance to both specification mining and attack detection

in the context of sequential control systems. The proposed approach is evaluated in a hardware-in-the-loop setting subject to targeted process-aware attacks. Overall, due to the explicit handling of process variables, the solution provides a better characterization of the alerts and a more meaningful understanding of false positives.

7.1.2. Illegal Information Flow Detection

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

Illegal Information Flow in Web-browser: In the context of the CominLabs SECCLOUD project, we were interested in implementing our approach to detect illegal information flow in web-browser. We have proposed a new secure information flow control model specifically designed for JavaScript [28]. In our approach, we augment the standard symbol table with a mechanism that replaces the reference address for secret values based on the current execution stack. This mechanism also ensures that the secret is stored in a dedicated memory location thereby protecting the secret from any unintended leakage or modification by a malicious JavaScript. This work on detection of illegal information flow in JavaScript has received the best paper award at the 9th International Conference on Security of Information and Networks (SIN 2016) [28].

Later Deepak Subramanian has improved this approach and optimized the computation time required to determine the legacy of information flows. An approach which begins with a learning phase allows to increase the accuracy of the proposed solution. Information about the modified variables are kept in memory to perform a more accurate analysis of the indirect information flows. This self-correcting information flow control model for a web-browser is described in [27].

Information Leaks: Qualitative information flow aims at detecting information leaks, whereas the emerging quantitative techniques target the estimation of information leaks. Quantifying information flow in the presence of low inputs is challenging, since the traditional techniques of approximating and counting the reachable states of a program no longer suffice. In [32], we propose an automated quantitative information flow analysis for imperative deterministic programs with low inputs. The approach relies on a novel abstract domain, the cardinal abstraction, in order to compute a precise upper-bound over the maximum leakage of batch-job programs. We prove the soundness of the cardinal abstract domain by relying on the framework of abstract interpretation. We also prove its precision with respect to a flow-sensitive type system for the two-point security lattice.

More generally, for his research activities during his PhD thesis, Mounir Assaf has received the 2016 thesis prize awarded by the GDR GPL (Engineering Programming and Software).

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

To fight repackaged applications containing malicious code, most official application marketplaces have implemented security analysis tools that try to detect and remove malware. Countermeasures adopted by the attackers to bypass these new controls can be divided into two main approaches: avoiding static analysis and avoiding dynamic analysis [39]. A static analysis of an application consists of analysing its code and its resources without executing it. Conversely, dynamic analysis stands for any kind of analysis that requires executing the application in order to observe its actions.

The Kharon project [19] goes a step further from classical dynamic analysis of malware (http://kharon.gforge.inria.fr). Funded by the Labex CominLabs and involving partners of Centrale-Supélec, Inria and INSA Centre Val de Loire, this project aims to capture a compact and comprehensive

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

representation of malware. To achieve such a goal we have developed tools to monitor operating systems' information flows induced by the execution of a marked application. We support the idea that the best way to understand malware impact is to observe it in its normal execution environment i.e., a real smartphone. Additionally, the main challenge is to be able to trigger malicious behaviours even if the malware tries to escape dynamic analysis.

In this context, we have developed an original solution that mainly consists of 'helping the malware to execute'. In other words we slightly modify the bytecode of the infected application in order to defeat the protection against dynamic analysis and we execute the suspicious code in its most favourable execution conditions. Thus, our software helps us understand malware's objectives and the consequences on the health of a user's device. In particular, we use a global control flow graph (CFG) to exhibit an execution path to reach specific parts of code [42].

To achieve stealthiness when attacking a mobile device, an effective approach is the use of a covert channel built by two colluding applications to locally exchange data. Since this process is tightly coupled with the used hiding method, its detection is a challenging task, also worsened by the very low transmission rates. Using general indicators such as the energy consumed by the device, we propose in [5] an approach to detect the hidden data exchange between colluding applications and show its feasibility and effectiveness through different experimental results.

Our main research direction and challenge is to develop new and original protections against malicious applications that try to defeat classical dynamic analysis.

7.1.3. Intrusion Detection in Low-Level Software Components

In order to protect the IDS itself, we have initiated different research activities in the domain of hardware security. Our goal is to use co-design software/hardware approaches against traditional software attacks. In a bilateral research project with HP Inc Research Labs, we investigate how dedicated hardware could be used to monitor the whole software stack (from the firmware to the user-mode applications). In the CominLabs HardBlare project, we study the use of a dedicated co-processor to enforce Dynamic Information Flow Control on the main CPU. Finally, in the context of the PhD thesis of Thomas Lethan (ANSSI), we investigate the use of formal methods to evaluate the security guarantees provided by hardware platforms, which combine different CPUs, chipsets and memories. Over time, hardware designs have constantly grown in complexity and modern platforms involve multiple interconnected hardware components. During the last decade, several vulnerability disclosures have proven that trust in hardware can be misplaced. In [21], [37], we give a formal definition of Hardware-based Security Enforcement (HSE) mechanisms, a class of security enforcement mechanisms such that a software component relies on the underlying hardware platform to enforce a security policy. We then model a subset of a x86-based hardware platform specifications and we prove the soundness of a realistic HSE mechanism within this model using Coq, a proof assistant system.

The HardBlare project proposes a software/hardware co-design methodology to ensure that security properties are preserved all along the execution of the system but also during files storage. It is based on the Dynamic Information Flow Tracking (DIFT) that generally consists in attaching tags to denote the type of information that are saved or generated within the system. These tags are then propagated when the system evolves and information flow control is performed in order to guarantee the safe execution and storage within the system monitored by security policies [43].

In [30] we introduce an efficient approach for DIFT (Dynamic Information Flow Tracking) implementations on reconfigurable chips. Existing solutions are either hardly portable or bring unsatisfactory time overheads. This work presents an innovative implementation for DIFT on reconfigurable SoCs such as Xilinx Zynq devices.

In [7], we detail a hardware-assisted approach for information flow tracking implemented on reconfigurable chips. Current solutions are either time-consuming or hardly portable (modifications of both software/hardware layers). This work takes benefits from debug components included in ARMv7 processors to retrieve details on instructions committed by the CPU. First results in terms of silicon area and time overheads are also given.

7.1.4. Vizualization

The large quantities of alerts generated by intrusion detection systems (IDS) make very difficult to distinguish on a network real threats from noise. To help solving this problem, we propose VEGAS [12], an alerts visualization and classification tool that allows first line security operators to group alerts visually based on their principal component analysis (PCA) representation. VEGAS is included in a workflow in such a way that once a set of similar alerts has been collected and diagnosed, a filter is generated that redirects forthcoming similar alerts to other security analysts that are specifically in charge of this set of alerts, in effect reducing the flow of raw undiagnosed alerts.

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

7.2. Privacy

7.2.1. Image Encryption

More and more users prefer to share their photos through image-sharing platforms of social networks than using e-mail or personal webpages. Since the provider of the image-sharing platform can clearly know the contents of any published images, the users have to trust the provider to respect their privacy or has to encrypt their images. In the context of the PhD of Kun He [18], [17], [16], we have proposed an IND-CPA image encryption algorithm that preserve the image format after encryption, and we have shown that our encryption algorithm can be used on several widely used image-sharing platforms such as Flickr, Pinterest, Google+ and Twitter.

7.2.2. Fingerprinting

Active fingerprinting schemes were originally invented to deter malicious users from illegally releasing an item, such as a movie or an image. To achieve this, each time an item is released, a different fingerprint is embedded in it. In the context of the PhD of Julien Lolive, we have defined the first privacy-preserving asymmetric fingerprinting protocol based on Tardos codes [2]. This protocol is optimal with respect to traitor tracing. We also formally proved that our protocol achieves the properties of correctness, anti-framing, traitor tracing, as well as buyer- and item-unlinkability.

7.3. Communication and Synchronization in Distributed Systems

7.3.1. Routing Protocol for Tactical Mobile Ad Hoc Networks

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

New generation military equipment, soldiers and vehicles, use wireless technology to communicate on the battlefield. During missions, they form a MANET. Since the battlefield includes coalition, each group may communicate with another group, and inter-MANET communication may be established. Inter-MANET (or inter-domain MANET) communication should allow communication, but maintain a control on the exchanged information. Several protocols have been proposed in order to handle inter-domain routing for tactical MANETs. In [14], [33], we describe and compare three solutions. Based on this analysis, we propose some preconizations to design Inter-domain protocols for MANET.

In [15], we present a coalition context and describe the functional hypothesis we used. Then, we propose a protocol that would fit such a network and conduct experimentation that tend to show that our proposition is quite efficient.

7.3.2. Communication and Synchronization Primitives

Use of Primitives to Limit Equivocation: We consider the approximate consensus problem in a partially connected network of n nodes where at most f nodes may suffer from Byzantine faults. In [22], we study under which conditions this problem can be solved using an iterative algorithm. A Byzantine node can equivocate: it may provide different values to its neighbors. To restrict the possibilities of equivocation, the 3-partial multicast primitive is considered. When a (correct or faulty) node uses this communication primitive, it provides necessarily the same value to the two identified receivers. Based on this communication primitive, a novel condition called f-resilient is proposed and proved to be necessary and sufficient to solve the approximate Byzantine consensus problem in a synchronous network.

The Test&Set Problem: In [35], we present a solution to the well-known problem of synchronization in a distributed asynchronous system prone to process crashes. This problem is also known as the Test&Set problem. The Test&Set is a distributed synchronization protocol that, when invoked by a set of processes, returns a unique winning process. This unique process is then allowed to use, for instance, a shared resource. Recently many advances in implementing Test&Set objects have been achieved, however all of them uniquely target the shared memory model. In this paper we propose an implementation of a Test&Set object for a message passing distributed system. This implementation can be invoked by any number $n \leq N$ of processes where N is the total number of processes in the system. We show in this paper, using a Markov model, that our implementation has an expected step complexity in O(log n) and we give an explicit formula for the distribution of the number of steps needed to solve the problem.

7.3.3. Dependability in Cloud Storage

The quantity of data in the world is steadily increasing bringing challenges to storage system providers to find ways to handle data efficiently in terms of dependability and in a cost-effectively manner. We have been interested in cloud storage which is a growing trend in data storage solution. For instance, the International Data Corporation (IDC) predicts that by 2020, nearly 40% of the data in the world will be stored or processed in a cloud. The thesis of Pierre Obame [3] addressed challenges around data access latency and dependability in cloud storage. We proposed Mistore, a distributed storage system that we designed to ensure data availability, durability, low access latency by leveraging the Digital Subscriber Line (xDSL) infrastructure of an Internet Service Provider (ISP). Mistore uses the available storage resources of a large number of home gateways, Points of Presence, and datacenters for content storage and caching facilities. Mistore also targets data consistency by providing multiple types of data consistency criteria and a versioning system. We also considered the data security and confidentiality in the context of storage systems applying data deduplication which is becoming one of the most popular data technologies to reduce the storage cost and we design a data deduplication method that is secure against malicious clients while remaining efficient in terms of network bandwidth and storage space savings.

7.3.4. Decentralized Cryptocurrency Systems

Decentralized cryptocurrency systems offer a medium of exchange secured by cryptography, without the need of a centralized banking authority. Among others, Bitcoin is considered as the most mature one [10]. Its popularity lies on the introduction of the concept of the blockchain, a public distributed ledger shared by all participants of the system. Double spending attacks and blockchain forks are two main issues in blockchain-based protocols. The first one refers to the ability of an adversary to use the very same bitcoin more than once, while blockchain forks cause transient inconsistencies in the blockchain. In [9], we show through probabilistic analysis that the reliability of recent solutions that exclusively rely on a particular type of Bitcoin actors, called miners, to guarantee the consistency of Bitcoin operations, drastically decreases with the size of the blockchain.

Some recent works have proposed to improve upon Bitcoin weaknesses. In [31], we analyze of one of these recent works, and shows through an analytical performance evaluation that new Bitcoin improvements are still needed.

7.3.5. Large Scale Systems

Population Protocol: the computational model of population protocols is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. Significant work has been done so far to determine which problems are solvable in this model and at which cost in terms of states used by the protocols and time needed to converge. The problem tackled in [23] is the population proportion problem: each agent starts independently from each other in one of two states, say A or B, and the objective is for each agent to determine the proportion of agents that initially started in state A, assuming that each agent only uses a finite set of state, and does not know the number n of agents. We propose a solution which guarantees with any high probability that after O(log n) interactions any agent outputs with a precision given in advance, the proportion of agents that start in state A. The population proportion problem is a generalization of both the majority and counting problems, and thus our solution solves both problems. We show that our solution is optimal in time and space. Simulation results illustrate our theoretical analysis.

Propagation Time of a Rumor: the context of this work 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. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In [24], we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at time unit, a single node interacts with a randomly chosen neighbor. We perform a thorough study of the total number of interactions needed for all the nodes of the network to discover the rumor.

Distributed Stream Processing Systems: shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tuples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. In [25], [26] 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.

Load shedding is a technique employed by stream processing systems to handle unpredictable spikes in the input load whenever available computing resources are not adequately provisioned. A load shedder drops tuples to keep the input load below a critical threshold and thus avoid unbounded queuing and system trashing. In [38] we propose Load-Aware Shedding (LAS), a novel load shedding solution that, unlike previous works, does not rely neither on a pre-defined cost model nor on any assumption on the tuple execution duration. Leveraging sketches, LAS efficiently builds and maintains at runtime a cost model to estimate the execution duration of each tuple with small error bounds. This estimation enables a proactive load shedding of the input stream at any operator that aims at limiting queuing latencies while dropping as few tuples as possible. We provide a theoretical analysis. Furthermore, through an extensive practical evaluation based on simulations and a prototype, we evaluate its impact on stream processing applications, which validate the robustness and accuracy of LAS.

COAST Project-Team

5. New Results

5.1. Evaluation and Design of Consistency Maintenance Algorithms for Complex Data

Participants: Luc André, Quang Vinh Dang, Claudia-Lavinia Ignat, Gérald Oster, Pascal Urso.

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 amounts of contributors and a high velocity of changes. This year we continued our work on the evaluation of existing collaborative editing approaches and on the design of new algorithms that overcome limitations of state of the art ones. We designed new optimistic replication algorithms for maintaining consistency for complex data such as wikis and strings and we evaluated existing algorithms in large scale settings.

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 [5]. Our merging solution is based on an operational transformation approach for which we defined operations with high-level semantics 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.

Over the last years we designed a CRDT-based consistency maintenance algorithm for strings [20] for peer-to-peer large scale collaboration that is used by our MUTE collaborative editor which will be integrated in the virtual desktop of the OpenPaaS::NG project. This algorithm called LogootSplit can be seen as an extension for variable-sized elements (e.g. strings) of one of the first basic CRDT algorithms for unit elements (e.g. characters) proposed by our team called Logoot [32]. Its principles are general and can be applied to other basic CRDT algorithms. This year we proposed another algorithm for strings based on the RGA algorithm [9].

By means of simulations we measured the delays in popular real-time collaborative editing systems such as GoogleDocs and Etherpad [12] in terms of the number of users that edit a shared document and their typing frequency. Delays exist between the execution of one user's modification and the visibility of this modification to the other users. Such delays are in part fundamental to the network, as well as arising from the consistency maintenance algorithms and underlying architecture of collaborative editors. Results of this study support our team assertion that delay associated with conventional consistency maintenance algorithms will impede group performance.

5.2. Probabilistic Partial Orderings

Participants: Jordi Martori Adrian, Pascal Urso.

Ensuring reliable and ordered communication between computers usually requires acknowledgment messages. In systems with a high rate of broadcast communication, the cost of such acknowledgment messages can be large. We propose to use the causal ordering information required by some applications to detect and request missing messages. To circumscribe the number of unnecessary requests we combine local awareness and probabilistic methods. Our model allows us to obtain reliable communication within a latency equivalent to unordered communication and lower network usage than acknowledgment systems [18].

5.3. Computational Trust based on User Behavior

Participants: Quang Vinh Dang, Claudia-Lavinia Ignat.

We continued our investigation on computing a trust score for each user according to their behaviour during a collaborative task. Previously we proposed a contract-based collaboration model [31] where trust in users is established and adjusted based on their compliance to the contracts specified by the data owners when they share the data.

We continued this work by proposing an experimental design for testing the proposed trust-based collaboration model. We studied the trust game, a money exchange game that has been widely used in behavioural economics for studying trust and collaboration between humans. In this game, exchange of money is entirely attributable to the existence of trust between users. In the context of the trust game we proposed a trust metric that reflects user behaviours during the collaboration [10]. This metric is robust against fluctuating user behaviour. Our trust metric is the first one that was proposed in the context of the trust game in order to predict user behaviour.

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 document content. As an initial work in this direction we investigated how to automatically assess the quality of Wikipedia articles in order to guide readers towards high quality articles and to suggest to authors which articles need to be improved. In this context we proposed two automatic assessment methods of the quality of Wikipedia articles. In the first approach we introduced readability features for a better prediction of quality [11]. The second approach is based on a deep-learning mechanism that automatically learns features from document contents rather than manually defining them [13], [4].

5.4. A model to secure collaborative resources within Enterprise Social Networks

Participants: Ahmed Bouchami, Olivier Perrin.

Enterprise social networks (ESN) are collaborative environments that raise major challenges to secure them. In his thesis [2], Ahmed Bouchami addressed the problem of authentication of digital identities within collaborative communities. He proposed an interoperable architecture for managing federated authentication, thus allowing each enterprise to preserve its (own) authentication mechanism and each principal to perform a single sign on authentication regarding different enterprises. He also proposed access control management. His flexible access control model is based on a set of identity attributes, and a formal language based on temporal logic. This model allows for checking the consistency of the policies defined. with the model.

Last, the access control system offers the ability to control the user-centric sharing policies through policies based on a risk management mechanism, which makes the access control mechanism dynamic. The risk mechanism is based on the NIST's risk definition with an alignment with a set of parameters that include access control in the ESN context. More precisely, the dynamic risk management includes, the collaborative resource's importance, the authentication system's vulnerabilities and trust level reflected through the behavior of each collaborative actor. On this latter aspect of trust, a reputation score is computed using the history of collaborative interactions of each subject of the collaborative environment. Finally, a prototype is available and was demonstrated within the OpenPaaS ESN project.

5.5. Risk management for the deployment of a business process in a multi-cloud context

Participants: Amina Ahmed Nacer, Claude Godart, Elio Goettelmann, Samir Youcef.

The lack of trust in cloud organizations is often seen as obstacle 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).
- At the framework level, we have integrated the ability to integrate fake BP fragments in the objective to increase the obfuscation of a deployed BP logic [15].

5.6. Cloud Provisioning for Elastic BPM

Participants: François Charoy, Samir Youcef, Guillaume Rosinosky.

Even though the cloud computing paradigm has proven benefits, it faces a serious problem that can compromise its commercial success. It concerns the lack of an efficient approach for using optimally the available resources. For this, several approaches have been proposed [29]. However, they suffer from several shortcomings. Often only one objective is taken into account, expressing all operations in terms of cost. Furthermore, business processes should be insured with elasticity and multi-tenancy mechanism while adjusting the available resources to the dynamic load distribution. We proposed to optimize two conflicting objectives, namely the number of migrations of tenants and the cost incurred using 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 [16]. In order to secure the scientific value of our findings we have set up a experimentation infrastructure for making repeatable experiments on the Cloud [17]

5.7. Orchestration of crowdsourcing activities

Participants: François Charoy, Kahina Bessai.

Crowdsourcing is an important paradigm in human problem solving using the Web. When they face a workload outburst, businesses may choose to outsource some or all of their process tasks to the crowd in order to maintain the quality of service promised for their customers. This may occur in situations like crisis management, when organizations are overloaded by a sudden event breakout. These tasks are generally difficult to implement as solution based on software service only. So, the use of crowdsourcing platform seems enticing. To ensure efficient and wise use of resources, methods assisting decision making need to be developed whose aim is to assist businesses in choosing the most knowledgeable workers. We addressed the resource allocation problem in crisis context by defining a delegation approach based on crowdsourcing as resource provider. We introduce a mathematical model for business process execution in crowd-sourcing context and an exact optimization algorithm. As the problem addressed is NP-complete, we proposed a more efficient algorithm that we validated through simulation [7]. Furthermore, to overcome the limitations of existing works we take into account the fact that business process tasks are ordered while optimizing the overall execution time of a given business process instance under budget constraint. We used a synthetic crowd model or valitation. We have also defined a model to validate our work for geo-crowdsourcing activities [8].

CTRL-A Team

7. New Results

7.1. Design and programming

7.1.1. Component-based approaches

Participants: Gwenaël Delaval, Eric Rutten.

Architecting in the context of variability has become a real need in today's software development. Modern software systems and their architecture must adapt dynamically to events coming from the environment (e.g., workload requested by users, changes in functionality) and the execution platform (e.g., resource availability). Component-based architectures have shown to be very suited for self-adaptation especially with their dynamical reconfiguration capabilities. However, existing solutions for reconfiguration often rely on low level, imperative, and non formal languages. We have defined Ctrl-F, a domain-specific language whose objective is to provide high-level support for describing adaptation behaviors and policies in component-based architectures. It relies on reactive programming for formal verification and control of reconfigurations. We integrate Ctrl-F with the FraSCAti Service Component Architecture middleware platform, and apply it to the Znn.com self-adaptive case study

We have obtained new results in the application of modular controller synthesis and BZR compilation integrated in Ctrl-F, in order to attack issues in scalability, and reusability. We are also considering integration at the DSL level of expressivity extensions, for which the compilation and controller synhesis is relying on the ReaX tool developed at Inria Rennes, in the Sumo team.

7.1.2. Rule-based systems

Participants: Adja Sylla, Eric Rutten.

We are starting a cooperation with CEA LETI/DACLE on the topic of a high-level language for safe rule-based programming in the LINC platform. The general context is that of the runtime redeployment of distributed applications, for example managing smart buildings. Motivations for redeployment can be diverse: load balancing, energy saving, upgrading, or fault tolerance. Redeployment involves changing the set of components in presence, or migrating them. The basic functionalities enabling to start, stop, migrate, or clone components, and the control managing their safe coordination, will have to be designed in the LINC middleware developed at CEA.

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. A case study taken in the field of building automation is treated to illustrate the proposed approach [18].

The PhD of Adja Sylla at CEA on this topic is co-advised with F. Pacull and M. Louvel.

7.2. Infrastructure-level support

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 [23], [22].

7.2.1. Autonomic Cloud and Big-Data systems

Participants: Soguy Mak Kare Gueye, Gwenaël Delaval, Eric Rutten.

Complex computing systems are increasingly self-adaptive, with an autonomic computing approach for their administration. Real systems require the co-existence of multiple autonomic management loops, each complex to design. However their uncoordinated co-existence leads to performance degradation and possibly to inconsistency. There is a need for methodological supports facilitating the coordination of multiple autonomic managers. To tackle this problem, we take a global view and underscore that Autonomic Management Systems (AMS) are intrinsically reactive, as they react to flows of monitoring data by emitting flows of reconfiguration actions. Therefore we propose a new approach for the design of AMSs, based on synchronous programming and discrete controller synthesis techniques. They provide us with high-level languages for modeling the system to manage, as well as means for statically guaranteeing the absence of logical coordination problems. Hence, they suit our main contribution, which is to obtain guarantees at design time about the absence of logical inconsistencies in the taken decisions. We detail our approach, illustrate it by designing an AMS for a realistic multi-tier application, and evaluate its practicality with an implementation [16].

We addressed these problems in the context of follow-ups of the ANR project Ctrl-Green, in cooperation with LIG (N. de Palma) in the framework of the PhD of S. Gueye [17] and the post-doc of N. Berthier.

7.2.2. Reconfiguration control in DPR FPGA

Participants: Soguy Mak Kare Gueye, Eric Rutten.

Dynamically reconfigurable hardware has been identified as a promising solution for the design of energy efficient embedded systems. However, its adoption is limited by the costly design effort including verification and validation, which is even more complex than for non dynamically reconfigurable systems. We worked on this topic in the context of a ensign environment, developed in the framework of the ANR project Famous, in cooperation with LabSticc in Lorient and Inria Lille (DaRT team). We proposed a tool-supported formal method to automatically design a correct-by-construction control of the reconfiguration. By representing system behaviors with automata, we exploit automated algorithms to synthesize controllers that safely enforce reconfiguration strategies formulated as properties to be satisfied by control. We design generic modeling patterns for a class of reconfigurable architectures, taking into account both hardware architecture and applications, as well as relevant control objectives. We validate our approach on two case studies implemented on FPGAs [3].

We are currently valorizing results in more publications [15], and extending the use of control techniques by evaluating the new tool ReaX developed at Inria Rennes (Sumo).

We are starting a new ANR project called HPeC, within which some of these topics will be extended, especially regarding hierarchical and modular control, and logico-numeric aspects.

7.2.3. Autonomic memory management in HPC

Participants: Naweiluo Zhou, Gwenaël Delaval, Bogdan Robu, Eric Rutten.

Parallel programs need to manage the time trade-off between synchronization and computation. A high parallelism may decrease computing time but meanwhile increase synchronization cost among threads. Software Transactional Memory (STM) has emerged as a promising technique, which bypasses locks, to address synchronization issues through transactions. A way to reduce conflicts is by adjusting the parallelism, as a suitable parallelism can maximize program performance. However, there is no universal rule to decide the best parallelism for a program from an offline view. Furthermore, an offline tuning is costly and error-prone. Hence, it becomes necessary to adopt a dynamical tuning-configuration strategy to better manage a STM system. Autonomic control techniques begin to receive attention in computing systems recently. Control technologies offer designers a framework of methods and techniques to build autonomic systems with well-mastered behaviors. The key idea of autonomic control is to implement feedback control loops to design safe, efficient and predictable controllers, which enable monitoring and adjusting controlled systems dynamically while keeping overhead low. We propose to design feedback control loops to automate the choice of parallelism

at runtime and diminish program execution time [20], [24], [21]. It is then combined with another objective related to Thread Mapping Control [19]

In the context of the action-team HPES of the Labex Persyval-lab [0] (see 9.1), this work is performed in cooperation with LIG (J.F. Méhaut) in the framework of the PhD of N. Zhou [14].

7.2.4. Control of smart environments

Participants: Adja Sylla, Armando Ochoa, Eric Rutten, Stéphane Mocanu.

7.2.4.1. A service-oriented approach to smart home applications control with reactive programming

The need for adaptability in pervasive computing is growing, driven in part by the increasing number and variety of communication devices. In autonomic applications, however, the control architecture frequently becomes itself a complex system that needs to be adapted. Autonomic applications are often composed of multiple control loops ? each addressing a specific aspect ? whose execution needs to be coordinated for efficient and correct administration. We therefore propose to investigate the use of reactive control models with events and states to coordinate autonomic loops in service-oriented architectures. In this work, we illustrate our approach by integrating a controller based on discrete controller synthesis in an autonomic pervasive environment. The role of the controller is to influence the service-binding criteria of multiple control loops, while respecting logical constraints. In particular, we consider reconfiguration operations of known and dynamic service sets. This work constituted the M2R internship of Armando Ochoa, and was performed in cooperation with the Adele team at LIG, co-advised by E. Rutten and V. Lestideau, in the framework of the Labex Persyval-lab project CASE.

Another activity in this topic was the M2R internship of Ronak Feizimirkhani, co-advised by S. Mocanu and V. Lestideau, The context is the development of an application for a smart home in which automation devices are connected through a wireless communication protocol, Z-Wave, and controlled by a central controller, USB plug in. This involves methods and tools to design fail-safe controllers for autonomic, adaptive, reconfigurable computing systems by combining Computer Science and Control Theory techniques. For this purpose, it is necessary to access required information over the network, derive out a simplified model of the physical network, and then link it to the User interface application. According to the information achieved, there will be an estimation of the network diagnostics to find some probable solutions for. The final application is in a user media to do installing, maintaining or even optimizing the network and devices.

7.2.4.2. Rule-based specification of smart environments control

In the context of IoT applications like mart home environments, the rules for programming in the LINC framework are used as a flexible tool to govern the relations between sensors and actuators. Runtime coordination and formal analysis becomes a necessity to avoid side effects mainly when applications are critical. In cooperation with CEA LETI/DACLE, we are working on a case study for safe applications development in IoT and smart home environments.

New results from Section 7.1.2 are applied in case studies regarding smart environments (offices or homes) [18].

[0] https://persyval-lab.org/en/sites/hpes

MIMOVE Team

7. New Results

7.1. Introduction

MiMove's research activities in 2016 have focused on a set of areas directly related to the team's research topics. Hence, we have worked on QoS for Emergent Mobile Systems (§ 7.2) in relation to our research topic regarding Emergent Mobile Distributed Systems (§ 3.2). Furthermore, our effort on Ambiciti (§ 7.3) is linked to our research on Mobile Social Crowd-sensing (§ 3.4). Still in the context of Mobile Social Crowd-sensing (§ 3.4), we have developed AppCivist-PB (§ 7.4) related to our interest in social applications aiming to actively involve citizens (see § 4.1); this is further linked to our research on composition of Emergent Mobile Distributed Systems (§ 3.2). Finally, we have worked on the Fiesta-IoT ontology (§ 7.5) and on the Sarathi platform (§ 7.6), related to our research on both Large-scale Mobile Sensing & Actuation (§ 3.3) and Mobile Social Crowd-sensing (§ 3.4).

7.2. QoS for Emergent Mobile Systems

Participants: Georgios Bouloukakis, Nikolaos Georgantas, Siddhartha Dutta, Valérie Issarny.

With the emergence of Future Internet applications that connect web services, sensor-actuator networks and service feeds into open, dynamic, mobile choreographies, heterogeneity support of interaction paradigms is of critical importance. Heterogeneous interactions can be abstractly represented by client-server, publish/subscribe, tuple space and data streaming middleware connectors that are interconnected via bridging mechanisms providing interoperability among the choreography peers. We make use of the *eVolution Service Bus (VSB)* (see § 6.2) as the connector enabling interoperability among heterogeneous choreography participants [15]. VSB models interactions among peers through generic *post* and *get* operations that represent peer behavior with varying time/space coupling.

Within this context, we study end-to-end Quality of Service (QoS) properties of choreographies, where in particular we focus on the effect of middleware interactions on QoS. We consider both homogeneous and heterogeneous (via VSB) interactions. We report in the following our results in two complementary directions:

- Choreography peers deployed in mobile environments are typically characterized by intermittent connectivity and asynchronous sending/reception of data. In such environments, it is essential to guarantee acceptable levels of timeliness between sending and receiving mobile users. In order to provide QoS guarantees in different application scenarios and contexts, it is necessary to model the system performance by incorporating the intermittent connectivity. Queueing Network Models (QNMs) offer a simple modeling environment, which can be used to represent various application scenarios, and provide accurate analytical solutions for performance metrics, such as system response time. We provide an analytical solution regarding the end-to-end response time between users sending and receiving data by modeling the intermittent connectivity of mobile users with QNMs. We utilize the publish/subscribe middleware as the underlying communication infrastructure for the mobile users. To represent the user's connections/disconnections, we model and solve analytically an ON/OFF queueing system by applying a mean value approach. Finally, we validate our model using simulations with real-world workload traces. 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% for a variety of scenarios [16].
- Based on the QoS models and analyses outlined in the previous paragraph, we go one step further towards realistic QoS modeling and analysis of choreographies integrating heterogeneous interaction paradigms. We introduce QoS modeling patterns that correspond to each one of the interaction paradigms – client-server, publish/subscribe, tuple space and data streaming – and

for different interaction styles – one way, two way synchronous, two way asynchronous. Our patterns rely on Queueing Network Models (QNMs) and represent the following characteristics of choreography peers and their middleware protocols: (i) reliable or unreliable interactions supported by the middleware and underlying transport layers; (ii) application-level (user) and middleware-level disconnections; (iii) application-level and middleware-level buffering of messages with finite capacity; (iv) limited lifetime of messages; and (v) timing of synchronous interactions. These QoS patterns enable the analysis and evaluation of the performance and success rates characterizing the modeled interactions. By combining several QoS patterns, we can further evaluate the end-to-end QoS of choreography interactions among heterogeneous peers. Based on our QoS models, we statistically analyze through simulations the effects on QoS when varying the parameters found in (i) to (v). We can also in this way evaluate the interconnection effectiveness, i.e., the degree of mapping of QoS semantics and expectations, when interconnecting heterogeneous choreography peers.

7.3. Mobile Phone Sensing Middleware for Urban Pollution Monitoring

Participants: Valerie Issarny, Cong Kinh Nguyen, Pierre-Guillaume Raverdy, Fadwa Rebhi.

Mobile Phone Sensing (MPS) is a powerful solution for massive-scale sensing at low cost. The ubiquity of phones together with the rich set of sensors that they increasingly embed make mobile phones the devices of choice to sense our environment. Further, thanks to the – even sometimes unconscious – participation of people, MPS allows for leveraging both quantitative and qualitative sensing. And, still thanks to the participation of people who are moving across space, mobile phones may conveniently act as opportunistic proxies for the sensors in their communication range, which includes the fast developing wearables.

However, despite the numerous research work since the end 2000s, MPS keeps raising key challenges among which: How to make MPS resource-efficient? How to mitigate mobile sensing heterogeneities? How to involve and leverage the crowd? How to leverage prior experiences?

Addressing the above MPS challenges primarily lies in taming the high heterogeneity not only of the computing system but also the crowd. The latter introduces a new dimension compared to traditional middleware research that has been concentrating on overcoming the heterogeneities of the computing infrastructure. In order to tackle these two dimensions together, we have been conducting a large scale empirical study in cooperation with the city of Paris (see http://tinyurl.com/soundcity-paris). Our experiment revolves around the public release of a MPS app for noise pollution monitoring that is built upon our dedicated mobile crowd-sensing middleware. Building on the Paris experiment, we systematically studied the influence of resource-efficiency and sensing accuracy on the effectiveness of the crowd participation [18]. In a complementary way, we analyzed user participation across time, so as to derive participation patterns that MPS middleware and application design may leverage.

Key take-away for MPS middleware and application design following our analysis includes:

- While contributors exhibit high heterogeneity regarding the accuracy of their sensors, they overall exhibit similar patterns. Location accuracy leads to discard about 60% of the observations and most observations are in the $[20 - 50]$ meters accuracy range. Noise sensing accuracy varies but calibration may be achieved per model rather than per device; calibration may then combine a number of techniques from comparison using a high-quality reference sensor to automated techniques leveraging assimilation and machine learning. Although our experiment is focused on noise sensing, we may expect similar results for other physical sensors. Overall, MPS allows collecting and assimilating relevant observations/measures. Still, the number of contributed measures by the MPS system needs to be high enough to overcome the low accuracy of the phone sensors.
- Although not specifically related to heterogeneity, energy efficiency is critical for the adoption of MPS. Our study confirms that energy-delay tradeoffs is a valuable approach; hence, the middleware must enable the buffering of the observations while the frequency of the transfers must be tuned by the application. Still, we notice that 30% of the observations reach the server after 2 hours even when observations are not buffered and are sent every 5mns, which indicates long periods of disconnection.

Hence, if the timeliness of the observation is critical, then participatory sensing is most likely the approach to follow to ensure that the user is conscious about the sensing and activates appropriate network connection.

- The heterogeneity of the contributing crowd is obvious. However, it turns out to be an asset rather than a shortcoming of MPS. Indeed, the crowd overall exhibits similar contribution patterns across time. However, in the detail, each individual has different contribution patterns. This allows for the collection of complementary contributions over the whole day.
- The users appear to be still most of the time, while the user's activity cannot be qualified for 20% of the observations. This should be accounted for in the design of mobility-dependent MPS.
- One design issue that arises for MPS is whether to promote participatory or opportunistic sensing. It is our belief that a system (and thus supporting app) must support both. This enables to collect as many observations as possible from a large diversity of people, while participatory sensing guarantees contributions of higher quality.

7.4. Computer-mediated Social Communication Interoperability

Participants: Rafael Angarita, Nikolaos Georgantas, Valerie Issarny, Cristhian Parra Trepowski, Christelle Rohaut.

People increasingly rely on computer-mediated communication for their social interactions. This is a direct consequence of the global reach of the Internet combined with the massive adoption of social media and mobile technologies that make it easy for people to view, create and share information within their communities almost anywhere, anytime. The success of social media has further led – and is still leading – to the introduction of a large diversity of social communication services (e.g., Skype, Facebook, Google Plus, Telegram, Instagram, WhatsApp, Twitter, Slack, ...). These services differ according to the types of communities and interactions they primarily aim at supporting. However, existing services are not orthogonal and users ultimately adopt one service rather than another based on their personal experience. As a result, users who share similar interests from a social perspective may not be able to interact in a computer-mediated social sphere because they adopt different technologies. This is particularly exacerbated by the fact that the latest social media are proprietary services that offer an increasingly rich set of functionalities, and the function of one service does not easily translate -both socially and technically- into the function of another. As an illustration, compare the early and primitive social media that is the Email with the richer social network technology. Protocols associated with the former are rather simple and email communication between any two individuals is now trivial, independent of the mail servers used at both ends. On the other hand, protocols associated with today's social networks involve complex interaction processes, which prevent communication across social networks.

The above issue is no different than the long-standing issue of interoperability in distributed computing systems, which require to mediate (or translate) the protocols run by the interacting parties for them to be able to exchange meaningful messages and coordinate. And, while interoperability in the early days of distributed systems was essentially relying on the definition of standards, the increasing complexity and diversity of networked systems has led to the introduction of various interoperability solutions, among which the (Enterprise) Service Bus paradigm.

In the above context, we have specifically introduced the "*social communication bus*" paradigm so as to allow interoperability across computer-mediated social communication protocols. Our work is motivated by our research effort within the AppCivist project. AppCivist provides a software platform for participatory democracy that leverages the reach of the Internet and the powers of computation to enhance the experience and efficacy of civic participation. Its first instance, AppCivist-PB, targets participatory budgeting, an exemplary process of participatory democracy that let citizens prepare and select projects to be implemented with public funds by their cities [17]. For city-wide engagement, AppCivist-PB must enable citizens to participate with the Internet-based communication services they are the most confortable with. The need for interoperability in this context is indeed paramount since the idea is to include people in the participatory processes without leaving anyone behind. This has led us to revisit the service bus paradigm for the sake of social communication across communities, so as to gather together the many communities of our cities.

Our contributions span:

- *Social communication paradigm:* Based on the survey of the various forms of computer-mediated social communication supported by today's software services and tools, we have derived how the approaches to middleware interoperability may apply to social communication interoperability.
- *Social Communication Bus architecture:* We leverage the VSB bus (see § 6.2) that supports interoperability across interaction paradigms as opposed to interoperability across heterogeneous middleware protocols implementing the same paradigm. The proposed bus architecture features the traditional concepts of bus protocols and binding components, but those are customized for the sake of social interaction whose coupling differs along the social and presence dimensions.
- *Social Communication Bus instance for participatory democracy:* We have refined our bus architecture, introducing the Social-MQ implementation that leverages the RabbitMQ message broker. The resulting implementation has been integrated within the AppCivist-PB platform for evaluation.

In order to inform the further study of the "Social Communication Bus" paradigm, we have analyzed existing practices and supporting technologies promoting citizen collaboration. In relation with our work on the AppCivist-PB platform, our study has concentrated on Participatory Budgeting (PB) campaigns, with a special focus on US-related initiatives, as a mean to understand the current and future design space of ICT for participatory democracy. We then derived new design opportunities for ICT to facilitate citizen collaboration in the PB process, and by extension, to reflect on how these technologies could better foster deliberative decision-making at a scale that is both small and large.

This research is carried out in collaboration with the Social Apps Lab at CITRIS at UC Berkeley in the context of CityLab@Inria and Inria@SiliconValley.

7.5. FIESTA-IoT Ontology: Semantic Model for Federation & Interoperability among Platforms

Participants: Rachit Agarwal, Valérie Issarny, Nikolaos Georgantas.

Plethora of heterogeneous data is being generated and made available by diverse platforms. Such platforms can be those that are formed by the use of mobile application that act as interface between sensing devices and storage or between users and storage. The diversity and openness in the data generated isolate platforms and lead to interoperability issues between platforms, where much work has to be done in order to ensure compatibility. One has to understand the other's format, parse different data formats, and create the mapping between different data formats. One method to accomplish this interoperability is by attaching semantics to this data. Semantics provides meaning to the data and helps in (a) achieving common understanding and (b) performing analysis and reasoning. Many IoT-related semantic models [0] propose interoperability but have many issues like: observation graph is missing, are highly domain specific, and do not follow best practices. In order to address the above, we focused our research on: the identification of a unified semantic model that addresses the above, creation of a prototype application, and identification of guidelines for storing semantic data [13]. We report our following key results:

- *State of art survey of semantic models that are available in literature in the domain of the Internet of Things*: This survey gave us required knowledge needed for the semantic model from which concepts can be reused to create a unified ontology. This helps the semantic community by not overloading the domain with concepts similar to already existing concepts, and allows us to reuse concepts as much as possible. We identified that recent trends show more and more use of the SSN [35] and oneM2M [67] ontologies. However, these models are currently far from being able to address observation-related issues and lack domain taxonomy.
- *Unified semantic model for enabling interoperability and federation of testbeds*: Based on the analysis of the concepts from various onotologies identified, we unify specific concepts from these identified ontologies into one ontology. These ontologies being: SSN, oneM2M, IoT-lite [27],

[0] http://sensormeasurement.appspot.com/?p=ontologies

WGS84 [0], DUL [0], TIME [0] and M3-lite taxonomy (created as a part of this research). Such unification gives our ontology the power to define meta data about the sensor that is producing the observation and the observation itself. The federation is achieved by the use of the taxonomy that each platform should follow.

- *Best practices to publish data based on the unified model*: In order to enable full interoperability, federation and usage of data, it is essential that best practices are followed while storing the data based on the unified model. We identify various best practices which form our recommendations to the platform owners towards annotating the data with respect to the ontology. This is supported by a reference annotator that also acts as a guide for developers to publish data.

These above-mentioned results are currently applied in the frame of the EU funded H2020 FIESTA-IoT project (see § 8.2.1.2).

7.6. Sarathi: A Platform for Personalized Mobility Service for Urban Travellers

Participants: Rachit Agarwal, Garvita Bajaj, Georgios Bouloukakis, Valérie Issarny, Nikolaos Georgantas.

Thanks to the increased abundance of mobile phones, the recent field of mobile participatory sensing could be leveraged towards providing a more fine-grained and up-to-date view of a city's transportation system. Thus, in order to address problems like dynamicity (unexpected faults, stoppages, etc.) and unexpected load (number of people using the transportation), etc., in different societal contexts of France and India, we aimed to produce a middleware platform called "*Sarathi*" that is enriched with personalized mobility services for urban travelers and is evaluated via real-life demonstrators. Towards this, the key results include:

- *Identification of System Architecture* [14]: We first identify requirements for our system that would satisfy the objectives. The identified requirements are then mapped to specific components that would carry out specific tasks. A client-server system architecture is then created by connecting the identified components. Some components that we identified are: UI component that would run at the client side, recommendation system and knowledgebase component that would run at the server, and a communication component that would ensure communication of the client with the server. To realise these components, we also identify tools and techniques that would ensure best runtime performance.
- *Modeling Passenger convenience in Metro transit* [20]: This effort builds upon existing research in the area, studied during our joint survey of related work, and applies the work to the context of the Paris and New Delhi metro system. This work captures 'personalized' experience of passengers during a multi-leg journey and models the convenience for commuters. A leg in a journey is defined as a segment of a journey traveled on a metro line. The work proposes a mathematical model for commuter convenience and validates it using data collected from metro commuters. The convenience model uses 3 convenience measures namely *seat availability*, *wait time* and *comfort*. The work also aims to identify the best mobile interaction paradigm for enabling timely data collection and dissemination and outlines a middleware architecture to achieve this (aiming at acceptable response times for mobile apps).
- *Mobile Application*: An Android application called *MetroCognition* for gathering commuters convenience rating during their metro transit based on the three above described measures has been developed, deployed and made available on Google Play Store [0] for beta testing.

[0] https://www.w3.org/2003/01/geo/
[0] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl
[0] https://www.w3.org/TR/owl-time/
[0] https://play.google.com/apps/testing/edu.sarathi.metroCognition

MYRIADS Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Heterogeneous Resource Management

Participants: Baptiste Goupille-Lescar, Ancuta Iordache, Christine Morin, Manh Linh Pham, Nikos Parlavantzas, Guillaume Pierre, Arnab Sinha.

7.1.1.1. High performance in the cloud with FPGA virtualization

Participants: Ancuta Iordache, Guillaume Pierre.

Cloud platforms are becoming increasingly heterogeneous, with the availability of large numbers of virtual machine instance types as well as accelerator devices such as GPUs. In collaboration with Maxeler technologies, we have proposed a technique to virtualize FPGAs and make them available as first-class high-performance computation devices in the cloud [24]. The increasing variety of computation, storage and networking resources in the cloud is an opportunity for adjusting the provisioned resources to the individual needs of each application, but making an informed choice is extremely difficult. We therefore proposed application profiling techniques which can automatically identify the configuration which provides the best performance/cost tradeoff [49]. These two results were developed as part of the HARNESS European project, and they constitute Anca Iordache's PhD thesis [50]. FPGA virtualization is being further developed by Maxeler technologies toward commercial exploitation, and application profiling has been integrated in the open-source ConPaaS platform.

7.1.1.2. Multi-cloud application execution

Participants: Manh Linh Pham, Nikos Parlavantzas, Arnab Sinha.

Within the PaaSage European project, we improved and extended the Adapter subsystem, the part of the PaaSage platform that dynamically adapts the application deployment to changes in current runtime conditions [45]. Specifically, we added full support for causal connection between the running system and the runtime model and extended the plan validation functionality to use historical reconfiguration information. Moreover, we assisted industrial PaaSage partners with applying the PaaSage platform in diverse business scenarios.

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

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

In the context of our collaboration with Thales Research and Technology, 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 2016, we focused on characterising the targeted applications and platforms and developing a simulator in order to explore relevant resource management solutions. This work is performed in the context of Baptiste Goupille-Lescar's PhD work.

7.1.2. Distributed Cloud Computing

Participants: Nikos Parlavantzas, Jean-Louis Pazat, Guillaume Pierre, Genc Tato, Cédric Tedeschi, Alexandre Van Kempen.

7.1.2.1. Application self-optimization in multi-cloud environments

Participant: Nikos Parlavantzas.

Current approaches to application adaptation in multi-cloud environments are typically static, platform dependent, complex, and error prone. To address these limitations, we are combining the use of software product lines (SPLs) with models@run-time 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 focuses on the development of an SPL-based framework supporting initial cloud configuration as well as proactive, dynamic adaptation in a systematic, platform-independent way. The evaluation of this framework is currently in progress.

7.1.2.2. Edge clouds

Participants: Guillaume Pierre, Genc Tato, Cédric Tedeschi, Alexandre Van Kempen.

Mobile edge cloud computing aims to deploy cloud resources even closer to the end users, typically within mobile network access points. This is useful for hyper-interactive applications such as augmented reality which demand ultra-low network latencies (2-5 ms) between the end-user device and the cloud instances serving it. In contrast, current mobile networks exhibit network latencies in the order of 50-150 ms between the device and any cloud. We extended the ConPaaS open-source cloud platform to support the deployment of cloud applications in a distributed set of Raspberry Pi machines: instead of reaching the cloud through a wide-area network, in this setup each cloud node is also equipped with a wifi hotspot which allows local users to access it directly [53]. This work is ongoing, and a paper on this topic is currently being reviewed.

Getting closer to the edge user can be done through provisioning computing resources in Points of Presence (PoPs) within the telco's backbone network. The Discovery project [52] aims at revisiting the OpenStack Cloud stack to allow to disperse several smaller cloud facilities and connect them together to make them appear as a single Cloud entity. Genc Tato's PhD aims at proposing the building blocks on top of such an infrastructure to abstract out the network, route queries, store and retrieve objects (VMs and data). We have devised an overlay network to support such functionalities keeping in mind to maximise the laziness of the maintenance protocol to avoid any useless cost. A paper is being written on the subject.

7.1.2.3. Community Clouds

Participant: Jean-Louis Pazat.

Hosting services on an edge infrastructure based on devices owned and operated by end-users may be interesting for serving a community of users. However, these devices (such as internet boxes, disks or small computers have heterogeneous capabilities and no guaranteed availability. It is therefore challenging to ensure to the guest application a minimal hosting service level, like availability or Quality of Service. The management of the hosting service should adapt to the characteristics of the infrastructure. We are designing an architecture for a middleware capable of adapting the deployment of services on edge devices to ensure a given Quality of Service to access the service. While the middleware requires a minimal knowledge of the underlying infrastructure, its adaptation decisions are based on the feedbacks of users of the deployed service, like measured network latency. The environment relies on the use of micro-services which are composed to build the end-user services. This allows many adaptation strategies to adapt the system during run-time.

7.1.3. Scaling workflows with GinFlow

Participants: Matthieu Simonin, Cédric Tedeschi.

In 2016, we deployed GinFlow over 800 cores of the Grid'5000 platform, running Montage workflows comprising 118 tasks, and artificial workflows made of more than 3000 tasks. The ability of GinFlow to support adaptation and versioning of workflow with seamless transitions between workflow alternatives at runtime has been validated experimentally and presented on the Inria booth at SuperComputing in November 2016. These results have been presented at the IPDPS conference [32], and have been submitted to a journal special issue on workflows.

7.2. Greening Clouds

7.2.1. Energy Models

Participants: Yvon Jégou, Anne-Cécile Orgerie, Edouard Outin, Jean-Louis Pazat, Martin Quinson.

Simulating the impact of DVFS within SimGrid Simulation is a a popular approach for studying the performance of HPC applications in a variety of scenarios. However, simulators do not typically provide insights on the energy consumption of the simulated platforms. The goal of this ongoing work is to enable energy-aware experimentation within the SimGrid simulation toolkit, by introducing a model of energy consumption for computing applications making use of Dynamic Voltage and Frequency Scaling (DVFS) techniques.

Simulating Energy Consumption of Wired Networks In this work, we aim at simulating the energy consumption of wired networks which receive little attention in the Cloud computing community even though they represent key elements of these distributed architectures. To this end, we are contributing to the well-known open-source simulator ns3 by developing an energy consumption module named ECOFEN. This simulator embeds green levers: low power idle (IEEE 802.3az) and adaptive link rate. An article is currently under review on this topic.

Multicriteria scheduling for large-scale HPC environments Energy consumption is one of the main limiting factor for the design and deployment of large scale numerical infrastructures. The road towards "Sustainable Exascale" is a challenge with a target of 50 Gflops per watt. As platforms become more and more heterogeneous (co-processors, GPUs, low power processors...), an efficient scheduling of applications and services at large scale remains a challenge. In this context, we explore a multicriteria scheduling model and framework for large scale HPC systems. This work is done in collaboration with ROMA and Avalon teams from LIP in Lyon [29], [37].

Dynamic resource management for energy-efficiency The B-Com project, a joint private/public focusing on transfer, targets the design and the implementation of Watcher, a software module used to optimize an OpenStack cloud (in terms of performance, storage optimization or energy savings). This Software module is in the "Big Tent" software development process of OpenStack. In cooperation with Olivier Barais (Diverse Inria Team), we focus on dynamic management of cloud resources for energy-efficiency. Our approach relies on machine learning techniques, models@run-time and dynamic adaptation, and is intended to be included in Watcher. At regular intervals of time, we optimize the use of cloud resources by checking if a better placement of Virtual Machines on physical resources can be achieved, taking into account the migration cost. To achieve this, we have an energy model of the resources which is regularly updated using machine learning techniques that helps optimization algorithms to check if a better configuration can be reached energy-wise. This year we worked on the evaluation of the energy model [28].

7.2.2. Involving users in Energy Saving

Participants: Deborah Agarwal, Ismael Cuadrado Cordero, David Guyon, Christine Morin, Anne-Cécile Orgerie.

Energy-efficient cloud elasticity for data-driven applications Data centers hosting cloud systems consume enormous amounts of energy. Reducing this consumption becomes an urgent challenge with the rapid growth of cloud utilization. An existing solution to lower this consumption is to turn off as many servers as possible, but these solutions do not involve the user as a main lever to save energy. We introduce a system that proposes to the user to run her application with degraded performance in order to promote a better consolidation and thus to turn off more servers. Experimentation results using the Montage workflow show promising outcomes [47], [48]. We also performed a simulation-based evaluation on how much an energy-aware cloud system could save in energy consumed depending on the proportion of users selecting a green execution mode. These results based on the simulation of two typical daily uses of a data center running 3 real scientific applications will be published in Euromicro PDP 2017.

Energy-efficient and network-aware resource allocation in Cloud infrastructures The ever-growing appetite of new applications for network resources leads to an unprecedented electricity bill, and for these bandwidth-hungry applications, networks can become a significant bottleneck. Towards this end, we proposed microclouds, a fully autonomous energy-efficient subnetwork of clients of the same service, designed to keep the greenest path between its node. This semi-decentralized PaaS architecture for real-time multiple-users applications geographically distributes the computation among the clients of the cloud, moving the

computation away from the datacenter to save energy - by shutting down or downgrading non utilized resources such as routers and switches, servers, etc. - and provides lower latencies for users. In this work, we have also analyzed the use of incentives for Mobile Clouds, and proposed a new auction system adapted to the high dynamism and heterogeneity of these systems [20], [19] [46].

7.2.3. Exploiting Renewable Energy in Datacenters

Participants: Sabbir Hasan Rochi, Yunbo Li, Anne-Cécile Orgerie, Jean-Louis Pazat.

Resource allocation in a Cloud partially powered by renewable energy sources We propose here to design a disruptive approach to Cloud resource management which takes advantage of renewable energy availability to perform opportunistic tasks. This Cloud receives a fixed amount of power from the regular electric Grid. This power allows it to run usual tasks. In addition, this Cloud is also connected to renewable energy sources (such as windmills or solar cells) and when these sources produce electricity, the Cloud can use it to run more tasks. The proposed resource management system integrates a prediction model to be able to forecast these extra-power periods of time in order to schedule more work during these periods. This work is done in collaboration with Ascola team from LINA in Nantes [44], [51][9].

Creating green-energy adaptivity awareness in SaaS application In addition to "green" resource allocation at the IaaS level in Datacenters, we think that users should be involved in "greening" their energy use (SaaS level). We propose that applications should have multiple "modes" of execution, each mode using a different level of energy and providing a different service level. For example, a B2C application may provide more or less recommandations . If this applcation can be dynamically swiched between these modes depending on the availability of green energy, the IaaS can optimize resource allocation better. To enforce his, we have designed green energy aware controllers.

This work is done in collaboration with Ascola team [23], [9].

7.3. Securing Clouds

7.3.1. Security monitoring in clouds

Participants: Jean Leon Cusinato, Anna Giannakou, Fergal Martin-Tricot, Christine Morin, Jean-Louis Pazat, Louis Rilling, Amir Teshome Wonjiga.

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 2016 we improved 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 particular, we validated that the approach is generic enough to handle signature-based NIDSs (support for Snort and Suricata was implemented) as well as event-based NIDSs (support for Bro was implemented). An experimental evaluation of SAIDS has also been started in order to submit a full paper for publication in 2017. Jean-Léon Cusinato contributed to this work during his master internship.

We also improved the AL-SAFE approach, that we proposed in 2015, and that secures an application-level firewall by isolating it from the customer virtual machine and makes it self-adaptable [36], [35]. In particular, we validated that the self-adaptation architecture introduced for SAIDS could be reused to address firewalls, and the prototype was improved to implement stateful filtering. Fergal Martin-Tricot contributed to this work during his master internship. We also evaluated AL-SAFE experimentally on the prototype as well as analytically regarding the security correctness. The design and the evaluation of AL-SAFE were published in the CloudCom 2016 conference [21].

Regarding SLA definition and enforcement, in 2016 we have studied a verification method to enable 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. A simple example of SLO is being used for this study, and further work should address more complete SLO regarding NIDSs. A prototype of the proposed verification method was implemented on OpenStack and Open vSwitch, and the NIDS software used is Snort. An evaluation of the verification method has been started and will include both experiments on the Grid'5000 platform and a correctness analysis. The design and evaluation of the verification method will be submitted in a full paper for publication in 2017.

7.3.2. Risk assessment in clouds

Participant: Christine Morin.

Attack graphs are leveraged in networks to exhibit the various scenarios available to compromise the system. They allow to uncover vulnerabilities chains exploitable by attackers based on network connectivity and vulnerabilities pre-requisites. In physical infrastructures, the acquisition of the topology has been vastly addressed in existing works with either passive or active discovery methods. Considering the Cloud context, in which virtualization attacks and virtual infrastructure dynamism are introduced, new methods need to be developed. We have designed a topology builder able to keep the topology and connectivity up to date in cloud environments. Based on the use of an IaaS cloud management system and a SDN (Software-Defined Networking) controller, our approach encompasses two steps: (i) when plugged into a running system, the topology builder retrieves the current topology and builds the associated connectivity: this represents the static topology and connectivity retrieval, in which we assume the network configuration to be fixed ; (ii) the topology builder listens to change events generated inside the infrastructure and within the SDN controller in order to update the topology and connectivity previously built: this represents the dynamic topology and connectivity retrieval. A prototype has been developed based on OpenStack cloud management system and ONOS SDN open source technologies. This work is carried out in the context of Pernelle Mensah's PhD thesis and in collaboration with Nokia and CIDRE Inria project-team.

7.4. Experimenting with Clouds

7.4.1. Simulation

Participants: Simon Bihel, Martin Quinson.

Providing better interfaces to the users for Cloud Studies. Aware that the current user interface is a impediment to the adoption of our framework by the scientific community, we tried to propose a new, simplified API through the internship of Simon Bihel this summer. We identified several use cases and usage scenario that relevant to our context, and started implementing the new interface that we will provide. This work is still under progress.

Production-ready simulator of large-scale distributed systems. We are currently involved in a complete reorganization of the SimGrid implementation. The goal is two-fold: first we want reduce the tool's learning curve to help beginners. At the same time, we want to normalize the tool's internals so that power users can modify it and/or script the kernel behavior easily. Eventually, we are targeting usages in production and teaching contexts. This long term overhaul is still underway.

7.4.2. Experimentation Testbed

Participants: Anirvan Basu, Julien Lefeuvre, David Margery, Pascal Morillon.

Providing ready to use scripts to deploy popular and complex stacks. The study of complex software stacks on Grid'5000 has always been possible due to the reconfigurability properties of the testbed. Nevertheless, for newcomers with little background in system administration, automating the deployment of these stacks on Grid'5000 has always proved difficult. In 2016, we have provided scripts, that users can fork on github to customise to their needs, to deploy OpenStack, Ceph, Hadoop over Ceph or Sparkle. These have been presented to users during the 2016 winter school.

7.4.3. Use cases

Participants: Deborah Agarwal, Yvon Jégou, Nikos Parlavantzas, Manh Linh Pham, Christine Morin, Kartik Sathyanarayanan, Arnab Sinha.

7.4.3.1. Experimental Evaluation of Data Stream Processing Frameworks

We worked on evaluating data stream processing environments deployed in clouds. We compared the throughput, latency and energy consumption of Spark Streaming, Storm and Heron real-time data processing environments executed on top of Linux clusters and on top of virtual clusters deployed on top of the OpenStack IaaS cloud. The preliminary evaluation was conducted using the word count application on the twitter data stream. All experiments were conducted on Grid'5000 experimentation platform. The experimental results are described in a technical report to be published in 2017. This work was carried out by Kartik Sathyanarayanan, a student intern in Myriads team in the framework of DALHIS associate team.

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

In our collaboration with Inra in the context of the Mihmes project, we worked on the design of decision tools to evaluate the epidemio-economic effectiveness of disease prevention and control strategies at the scales of the herd, the region and the supply chain. We developed a generic service-based framework to efficiently execute models of infection dynamics in a metapopulation of cattle herds on large-scale computing infrastructures. Our framework has been designed to execute complex regional models combining within-herds epidemiological models. The framework automatically distributes the simulation runs on multiple servers in a cluster and exploits the parallelism of the multicore servers. It relies on OpenMP for parallelizing simulation loops and deals with server heterogeneity and failures. We leveraged PaaSage software stack to deploy the framework on several IaaS clouds.

7.4.3.3. Mobile application for reliable collection of field data for Fluxnet

Critical to the interpretation of Fluxnet carbon flux data is the ancillary information and measurements taken at the tower sites. The submission and update of this data using excel sheets is difficult and error prone. In partnership with ICOS in the framework of DALHIS associate team, we are innovating the data submission and organization method through a responsive web User Interface able to run on desktop, mobile etc.; thus easing the data lookup and entry process from anywhere including the field sites. Continuing with our initial usability feedback experiences gathered last year on the application interface designs, we decided on the mobile application workflow for implementation. We developed a first prototype based on the PhoneGap [0] platform which provided the advantage of the same development code generating mobile application for IOS, Android and Windows platform simultaneously. The main functionality realized in the application prototype is that the user can download all the site data required by logging in through the application; and then view/edit them at the tower site (even in offline mode). The next logical step would be developing the synchronization and validation of data held locally in the application with the servers.

[0] http://phonegap.com/

REGAL Project-Team

6. New Results

6.1. Distributed Algorithms for Dynamic Networks and Fault Tolerance

Participants: Luciana Bezerra Arantes [correspondent], Sébastien Bouchart, Marjorie Bournat, Swan Dubois, Denis Jeanneau, Mohamed Hamza Kaaouachi, Sébastien Monnet, 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.

We address both system dynamic and fault tolerance through various aspects: (1) Fault Detection, (2) Self-Stabilization, and (3) Dynamic System Design. Our approach covers the whole spectrum from theory to experimentation. We design algorithms, prove them correct, implement them, and evaluate them within simulation plateforms.

6.1.1. Failure detection

Since 2013, we address both theoretical and practical aspects of failure detector. The failure detector (FD) abstraction has been used to solve agreement problems in asynchronous systems prone to crash failures, but so far it has mostly been used in static and complete networks. FDs are distributed oracles that provide processes with unreliable information on process failures, often in the form of a list of trusted process identities. In 2016 we obtain the following results.

We propose in [31] a new failure detector that expresses the confidence with regard to the system as a whole. Similarly to a reputation approach, it is possible to indicate the relative importance of each process of the system, while a threshold offers a degree of flexibility for failures and false suspicions. Performance evaluation results, based on real PlanetLab traces, confirm the degree of flexible of the failure detector. By logically organizing nodes in a distributed hypercube, denoted VCube, which dynamically re-organizes itself in case of node failures, detected by a hierarchical perfect failure, we have proposed a autonomic distributed quorum algorithm [35]. By replacing the perfect failure detector by another one that offers eventual strong completeness, we have presented in [33] a second autonomic reliable broadcast protocol.

In the context of large networks, we propose Internet Failure Detector Service (IFDS) [16] for processes running in the Internet on multiple autonomous systems. The failure detection service is adaptive, and can be easily integrated into applications that require configurable QoS guarantees. The service is based on monitors which are capable of providing global process state information through a SNMP MIB. Monitors at different networks communicate across the Internet using Web Services. The system was implemented and evaluated for monitored processes running both on single LAN and on PlanetLab. Experimental results are presented, showing the performance of the detector, in particular the advantages of using the self-tuning strategies to address the requirements of multiple concurrent applications running on a dynamic environment.

Finally, in collaboration with ICL Lab. (University of Tennessee), we study failure detection in the context of ExaScale computing. We designed and evaluated a new robust failure detector, able to 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 has been published at SC 2016 conference [26].

6.1.2. Self-Stabilization

Regardless its initial state, a *self-stabilizing* system has the ability to reach a correct behavior in finite time. Self-stabilization is a generic paradigm to tolerate transient faults (*i.e.*, faults of finite duration) in distributed systems. Self-stabilization is also a suitable approach to design reliable solutions for dynamic systems. Results obtained in this area by Regal members in 2016 follow.

In [8], we address the ability to maintain distributed structures at large scale. Among the many different structures proposed in this context, The prefix tree structure is a good candidate for indexing and retrieving information. One weakness of using such a distributed structure stands in its poor native fault tolerance, leading to the use of preventive costly mechanisms such as replication. We focus on making tries self-stabilizing over such platforms, and propose a self-stabilizing maintenance algorithm for a prefix tree using a message passing model. The proof of self-stabilization is provided, and simulation results are given, to better capture its performances.

In [4], we propose a silent self-stabilizing leader election algorithm for bidirectional connected identified networks of arbitrary topology. Written in the locally shared memory model, it assumes the distributed unfair daemon, *i.e.*, the most general scheduling hypothesis of the model. Our algorithm requires no global knowledge on the network (such as an upper bound on the diameter or the number of processes). We show that 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 asynchronous 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 stabilize in a non polynomial number of steps in the worst case.

A *snap-stabilizing* protocol, regardless of the initial configuration of the system, guarantees that it always behaves according to its specification. In [9], we consider the locally shared memory model. In this model, we propose a snap-stabilizing Propagation of Information with Feedback (PIF) protocol for rooted networks of arbitrary topology. Then, we use the proposed PIF protocol as a key module in the design of snap-stabilizing solutions for some fundamental problems in distributed systems, such as Leader Election, Reset, Snapshot, and Termination Detection. Finally, we show that in the locally shared memory model, snap-stabilization is as expressive as self-stabilization by designing a universal transformer to provide a snap-stabilizing version of any protocol that can be (automatically) self-stabilized. Since by definition, a snap-stabilizing algorithm is self-stabilizing, self- and snap-stabilization have the same expressiveness in the locally shared memory model.

In [6], we address the *committee coordination problem*: A committee consists of a set of professors and committee meetings are synchronized, so that each professor participates in at most one committee meeting at a time. We propose two snap-stabilizing distributed algorithms for the committee coordination. They are enriched with some desirable properties related to concurrency, (weak) fairness, and a stronger synchronization mechanism called 2-Phase Discussion. Existing work in the literature has shown that (1) in general, fairness cannot be achieved in committee coordination, and (2) it becomes feasible if each professor waits for meetings infinitely often. Nevertheless, we show that even under this latter assumption, it is impossible to implement a

fair solution that allows maximal concurrency. Hence, we propose two orthogonal snap-stabilizing algorithms, each satisfying 2-phase discussion, and either maximal concurrency or fairness.

6.1.3. Dynamic Distributed Systems

In [19], we introduce the notion of *gradually stabilizing* algorithm as any self-stabilizing algorithm with the following additional feature: if at most τ*dynamic steps*—a dynamic step is a step containing topological changes—occur starting from a legitimate configuration, it first quickly recovers to a configuration from which a minimum quality of service is satisfied and then gradually converges to stronger and stronger safety guarantees until reaching a legitimate configuration again. We illustrate this new property by proposing a gradually stabilizing unison algorithm, that consists in synchronizing logical clocks locally maintained by the processes.

The next results consider highly dynamic distributed systems modelled by time-varying graphs (TVGs). In [7], we first address proof of impossibility results that often use informal arguments about convergence. We provide a general framework that formally proves the convergence of the sequence of executions of any deterministic algorithm over TVGs of any convergent sequence of TVGs. Next, we focus of the weakest class of long-lived TVGs, *i.e.*, the class of TVGs where any node can communicate any other node infinitely often. We illustrate the relevance of our result by showing that no deterministic algorithm is able to compute various distributed covering structure on any TVG of this class. Namely, our impossibility results focus on the eventual footprint, the minimal dominating set and the maximal matching problems.

We also study the k-set agreement problem, a generalization of the consensus problem where processes can decide up to k different values. Very few papers have tackled this problem in dynamic networks. Exploiting the formalism of TVGs, we propose in [11] a new quorum-based failure detector for solving k-set agreement in dynamic networks with asynchronous communications. We present two algorithms that implement this new failure detector using graph connectivity and message pattern assumptions. We also provide an algorithm for solving k-set agreement using our new failure detector.

Finally, in [22], we deal with the classical problem of exploring a ring by a cohort of synchronous robots. We focus on the perpetual version of this problem in which it is required that each node of the ring is visited by a robot infinitely often. We assume that the robots evolve in ring-shape TVGs, *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. We also assume that each node is infinitely often reachable from any other node. In this context, we aim at providing a self-stabilizing algorithm to the robots (*i.e.*, the algorithm must guarantee an eventual correct behavior regardless of the initial state and positions of the robots). We show that this problem is deterministically solvable in this harsh environment by providing a self-stabilizing algorithm for three robots.

6.2. Large scale data distribution

Participants: Luciana Arantes [correspondent], Rudyar Cortes, Mesaac Makpangou, Sébastien Monnet, Pierre Sens.

The proliferation of GPS-enabled devices leads to the massive generation of geotagged data sets recently known as Big Location Data. It allows users to explore and analyse data in space and time, and requires an architecture that scales with the insertions and location-temporal queries workload from thousands to millions of users. Most large scale key-value data storage solutions only provide a single one-dimensional index which does not natively support efficient multidimensional queries. In 2016, we propose GeoTrie [29], a scalable architecture built by coalescing any number of machines organized on top of a Distributed Hash Table. The key idea of our approach is to provide a distributed global index which scales with the number of nodes and provides natural load balancing for insertions and location-temporal range queries. We assess our solution using the largest public multimedia data set released by Yahoo! which includes millions of geotagged multimedia files.

We also propose ECHO [10], a novel and lightweight solution that efficiently supports range queries over a ring-like Distributed Hash Table (DHT) structure. By implementing a tree-based index structure and an effective query routing strategy, ECHO provides low-latency and low-overhead query searches by exploiting the Tabu Search principle. Load balancing is also improved reducing the traditional bottleneck problems arising in upper level nodes of tree-based index structures such as PHT. Furthermore, ECHO copes with DHT churn problems as its index exploits logical information as opposed to static reference cache approaches or replication techniques. The performance evaluation results obtained using PeerSim simulator show that ECHO achieves efficient performance compared other solutions such as the PHT strategy and its optimized version which includes a query cache.

6.3. Consistency protocols

Participants: Marc Shapiro [correspondent], Tyler Crain, Mahsa Najafzadeh, Marek Zawirski, Alejandro Tomsic.

6.3.1. Static Reasoning About Consistency, and associated tools

Large-scale distributed systems often rely on replicated databases that allow a programmer to request different data consistency guarantees for different operations, and thereby control their performance. Using such databases is far from trivial: requesting stronger consistency in too many places may hurt performance, and requesting it in too few places may violate correctness. To help programmers in this task, we propose the first proof rule for establishing that a particular choice of consistency guarantees for various operations on a replicated database is enough to ensure the preservation of a given data integrity invariant. Our rule is modular: it allows reasoning about the behaviour of every operation separately under some assumption on the behaviour of other operations. This leads to simple reasoning, which we have automated in an SMT-based tool. We present a nontrivial proof of soundness of our rule and illustrate its use on several examples.

The intuition was presented at EuroSys 2015 [47]. We present the full theory and proofs in the POPL 2016 paper "'Cause I'm Strong Enough: Reasoning about Consistency Choices in Distributed Systems" [30]. The proof procedure and tool are described in PaPoC 2016 paper "The CISE Tool: Proving Weakly-Consistent Applications Correct" [34] and a YouTube video [48]. It is also the focus of Mahsa Najafzadeh's PhD thesis [3].

6.3.2. Scalable consistency protocols

Developers of cloud-scale applications face a difficult decision of which kind of storage to use, summarised by the CAP theorem. Currently the choice is between classical CP databases, which provide strong guarantees but are slow, expensive, and unavailable under partition; and NoSQL-style AP databases, which are fast and available, but too hard to program against. We present an alternative: Cure provides the highest level of guarantees that remains compatible with availability. These guarantees include: causal consistency (no ordering anomalies), atomicity (consistent multi-key updates), and support for high-level data types (developer friendly API) with safe resolution of concurrent updates (guaranteeing convergence). These guarantees minimise the anomalies caused by parallelism and distribution, thus facilitating the development of applications. This paper presents the protocols for highly available transactions, and an experimental evaluation showing that Cure is able to achieve scalability similar to eventually- consistent NoSQL databases, while providing stronger guarantees.

This work is published under the title "Cure: Strong semantics meets high availability and low latency" at ICDCS 2016 [18].

6.3.3. Lightweight, correct causal consistency

Non-Monotonic Snapshot Isolation (NMSI), a variant of the widely deployed Snapshot Isolation (SI), aims at improving scalability by relaxing snapshots. In contrast to SI, NMSI snapshots are causally consistent, which allows for more parallelism and a reduced abort rate.

This work documents the design of PhysiCS-NMSI, a transactional protocol implementing NMSI in a partitioned data store. It is the first protocol to rely on a single scalar taken from a physical clock for tracking causal dependencies and building causally consistent snapshots. Its commit protocol ensures atomicity and the absence of write-write conflicts. Our PhysiCS-NMSI approach increases concurrency and reduces abort rate and metadata overhead as compared to state-of-art systems.

The paper "PhysiCS-NMSI: efficient consistent snapshots for scalable snapshot isolation" is published at PaPoC 2016 [36].

6.3.4. Reconciling consistency and scalability

Geo-replicated storage systems are at the core of current Internet services. Unfortunately, there exists a fundamental tension between consistency and performance for offering scalable geo-replication. Weakening consistency semantics leads to less coordination and consequently a good user experience, but it may introduce anomalies such as state divergence and invariant violation. In contrast, maintaining stronger consistency precludes anomalies but requires more coordination. This paper discusses two main contributions to address this tension. First, RedBlue Consistency enables blue operations to be fast (and weakly consistent) while the remaining red operations are strongly consistent (and slow). We identify sufficient conditions for determining when operations can be blue or must be red. Second, Explicit Consistency further increases the space of operations that can be fast by restricting the concurrent execution of only the operations that can break application-defined invariants. We further show how to allow operations to complete locally in the common case, by relying on a reservation system that moves coordination off the critical path of operation execution.

The paper "Geo-Replication: Fast If Possible, Consistent If Necessary" is published in the IEEE CS Data Engineering Bulletin of March 2016 [5].

6.3.5. Consistency in 3D

Comparisons of different consistency models often try to place them in a linear strong-to-weak order. However this view is clearly inadequate, since it is well known, for instance, that Snapshot Isolation and Serialisability are incomparable. In the interest of a better understanding, we propose a new classification, along three dimensions, related to: a total order of writes, a causal order of reads, and transactional composition of multiple operations. A model may be stronger than another on one dimension and weaker on another. We believe that this new classification scheme is both scientifically sound and has good explicative value. We presents the three-dimensional design space intuitively.

This work was presented as an invited keynote paper at Concur 2016 [17].

6.3.6. Scalable consistency protocols

Collaborative text editing systems allow users to concurrently edit a shared document, inserting and deleting elements (e.g., characters or lines). There are a number of protocols for collaborative text editing, but so far there has been no precise specification of their desired behavior, and several of these protocols have been shown not to satisfy even basic expectations. This work provides a precise specification of a replicated list object, which models the core functionality of replicated systems for collaborative text editing. We define a strong list specification, which we prove is implemented by an existing protocol, as well as a weak list specification, which admits additional protocol behaviors.

A major factor determining the efficiency and practical feasibility of a collaborative text editing protocol is the space overhead of the metadata that the protocol must maintain to ensure correctness. We show that for a large class of list protocols, implementing either the strong or the weak list specification requires a metadata overhead that is at least linear in the number of elements deleted from the list. The class of protocols to which this lower bound applies includes all list protocols that we are aware of, and we show that one of these protocols almost matches the bound.

This work is published at PODC 2016 [21].

6.3.7. Highly-responsive CRDTs for group editing

Group editing is a crucial feature for many end-user applications. It requires high responsiveness, which can be provided only by optimistic replication algorithms, which come in two classes: classical Operational Transformation (OT), or more recent Conflict-Free Replicated Data Types (CRDTs).

Typically, CRDTs perform better on *downstream* operations, i.e., when merging concurrent operations than OT, because the former have logarithmic complexity and the latter quadratic. However, CRDTs are often less responsive, because their *upstream* complexity is linear. To improve this, this paper proposes to interpose an auxiliary data structure, called the *identifier data structure* in front of the base CRDT. The identifier structure ensures logarithmic complexity and does not require replication or synchronization. Combined with a block-wise storage approach, this approach improves upstream execution time by several orders of magnitude, with negligeable impact on memory occupation, network bandwidth, and downstream execution performance.

This work is published at ACM Group 2016 [27].

6.4. Memory management for multicores

Participants: Antoine Blin, Damien Carver, Maxime Lorrillere, Sébastien Monnet, Julien Sopena [correspondent].

Regal co-advises with Whisper team the PhD of Antoine Blin. The thesis focusses on modern complex embedded systems that involve a mix of real-time and best effort applications. The recent emergence of low-cost multicore processors raises the possibility of running both kinds of applications on a single machine, with virtualization ensuring isolation. Nevertheless, memory contention can introduce other sources of delay, that can lead to missed deadlines. We first investigated the source of memory contention for the Mibench benchmark in a paper published at ETYS 2016 [25]. Then, in a paper published at ECRTS 2016 [24], we present a combined offline/online memory bandwidth monitoring approach. Our approach estimates and limits the impact of the memory contention incurred by the best-effort applications on the execution time of the real-time application. Using our approach, the system designer can limit the overhead on the real-time application to under 5% of its expected execution time, while still enabling progress of the best-effort applications.

Another memory management challenge for multi-cores is the fragmentation induced by the virtualized environments. Previously, we proposed Puma (for Pooling Unused Memory in Virtual Machines) which allows I/O intensive applications running on top of VMs to benefit of large caches. This was realized by providing a remote caching mechanism that provides the ability for any VM to extend its cache using the memory of other VMs located either in the same or in a different host. This work was defended by Maxime Lorrillere in April 2016 [2].

More recently, we study the memory arbitration between containers. In the Damien Carver's PhD thesis (started in October 2015), we are designing ACDC (Advanced Consolidation for Dynamic Containers), a kernel-level mechanisms that automatically provides more memory to the most active containers.

SPIRALS Project-Team

7. New Results

7.1. Change Impact Analysis

In [21], we have proposed a novel evaluation technique for change impact analysis (CIA). CIA is a prediction problem that, given a source code element in a program, determines the other source code elements impacted if one changes this original source code element. Given the large size of the element space in complex programs, this prediction requires a trade-off between different dimensions: precision, completeness, time. The novelty of the result lies in the use of mutation analysis to study simultaneously these three dimensions. This result is backed by an empirical evaluation performed on 10 open-source Java programs and 5 mutation operators, which enabled to generate 17,000 mutants and study how the error they introduce propagates. This result has been achieved in the context of the PhD thesis, defended in November 2016, of Vicenzo Musco [15].

7.2. Learning Power Models for Distributed and Virtualized Environments

Energy efficiency is a major concern for modern ICT infrastructures. The a priori estimation of the level of energy consumed by a given service is a difficult problem given the intricate nature of hardware and software that are involved. Consequently, even before considering saving, measuring the exact amount of energy consumed by a given software service or process is required. Over the last few years, a dozen of ad hoc power models have been proposed in the literature. Nevertheless they cannot cope with the constant evolution of software and hardware architecture. We have therefore defined and implemented a toolkit that automatically learns the power models of a given architecture, independently of the features and the complexity it exhibits. This toolkit considers traditional distributed environment as well as virtualized, cloud-based ones. This result has been achieved in the context of the PhD thesis, defended in November 2016, of Maxime Colmant [11].

7.3. Crowdmining to Increase the Quality of Software Systems

Modern software systems, especially in the open source world, are more and more part of ecosystems where large quantities of data about these systems are available. These data may come for example from application stores (e.g. Google Play Store or Apple Store for mobile applications), forges (e.g. GitHub), or from the usage conditions experienced by users of these software systems. This large amount of data enables to unlock some specific challenges where knowledge about the software systems can be automatically mined and learnt. In this domain, we obtained new results on the mining of mobile software antipatterns on a crowd of mobile applications and their versions to study their impact on resource consumption [32]. This result has been achieved in the context of the PhD thesis, defended in November 2016, of Geoffrey Hecht [13]. We also consider the crowd of mobile devices and users to detect and reproduce application crashes in the wild. By leveraging our results in the domain of in-breath monitoring, we use the APISENSE® platform (see Section 6.1) to collect extended crash reports that can be aggregated to infer the minimal execution path that lead to a crash [28]. This result has been achieved in the context of the PhD thesis, defended in December 2016, of María Gomez Lacruz [12]. These results are also in relation with our activities in the context of the SOMCA associated team (see Section 9.4).

7.4. Self-Optimization of Virtualized Environments

Elasticity is a major property of virtualized computing environments. In this domain, we especially work at the infrastructure and platform levels of a cloud computing system where we obtained two results that enable to better self-optimize the consumed resources. At the infrastructure level, we proposed CloudGC, a new middleware service for suspending, resuming, and recycling idle virtual machines. The algorithm has been implemented on top of the OpenStack cloud operating system. At the platform level, we proposed a new self-balancing approach to dynamically optimize the performance of the Hadoop framework for the distributed storage and processing of large data sets. These results have been achieved in the context of the PhD thesis, defended in December 2016, of Bo Zhang [16].

WHISPER Project-Team

7. New Results

7.1. Software engineering for infrastructure software

Our main work in this area has focused on driver porting. We aim at fully automating the backporting (or symmetrically forward porting) process: given any driver for one Linux kernel version, one would like to obtain a driver that has the same functionality for another kernel version. This requires identifying the changes that are needed, obtaining examples of how to carry these changes out, and inferring from these examples a change that is appropriate for the given driver code. We have carried out a preliminary study in this direction with David Lo of Singapore Management University; this work, published at ICSME 2016 [17], is limited to a port from one version to the next one, in the case where the amount of change required is limited to a single line of code.

More general automation of backporting requires more extensive search for relevant examples. This raises issues of scalability, because the Linux kernel code history is very large, and of expressivity, because we need to be able to express complex patterns to obtain change examples that are most relevant to a particular backporting problem. To this end, we have been adapted the notation used by Coccinelle, which describes how a change should be carried out, into a *patch query language* that allows describing patterns of changes that have been previously performed. The associated tool, Prequel, can find patches that match a particular pattern among several hundred thousand commits, often in tens of seconds [20]. This work is supported in part by OSADL, a consortium of companies, mostly in Germany, supporting the use and development of open source software in automation and other industries.

We will continue research in this direction over the next three years as part of the ANR PRCI ITrans project, awarded in 2016 and to be carried out in 2017-2020.

7.2. Developing infrastructure software using Domain Specific Languages

To bootstrap our long-term effort in designing safe and composable domain-specific languages, we have initiated two exploratory actions involving a combination of advanced type-theoretic concepts and domain-specific compilation techniques. Both actions are complementary, the first adopts a bottom-up approach – going from low-level artifacts to high-level abstractions – while the second follows a top-down approach – offering a safe translation of high-level guarantees to low-level executable code.

Our first line of inquiry, of which some early results have been published at FLOPS 2016 [13], aims at bridging the formalization gap between low-level, bit-twiddling code and high-level, mathematical abstractions. As such, it provided us with an opportunity to experiment with using an interactive theorem prover to design abstractions in a bottom-up manner. We have developed a library (`ssrbit`, publicly available under an open-source license) for modeling and computing with bit vectors in the Coq [35] proof assistant. Because ease of proving and efficiency in computing are often incompatible objectives, this library offers a two pronged approach by offering an abstract specification for proving and an efficient implementation for computing; we have shown that the latter is correct with respect to the former. Using this model of bit-level operations, we have implemented a bitset library and proved its correctness with respect to the formalization of sets of finite types provided by the Ssreflect library [43], which is part of the Mathematical Components framework developed at the MSR-Inria joint center. This library thus enables a seamless interaction of sets for computing and sets for proving. This library also supports the trustworthy extraction of bitsets down to OCaml's machine integers: we gained greater confidence in our model by adopting a methodology based on exhaustive testing. This enabled us to implement three bit-twiddling applications in Coq (Bloom filter, n-queens, and the efficient enumeration of all k-combinations of a set), prove their correctness and obtain efficient low-level OCaml code.

Our second line of inquiry is influenced by the realization that domain-specific languages are often treating the symptoms rather than providing a cure. Infrastructure software is often developed in C, which suffers from many semantic kludges and is, as a result, hardly amenable to formal reasoning. Many domain-specific languages are born out of the frustration of being unable to guarantee static properties of one's code: more often than not, the resulting language is little more than a domain-specific variant of Pascal supporting custom static analyses and some form of transliteration to C. To achieve safety and composability, we believe that a more holistic approach is called for, involving not only the design of a domain-specific *syntax* but also of a domain-specific *semantics*. Concretely, we are exploring the design of *certified domain-specific compilers* that integrate, from the ground up, a denotational and domain-specific semantics as part of the design of a domain-specific language. This vision is illustrated by our work on the safe compilation of Coq programs into secure OCaml code [14], [18]. It combines ideas from gradual typing – through which types are compiled into run-time assertions – and the theory of ornaments [37] – through which Coq datatypes can be related to OCaml datatypes. Within this formal framework, we enable a secure interaction, termed *dependent interoperability*, between correct-by-construction software and untrusted programs, be it system calls or legacy libraries. To do so, we trade static guarantees for runtime checks, thus allowing OCaml values to be safely coerced to dependently-typed Coq values and, conversely, to expose dependently-typed Coq programs defensively as OCaml programs. Our framework is developed in Coq: it is constructive and verified in the strictest sense of the terms. It thus becomes possible to internalize and hand-tune the extraction of dependently-typed programs to interoperable OCaml programs within Coq itself. This work is part of a collaboration with Eric Tanter, from the University of Chile, and Nicolas Tabareau, from the Ascola Inria project-team.

To further explore the realm of domain-specific compilers, we have been involved in the design and implementation of a certified compiler for the Lustre [30] 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. This on-going work, in collaboration with members of the Parkas team and Gallium team of Inria Paris, formalizes the compilation of a synchronous data-flow language into an imperative sequential language, which is eventually translated to Cminor [54], one of CompCert's intermediate languages. This project illustrates perfectly our methodological position: 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 is part of an on-going collaboration with Marc Pouzet and Tim Bourke, from the Parkas team of Inria Paris, Lionel Rieg, postdoc at Collège de France, and Xavier Leroy, from the Gallium Inria project-team.

In terms of DSL design for domains where correctness is critical, our current focus is on process scheduling and multicore architectures. 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.* [59], 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 allow expressing load balancing properties, and to enable verification of critical scheduling properties such as liveness; for the latter, we are exploring the use of tools such as the Z3 theorem prover from Microsoft, and the Leon theorem prover from EPFL. A first version of the language has been designed and we expect to have 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) will be supported in 2017 as a postdoc as part of the Inria-EPFL joint laboratory.

Finally, in the context of the Multicore IPL, we are working with Jens Gustedt and Mariem Saeid of the Inria Camus project-team on developing a domain-specific language that eases programming with the ordered read-

write lock (ORWL) execution model. The goal of this work is to provide a single execution model for parallel programs and to allow them to be deployed on multicore machines with varying architectures [16].

7.3. Run-time environments for multicore architectures

In the recent past, we acquired a solid expertise in multicore systems through the PhD of Jean-Pierre Lozi [60] and Florian David [38]. This expertise has led us to initiate several collaborations with industry partners, in the form of CIFRE PhD support. We first targeted real-time multicore systems with the goal of improving resource usage, through a cooperation with Renault and the PhD of Antoine Blin. Recently, we have started another cooperation on multicore real-time systems for avionics and space with Thales TRT, that is the topic of the PhD of Cédric Courtaud.

The PhD of Jean-Pierre Lozi [60] was on improving the performance locks on large multicore architectures. In an paper published at Usenix ATC 2012 [58], and more recently in an article published in 2016 in ACM Transactions on Computer Systems (TOCS) [10], we proposed a new locking technique, Remote Core Locking (RCL), that aims to accelerate the execution of critical sections in legacy applications on multicore architectures. RCL is currently one of the most efficient locking technique and the ATC 2012 paper has currently 67 citations on Google scholar. The idea of RCL is to replace lock acquisitions by optimized remote procedure calls to a dedicated server hardware thread. RCL limits the performance collapse observed with other lock algorithms when many threads try to acquire a lock concurrently and removes the need to transfer lock-protected shared data to the hardware thread acquiring the lock because such data can typically remain in the server's cache. Eighteen applications were used to evaluate RCL from standard multicore benchmark suites, such as SPLASH-2 and Phoenix 2. By using RCL instead of Linux POSIX locks, performance is improved by up to 2.5 times on Memcached, and up to 11.6 times on Berkeley DB with the TPC-C client. On a SPARC machine with two Sun Ultrasparc T2+ processors and 128 hardware threads, performance is improved by up to 1.3 times with respect to Solaris POSIX locks on Memcached, and up to 7.9 times on Berkeley DB with the TPC-C client.

The PhD of Antoine Blin is on modern complex embedded systems that involve a mix of real-time and best-effort applications. The recent emergence of low-cost multicore processors raises the possibility of running both kinds of applications on a single machine, with virtualization ensuring isolation. Nevertheless, memory contention can introduce other sources of delay, that can lead to missed deadlines. We first investigated the source of memory contention for the Mibench benchmark in a paper published at NETYS 2016 [12]. Then, in a paper published at ECRTS 2016 [11], we present a combined offline/online memory bandwidth monitoring approach. Our approach estimates and limits the impact of the memory contention incurred by the best-effort applications on the execution time of the real-time application. Using our approach, the system designer can limit the overhead on the real-time application to under 5% of its expected execution time, while still enabling progress of the best-effort applications.

ALPINES Project-Team

7. New Results

7.1. Communication avoiding algorithms

Our group continues to work on algorithms for dense and sparse linear algebra operations that minimize communication. 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.

In [9], we discuss sparse matrix-matrix multiplication (or SpGEMM), which is an important operation for many algorithms in scientific computing. In our previous work we have identified lower bounds on communication for this operation, which is the limiting factor of SpGEMM. Even though 3D (or 2.5D) algorithms have been proposed and theoretically analyzed in the flat MPI model on Erdos–Renyi matrices, those algorithms had not been implemented in practice and their complexities had not been analyzed for the general case. In this work, we present the first implementation of the 3D SpGEMM formulation that exploits multiple (intranode and internode) levels of parallelism, achieving significant speedups over the state-of-the-art publicly available codes at all levels of concurrencies. We extensively evaluate our implementation and identify bottlenecks that should be subject to further research.

In [10] we discuss algorithms that not only aim at minimizing communication, but they also aim at reducing the number of writes to secondary storage. Most of the prior work does not distinguish between loads and stores, i.e., between reads and writes to a particular memory unit. But in fact there are some current and emerging nonvolatile memory technologies (NVM) where writes can be much more expensive (in time and energy) than reads. NVM technologies are being considered for scientific applications on extreme scale computers and for cluster computing platforms, in addition to commodity computers.

This motivates us to first refine prior work on communication lower bounds of algorithms which did not distinguish between loads and stores to derive new lower bounds on writes to different levels of a memory hierarchy. When these new lower bounds on writes are asymptotically smaller than the previous bounds on the total number of loads and stores, we ask whether there are algorithms that attain them. We call such algorithms, that both minimize the total number of loads and stores (i.e., are CA), and also do asymptotically fewer writes than reads, *write-avoiding (WA)*. In this paper, we identify several classes of problems where either sequential or parallel WA algorithms exist, or provably cannot.

In [7] we introduce a new approach for reducing communication in Krylov subspace methods that consists of enlarging the Krylov subspace by a maximum of t vectors per iteration, based on the domain decomposition of the graph of A.We show in this paper that 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 this paper we focus on Conjugate Gradient (CG), a Krylov projection method for symmetric (Hermitian) positive definite matrices. We discuss two new versions of Conjugate Gradient. The first method, multiple search direction with orthogonalization CG (MSDO-CG), is an adapted version of MSD-CG with the A-orthonormalization of the search directions to obtain a projection method that guarentees convergence at least as fast as CG. The second projection method that we propose here, long recurrence enlarged CG (LRE-CG), is similar to GMRES in that we build an orthonormal basis for the enlarged Krylov subspace rather than finding search directions. Then, we use the whole basis to update the solution and the residual. We compare the convergence behavior of both methods using different A-orthonormalization and orthonormalization methods and then we compare the most stable versions with CG and other related methods. Both methods converge faster than CG in terms of iterations, but LRE-CG converges faster than MSDO-CG since it uses the whole basis to update the solution rather than only t search directions. And the more subdomains are introduced or the larger t is, the faster is the convergence of both methods with respect to CG in terms of iterations. For example, for $t = 64$ the MSDO-CG and LRE-CG methods converge in 75% up to 98 less iteration with respect to CG for the different test matrices.

In [12] we present an algorithm for computing a low rank approximation of a sparse matrix based on a truncated LU factorization with column and row permutations. We present various approaches for determining the column and row permutations that show a trade-off between speed versus deterministic/probabilistic accuracy. We show that if the permutations are chosen by using tournament pivoting based on QR factorization, then the obtained truncated LU factorization with column/row tournament pivoting, LU_CRTP, satisfies bounds on the singular values which have similarities with the ones obtained by a communication avoiding rank revealing QR factorization. Experiments on challenging matrices show that LU_CRTP provides a good low rank approximation of the input matrix and it is less expensive than the rank revealing QR factorization in terms of computational and memory usage costs, while also minimizing the communication cost. We also compare the computational complexity of our algorithm with randomized algorithms and show that for sparse matrices and high enough but still modest accuracies, our approach is faster.

7.2. Integral equation based domain decomposition

We kept on studying the convergence of classical domain decomposition strategies applied to multi-trace formulations (MTF). In the contribution [18], we present a gentle introduction to multi-trace formalism aimed at the domain decomposition community as well as analytical calculations in simple geometrical configuration where a full analysis of block-Jacoobi applied to MTF is possible. We only consider transmission problems in 1D with one or two interfaces. In [5], we generalize this analysis to arbitrary 2D or 3D transmission problems with arbitrary subdomain partitionning, only assuming that there is no junction point. The analysis holds mainly for completely homogeneous media with no material constrast, and in such a case we determine the spectrum of the multi-trace operator, as well as the spectrum of the Jacobi operator. We show that this spectrum only consists in a finite number of point values. In the more general case where the propagation medium is piecewise constant, this analysis still yields the location of the essential spectrum of the MTF and the Jacobi operator.

This analysis also led to an explicit expression for the inverse of the MTF operators for transmission problems in the case of perfectly homogeneous media. This was studied during the intership of Alan Ayala, and was described and tested numerically in 3D in the proceedings.

The analysis presented in [5] also shows that, in the case of purely homogeneous media, a block Jacobi strategy converges in a number of steps that exactly corresponds to the depth of the adjacency graph of the subdomain partition under consideration, which suggests a close relationship with Optimized Schwarz Methods (OSM), following the ideas of [20]. We investigated this point during the internship of Pierre Marchand, and we exhibited fully explicitely the exact relationship between block-Jacobi-MTF and OSM. Besides, we also generalized the analysis presented in [5] to the case of a competely heterogeneous problem, which involves abstract boundary integral operators that are not easily computable.

7.3. Multi-subdomain integral equations

In the context of boundary integral equations adapted to wave scattering in piecewise constant media in harmonic regime, we also made significant progress in the study of the single trace boundary integral formulation (STF) of the second kind originally introduced in [17]. This work was achieved in collaboration with Ralf Hiptmair and Elke Spindler (ETH Zürich). First of all, we proposed a version of this formulation for the solution to Maxwell's equations whereas, so far, it had been studied only in the context of scalar wave scattering (Helmholtz equation). In this direction, we conducted numerical experiments which confirmed the attractive properties of the matrices obtained when discretising such formulations (good accuracy, and good conditionning independent of discretisation parameters). For Maxwell's equations, we also established elementary theoretical results of STF 2nd kind such as Fredholmness of the corresponding integral operator.

So far, second kind STF had been stuudied for wave scatering problems where material contrasts only enter in the compact part of the partial differential operator, which is harmless regarding the Fredholmness of the corresponding boundary integral operator. Thus, in [19], we investigated the case where material contrasts come into play in the principal part of the operator, considering a pure diffusion-transmission problem. In

this case, we have been able to establish well-posedness (hence Fredholmness). A rather naive approach leads to choose Sobolev spaces of fractionnal order (half-integer) as main functional setting for this formulation. We showed that this formulation can be extended so as to make sense in the space of square integrable trace functions. This is much more handy a functional setting that allows in particular discontinuous Galerkin discretisations of the corresponding boundary integral equations.

7.4. Asymptotics for a semi-linear convex problem with small inclusion

In [16], in collaboration with Lucas Chesnel (Inria Defi) and Sergei Nazarov (Saint-Petersbourg University), we recently investigated the asymptotics of the solution to a semi-linear problem in 2D with Dirichlet boundary condition. The partial differential operator under consideration was $-\Delta u + (u)^{2p+1}$ where p is a positive integer. The computational domain is assumed to contain a small Dirichlet obstacle of size $\delta > 0$. Using the method of matched asymptotic expansions, we compute an asymptotic expansion of the solution as δ tends to zero. Its relevance was justified by proving a rigorous error estimate. Then we construct an approximate model, based on an equation set in the limit domain without the small obstacle, which provides a good approximation of the far field of the solution of the original problem. The interest of this approximate model lies in the fact that it leads to a variational formulation which is very simple to discretize. We obtained numerical experiments to illustrate the analysis.

7.5. Time-dependent wave splitting and source separation

Starting from classical absorbing boundary conditions, we (M. Grote, M. Kray, F. Nataf and F. Assous) propose a method for the separation of time-dependent scattered wave fields due to multiple sources or obstacles. More precisely, we propose a method to determine the separate outgoing components of the incident and scattered wave fields for time-dependent scattering problems. In the case of two superposed wave fields, our method applies to the following three typical configurations: two distinct localized sources with unknown time history each, a single (unknown) localized source with a nearby scatterer, or two separate scatterers illuminated by a known incident wave field. In all three cases, our method permits to recover the individual outgoing components from measurements of the total scattered field at a distance. In doing so, the particular nature of the scatterer, be it an im- penetrable well-defined obstacle or a penetrable localized inhomogeneity, is immaterial; only the purely outgoing character of the individual wave fields matters. In contrast to previous work, our approach is local in space and time, deterministic, and also avoids any a priori assumptions on the frequency spectrum of the signal. Numerical simulations in FreeFem++ in two space dimensions illustrate the usefulness of wave splitting for time-dependent scattering problems. This work was presented to several international conferences and was published in J. Comput. Phys. (2016).

7.6. SORAS GenEO-2

Optimized Schwarz methods (OSM) are very popular methods which were introduced by P.L. Lions (1989) for elliptic problems and by B. Després (1990) for propagative wave phenomena. We (R. Haferssas, P. Jolivet and F. Nataf) give here a theory for Lions' algorithm that is the genuine counterpart of the theory developed over the years for the Schwarz algorithm. The first step is to introduce a new symmetric variant of the ORAS (Optimized Restricted Additive Schwarz) algorithm that is suitable for the analysis of a two-level method. Then we build a coarse space for which the convergence rate of the two-level method is guaranteed regardless of the regularity of the coefficients. We show scalability results for thousands of cores for nearly incompressible elasticity and the Stokes systems with a continuous discretization of the pressure.

7.7. Numerical modeling and high speed parallel computing: new perspectives for tomographic microwave imaging for brain stroke detection and monitoring

These works deals with microwave tomography for brain stroke imaging using state-of-the-art numerical modeling and massively parallel computing. Iterative microwave tomographic imaging requires the solution

of an inverse problem based on a minimization algorithm (e.g. gradient based) with successive solutions of a direct problem such as the accurate modeling of a whole-microwave measurement system. Moreover, a sufficiently high number of unknowns is required to accurately represent the solution. As the system will be used for detecting the brain stroke (ischemic or hemorrhagic) as well as for monitoring during the treatment, running times for the reconstructions should be reasonable. The method used is based on high-order finite elements, parallel preconditioners from the Domain Decomposition method and Domain Specific Language with open source FreeFem++ solver. This work, for which we got the Joseph Fourier-Bull prize, is supported by ANR grant MEDIMAX (ANR-13-MONU-0012) and was granted access to the HPC resources of TGCC@CEA under the allocations 2016-067519 and 2016- 067730 made by GENCI.

AVALON Project-Team

6. New Results

6.1. Energy Efficiency of Large Scale Distributed Systems

Participants: Laurent Lefevre, Daniel Balouek-Thomert, Eddy Caron, Radu Carpa, Marcos Dias de Assunção, Jean-Patrick Gelas, Olivier Glück, Jean-Christophe Mignot, Violaine Villebonnet.

6.1.1. Energy Efficient Core Networks with SDN

This work [14], [15] seeks to improve the energy efficiency of backbone networks by providing an intra-domain Software Defined Network (SDN) approach to selectively and dynamically turn off and on a subset of links. We proposed the STREETE framework (SegmenT Routing based Energy Efficient Traffic Engineering) that represents an online method to switch some links off/on dynamically according to the network load. We have implemented a working prototype in the OMNET++ simulator and design a validation platform [15] based on NetFPGA and Raspberry equipment with SDN frameworks (ONOS).

6.1.2. Energy Proportionality in HPC Systems

Energy savings are among the most important topics concerning Cloud and HPC infrastructures nowadays. Servers consume a large amount of energy, even when their computing power is not fully utilized. These static costs represent quite a concern, mostly because many datacenter managers are over-provisioning their infrastructures compared to the actual needs. This results in a high part of wasted power consumption. In this work [25], [24], [23] , we proposed the BML ("Big, Medium, Little") infrastructure, composed of heterogeneous architectures, and a scheduling framework dealing with energy proportionality. We introduce heterogeneous power processors inside datacenters as a way to reduce energy consumption when processing variable workloads. Our framework brings an intelligent utilization of the infrastructure by dynamically executing applications on the architecture that suits their needs, while minimizing energy consumption. Our first validation process focuses on distributed stateless web servers scenario and we analyze the energy savings achieved through energy proportionality. This research activity is performed with the collaboration of Sepia Team (IRIT, Toulouse) through the co-advising of Violaine Villebonnet.

6.1.3. Energy-Aware Server Provisioning

Several approaches to reduce the power consumption of datacenters have been described in the literature, most of which aim to improve energy efficiency by trading off performance for reducing power consumption. However, these approaches do not always provide means for administrators and users to specify how they want to explore such trade-offs. This work [11] provides techniques for assigning jobs to distributed resources, exploring energy efficient resource provisioning. We use middleware-level mechanisms to adapt resource allocation according to energy-related events and user-defined rules. A proposed framework enables developers, users and system administrators to specify and explore energy efficiency and performance trade-offs without detailed knowledge of the underlying hardware platform. Evaluation of the proposed solution under three scheduling policies shows gains of 25% in energy-efficiency with minimal impact on the overall application performance. We also evaluate reactivity in the adaptive resource provisioning This research activity is performed with the collaboration of NewGen SR society through the co-advising of Daniel Balouek-Thomert.

6.1.4. Improving Energy Re-Usage of Large Scale Computing Systems

The heat induced by computing resources is generally a waste of energy in supercomputers. This is especially true in very large scale supercomputers, where the produced heat has to be compensated with expensive and energy consuming cooling systems. Energy is a critical point for future supercomputing trends that currently try to achieve exascale, without having its energy consumption reaching an important fraction of a nuclear power plant. Thus, new ways of generating or recovering energy have to be explored. Energy harvesting consists in recovering wasted energy. ThermoElectric Generators (TEGs) aim to recover energy by converting wasted dissipated energy into usable electricity. By combining computing units (CU) and TEGs at very large scale, we spotted a potential way to recover energy from wasted heat generated by computations on supercomputers. In this work [30], [20], we study the potential gains in combining TEGs with computational units at petascale and exascale. We explored the technology behind TEGs, and finally our results concerning binding TEGs and computational units in a petascale and exascale system. With the available technology, we demonstrate that the use of TEGs in a supercomputer environment could be realistic and quickly profitable, and hence have a positive environmental impact.

6.2. MPI Application Simulation

Participant: Frédéric Suter.

6.2.1. The SMPI approach

In [37], 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.

6.3. MapReduce Computations on Hybrid Distributed Computations Infrastructures

Participant: Gilles Fedak.

6.3.1. Availability and Network-Aware MapReduce Task Scheduling over the Internet

MapReduce offers an easy-to-use programming paradigm for processing large datasets. In our previous work, we have designed a MapReduce framework called BitDew-MapReduce for desktop grid and volunteer computing environment, that allows non-expert users to run data-intensive MapReduce jobs on top of volunteer resources over the Internet. However, network distance and resource availability have great impact on MapReduce applications running over the Internet. To address this, an availability and network-aware MapReduce framework over the Internet is proposed in [9]. Simulation results show that the MapReduce job response time could be decreased by 27.15%, thanks to Naive Bayes Classifier-based availability prediction and landmark-based network estimation.

6.4. Managing Big Data Life Cycle

Participants: Gilles Fedak, Valentin Lorentz, Laurent Lefevre.

6.4.1. Data Energy Traceability

In this work, we have opened a new research topic around the energy traceability of data. The objective is to answer the question of how many energy has been consumed to produce a particular data. This work is partially based on the concept of data life cycle, that is extended to record each step of the data life cycle.

6.5. Desktop Grid Computing

Participant: Gilles Fedak.

6.5.1. Multi-Criteria and Satisfaction Oriented Scheduling for Hybrid Distributed Computing Infrastructures

Assembling and simultaneously using different types of distributed computing infrastructures (DCI) like Grids and Clouds is an increasingly common situation. Because infrastructures are characterized by different attributes such as price, performance, trust, greenness, the task scheduling problem becomes more complex and challenging. In [7], we presented the design for a fault-tolerant and trust-aware scheduler, which allows to execute Bag-of-Tasks applications on elastic and hybrid DCI, following user-defined scheduling strategies. Our approach, named Promethee scheduler, combines a pull-based scheduler with multi-criteria Promethee decision making algorithm. Because multi-criteria scheduling leads to the multiplication of the possible scheduling strategies, we proposed SOFT, a methodology that allows to find the optimal scheduling strategies given a set of application requirements. The validation of this method is performed with a simulator that fully implements the Promethee scheduler and recreates an hybrid DCI environment including Internet Desktop Grid, Cloud and Best Effort Grid based on real failure traces. A set of experiments shows that the Promethee scheduler is able to maximize user satisfaction expressed accordingly to three distinct criteria: price, expected completion time and trust, while maximizing the infrastructure useful employment from the resources owner point of view. Finally, we present an optimization which bounds the computation time of the Promethee algorithm, making realistic the possible integration of the scheduler to a wide range of resource management software.

6.6. HPC Component Models and OpenMP

Participants: Hélène Coullon, Vincent Lanore, Christian Perez, Jérôme Richard, Thierry Gautier.

6.6.1. Combining Both a Component Model and a Task-based Model

We have 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. We have designed 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. Part of this work is described in [38].

6.6.2. OpenMP Scheduling Heuristic for NUMA Architecture

The recent addition of data dependencies to the OpenMP 4.0 standard provides the application programmer with a more flexible way of synchronizing tasks. Using such an approach allows both the compiler and the runtime system to know exactly which data are read or written by a given task, and how these data will be used through the program lifetime. Data placement and task scheduling strategies have a significant impact on performances when considering NUMA architectures. While numerous papers focus on these topics, none of them has made extensive use of the information available through dependencies. One can use this information to modify the behavior of the application at several levels: during initialization to control data placement and during the application execution to dynamically control both the task placement and the tasks stealing strategy, depending on the topology. This paper [26] introduces several heuristics for these strategies and their implementations in our OpenMP runtime XKAAPI. We also evaluate their performances on linear algebra applications executed on a 192-core NUMA machine, reporting noticeable performance improvement when considering both the architecture topology and the tasks data dependencies. We finally compare them to strategies presented previously by related works.

6.6.3. Extending OpenMP with Affinity Clause: Design and Implementation

OpenMP 4.0 introduced dependent tasks, which give the programmer a way to express fine grain parallelism. Using appropriate OS support (such as NUMA libraries), the runtime can rely on the information in the depend clause to dynamically map the tasks to the architecture topology. Controlling data locality is one of the key factors to reach a high level of performance when targeting NUMA architectures. On this topic, OpenMP does not provide a lot of flexibility to the programmer yet, which lets the runtime decide where a task should be executed. In [27], we present a class of applications which would benefit from having such a control and flexibility over tasks and data placement. We also propose our own interpretation of the new affinity clause for the task directive, which is being discussed by the OpenMP Architecture Review Board. This clause enables the programmer to give hints to the runtime about tasks placement during the program execution, which can be used to control the data mapping on the architecture. In our proposal, the programmer can express affinity between a task and the following resources: a thread, a NUMA node, and a data. We then present an implementation of this proposal in the Clang-3.8 compiler, and an implementation of the corresponding extensions in our OpenMP runtime LIBKOMP. Finally , we present a preliminary evaluation of this work running two task-based OpenMP kernels on a 192-core NUMA architecture, that shows noticeable improvements both in terms of performance and scalability.

6.6.4. Support of High Task Throughput for Complex OpenMP Application

In [4], we present block algorithms and their implementation for the parallelization of sub-cubic Gaussian elimination on shared memory architectures using OpenMP standard. Contrarily to the classical cubic algorithms in parallel numerical linear algebra, we focus here on recursive algorithms and coarse grain parallelization. Indeed, sub-cubic matrix arithmetic can only be achieved through recursive algorithms making coarse grain block algorithms perform more efficiently than fine grain ones. This work is motivated by the design and implementation of dense linear algebra over a finite field, where fast matrix multiplication is used extensively and where costly modular reductions also advocate for coarse grain block decomposition. We incrementally build efficient kernels, for matrix multiplication first, then triangular system solving, on top of which a recursive PLUQ decomposition algorithm is built. We study the parallelization of these kernels using several algorithmic variants: either iterative or recursive and using different splitting strategies. Experiments show that recursive adaptive methods for matrix multiplication, hybrid recursive-iterative methods for triangular system solve and tile recursive versions of the PLUQ decomposition, together with various data mapping policies, provide the best performance on a 32 cores NUMA architecture. Overall, we show that the overhead of modular reductions is more than compensated by the fast linear algebra algorithms and that exact dense linear algebra matches the performance of full rank reference numerical software even in the presence of rank deficiencies.

6.7. Security for Virtualization and Clouds

Participants: Eddy Caron, Arnaud Lefray.

6.7.1. Secured Systems in Clouds with Model-Driven Orchestration

As its complexity grows, securing a system is harder than it looks. Even with efficient security mechanisms, their configuration remains a complex task. Indeed, the current practice is the hand-made configuration of these mechanisms to protect systems about which we generally lack information. Cloud computing brings its share of new security concerns but it may also be considered as leverage to overcome these issues. In [13] we adressed the key challenge of achieving global security of Cloud systems and advocate for a new approach: Model-Driven Orchestration. We have designed an implementation of this new approach called Security-Aware Models for Clouds. With this approach an industrial use-case has been deployed and secured using the Sam4C software.

6.8. Large Scale Cloud Deployment

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

6.8.1. Efficient Heuristics for Placing Large-Scale Distributed Applications on Multiple Clouds

With the fast growth of the demand for Cloud computing services, the Cloud has become a very popular platform to develop distributed applications. Features that in the past were available only to big corporations, like fast scalability, availability, and reliability, are now accessible to any customer, including individuals and small companies, thanks to Cloud computing. In order to place an application, a designer must choose among VM types, from private and public cloud providers, those that are capable of hosting her application or its parts using as criteria application requirements, VM prices, and VM resources. This procedure becomes more complicated when the objective is to place large component based applications on multiple clouds. In this case, the number of possible configurations explodes making necessary the automation of the placement. In this context, scalability has a central role since the placement problem is a generalization of the NP-Hard multi-dimensional bin packing problem.

In this work [22], we propose efficient greedy heuristics based on first fit decreasing and best fit algorithms, which are capable of computing near optimal solutions for very large applications, with the objective of minimizing costs and meeting application performance requirements. Through a meticulous evaluation, we show that the greedy heuristics took a few seconds to calculate near optimal solutions to placements that would require hours or even days when calculated using state of the art solutions, namely exact algorithms or meta-heuristics.

6.8.2. Multi-Criteria Malleable Task Management for Hybrid-Cloud Platforms

The use of large distributed computing infrastructure is a mean to address the ever increasing resource demands of scientific and commercial applications. The scale of current large-scale computing infrastructures and their heterogeneity make scheduling applications an increasingly complex task. Cloud computing minimises the heterogeneity by using virtualization mechanisms, but poses new challenges to middleware developers, such as the management of virtualization, elasticity and economic models. In this context, we proposed algorithms for efficient scheduling and execution of malleable computing tasks with high granularity while taking into account multiple optimisation criteria such as resource cost and computation time. We focused on hybrid platforms that comprise both clusters and cloud providers. In [12] we defined and formalized the main aspects of the problem, introduced the difference between local and global scheduling algorithms and evaluate their efficiency using discrete-event simulation.

6.9. Workflow management on Cloud environment

Participants: Daniel Balouek-Thomert, Eddy Caron, Laurent Lefevre.

6.9.1. Multi-objective workflow placements in Clouds

The recent rapid expansion of Cloud computing facilities triggers an attendant challenge to facility providers and users for methods for optimal placement of workflows on distributed resources, under the often-contradictory impulses of minimizing makespan, energy consumption, and other metrics. Evolutionary Optimization techniques that from theoretical principles are guaranteed to provide globally optimum solutions, are among the most powerful tools to achieve such optimal placements. Multi-Objective Evolutionary algorithms by design work upon contradictory objectives, gradually evolving across generations towards a converged Pareto front representing optimal decision variables – in this case the mapping of tasks to resources on clusters. However the computation time taken by such algorithms for convergence makes them prohibitive for real time placements because of the adverse impact on makespan. In [11], we described parallelization, on the same cluster, of a Multi-objective Differential Evolution method (NSDE-2) for optimization of workflow placement, and the attendant speedups that bring the implicit accuracy of the method into the realm of practical utility. We did experimental validation on a reallife testbed using diverse Cloud traces. The solutions under different scheduling policies demonstrate significant reduction in energy consumption with some improvement in makespan. We designed, implementation and evaluation of an energy-efficient resource management system that builds upon DIET, an open source middleware and NSDE-divisible tasks with precedence constraints. Real-life experiment of this approach on the Grid'5000 testbed demonstrates its effectiveness in a dynamic environment.

DATAMOVE Team

6. New Results

6.1. In Situ Statistical Analysis for Parametric Studies

In situ processing proposes to reduce storage needs and I/O traffic by processing results of parallel simulations as soon as they are available in the memory of the compute processes. We focus in this paper [11] on computing in situ statistics on the results of N simulations from a parametric study. The classical approach consists in running various instances of the same simulation with different values of input parameters. Results are then saved to disks and statistics are computed post mortem, leading to very I/O intensive applications. Our solution is to develop Melissa, an in situ library running on staging nodes as a parallel server. When starting, simulations connect to Melissa and send the results of each time step to Melissa as soon as they are available. Melissa implements iterative versions of classical statistical operations, enabling to update results as soon as a new time step from a simulation is available. Once all statistics ar updated, the time step can be discarded. We also discuss two different approaches for scheduling simulation runs: the jobs-in-job and the multi-jobs approaches. Experiments run instances of the Computational Fluid Dynamics Open Source solver Code_Saturne. They confirm that our approach enables one to avoid storing simulation results to disk or in memory.

6.2. Online Non-preemptive Scheduling in a Resource Augmentation Model based on Duality

Resource augmentation is a well-established model for analyzing algorithms, particularly in the online setting. It has been successfully used for providing theoretical evidence for several heuristics in scheduling with good performance in practice. According to this model, the algorithm is applied to a more powerful environment than that of the adversary. Several types of resource augmentation for scheduling problems have been proposed up to now, including speed augmentation, machine augmentation and more recently rejection. In this paper [7], we present a framework that unifies the various types of resource augmentation. Moreover, it allows generalize the notion of resource augmentation for other types of resources. Our framework is based on mathematical programming and it consists of extending the domain of feasible solutions for the algorithm with respect to the domain of the adversary. This, in turn allows the natural concept of duality for mathematical programming to be used as a tool for the analysis of the algorithm's performance. As an illustration of the above ideas, we apply this framework and we propose a primal-dual algorithm for the online scheduling problem of minimizing the total weighted flow time of jobs on unrelated machines when the preemption of jobs is not allowed. This is a well representative problem for which no online algorithm with performance guarantee is known. Specifically, a strong lower bound of $\Omega(\sqrt{n})$ exists even for the offline unweighted version of the problem on a single machine. In this paper, we first show a strong negative result even when speed augmentation is used in the online setting. Then, using the generalized framework for resource augmentation and by combining speed augmentation and rejection, we present an $(1+\epsilon_s)$-speed $O(\frac{1}{\epsilon_s \epsilon_r})$-competitive algorithm if we are allowed to reject jobs whose total weight is an ϵ_r-fraction of the weights of all jobs, for any $\epsilon_s > 0$ and $\epsilon_r \in (0,1)$. Furthermore, we extend the idea for analysis of the above problem and we propose an $(1+\epsilon_s)$-speed ϵ_r-rejection $O(\frac{k^{\frac{(k+3}{k}}}{\epsilon_r^{1/k} \epsilon_s^{\frac{(k+2)}{k}}})$-competitive algorithm for the more general objective of minimizing the weighted l_k-norm of the flow times of jobs.

6.3. Batsim: a Realistic Language-Independent Resources and Jobs Management Systems Simulator

As large scale computation systems are growing to exascale, Resources and Jobs Management Systems (RJMS) need to evolve to manage this scale modification. However, their study is problematic since they

are critical production systems, where experimenting is extremely costly due to downtime and energy costs. Meanwhile, many scheduling algorithms emerging from theoretical studies have not been transferred to production tools for lack of realistic experimental validation. To tackle these problems we propose Batsim [6], an extendable, language-independent and scalable RJMS simulator. It allows researchers and engineers to test and compare any scheduling algorithm, using a simple event-based communication interface, which allows different levels of realism. In this paper we show that Batsim's behavior matches the one of the real RJMS OAR. Our evaluation process was made with reproducibility in mind and all the experiment material is freely available.

HIEPACS Project-Team

7. New Results

7.1. High-performance computing on next generation architectures

7.1.1. Numerical recovery strategies for parallel resilient Krylov linear solvers

As the computational power of high performance computing (HPC) systems continues to increase by using a huge number of cores or specialized processing units, HPC applications are increasingly prone to faults. In this paper, we present a new class of numerical fault tolerance algorithms to cope with node crashes in parallel distributed environments. This new resilient scheme is designed at application level and does not require extra resources, i.e., computational unit or computing time, when no fault occurs. In the framework of iterative methods for the solution of sparse linear systems, we present numerical algorithms to extract relevant information from available data after a fault, assuming a separate mechanism ensures the fault detection. After data extraction, a well chosen part of missing data is regenerated through interpolation strategies to constitute meaningful inputs to restart the iterative scheme. We have developed these methods, referred to as Interpolation-Restart techniques, for Krylov subspace linear solvers. After a fault, lost entries of the current iterate computed by the solver are interpolated to define a new initial guess to restart the Krylov method. A well suited initial guess is computed by using the entries of the faulty iterate available on surviving nodes. We present two interpolation policies that preserve key numerical properties of well-known linear solvers, namely the monotonic decrease of the A-norm of the error of the conjugate gradient or the residual norm decrease of GMRES. The qualitative numerical behavior of the resulting scheme have been validated with sequential simulations, when the number of faults and the amount of data losses are varied. Finally, the computational costs associated with the recovery mechanism have been evaluated through parallel experiments.

More details on this work can be found in [7].

7.1.2. Interpolation-restart strategies for resilient eigensolvers

The solution of large eigenproblems is involved in many scientific and engineering applications when for instance, stability analysis is a concern. For large simulation in material physics or thermo-acoustics, the calculation can last for many hours on large parallel platforms. On future large-scale systems, the mean time between failures (MTBF) of the system is expected to decrease so that many faults could occur during the solution of large eigenproblems. Consequently, it becomes critical to design parallel eigensolvers that can survive faults. In that framework, we investigate the relevance of approaches relying on numerical techniques, which might be combined with more classical techniques for real large-scale parallel implementations. Because we focus on numerical remedies we do not consider parallel implementations nor parallel experiments but only numerical experiments. We assume that a separate mechanism ensures the fault detection and that a system layer provides support for setting back the environment (processes, . . .) in a running state. Once the system is in a running state, after a fault, our main objective is to provide robust resilient schemes so that the eigensolver may keep converging in the presence of the fault without restarting the calculation from scratch. For this purpose, we extend the interpolation-restart (IR) strategies initially introduced for the solution of linear systems in a previous work to the solution of eigenproblems in this paper. For a given numerical scheme, the IR strategies consist of extracting relevant spectral information from available data after a fault. After data extraction, a well-selected part of the missing data is regenerated through interpolation strategies to constitute a meaningful input to restart the numerical algorithm. One of the main features of this numerical remedy is that it does not require extra resources, i.e. , computational unit or computing time, when no fault occurs. In this paper, we revisit a few state-of-the-art methods for solving large sparse eigenvalue problems namely the Arnoldi methods, subspace iteration methods and the Jacobi-Davidson method, in the light of our IR strategies. For each considered eigensolver, we adapt the IR strategies to regenerate as much spectral information as possible. Through extensive numerical experiments, we study the respective robustness of the resulting resilient schemes with respect to the MTBF and to the amount of data loss via qualitative and quantitative illustrations.

More details on this work can be found in [8].

7.2. High performance solvers for large linear algebra problems

7.2.1. Exploiting Kepler architecture in sparse direct solver with runtime systems

Many works have addressed heterogeneous architectures to exploit accelerators such as GPUs or Intel Xeon Phi with interesting speedup. Despite researches towards generic solutions to efficiently exploit those accelerators, their hardware evolution requires continual adaptation of the kernels running on those architectures. The recent Nvidia architectures, as Kepler, present a larger number of parallel units thus requiring more data to feed every computational units. A solution considered to supply enough computation has been to study problems with large number of small computations. The batched BLAS libraries proposed by Intel, Nvidia, or the University of Tennessee are examples of this solution. We have investigated the use of the variable size batched matrix-matrix multiply to improve the performance of a the PaStiX sparse direct solver. Indeed, this kernel suits the super-nodal method of the solver, and the multiple updates of variable sizes that occur during the numerical factorization.

These contributions have been presented at the PMAA'16 conference [28].

7.2.2. 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 computations 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 10% on CPUs and up to 40% on GPUs.

These contributions have been presented at the SIAM PP'16 conference [35] and an extended paper has been submitted to SIAM Journal on Matrix Analysis and Applications [49].

7.2.3. Sparse supernodal solver using hierarchical compression

In the context of FASTLA associate team, during the last 3 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.

Sparse direct solvers such as PaStiX are currently limited by their memory requirements and computational cost. They are competitive for small matrices but are often less efficient than iterative methods for large matrices in terms of memory. We are currently accelerating the dense algebra components of direct solvers using hierarchical matrices algebra. In the first step, we are targeting an $O(N^{4/3})$ solver. Preliminary benchmarks indicate that a speed up of 2x to 10x is possible (on the largest test cases).

In the context of the FASTLA team, 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 a new implementation of the PaStiX solver using hierarchical compression to reduce the burden on large blocks appearing during the nested dissection process. To improve the efficiency of our sparse update kernel for both BLR (block low-rank) and HODLR (hierarchically off-diagonal low-rank), we are now investigating to BDLR (boundary distance low-rank) approximation scheme to preselect rows and columns in the low-rank approximation algorithm. We also have to improve our ordering strategies to enhance data locality and compressibility. The implementation is based on runtime systems to exploit parallelism.

Some contributions have already been presented at the workshops on Fast Solvers [32], [31], [30]. This work is a joint effort between Professor Darve's group at Stanford and the Inria HiePACS team within FASTLA.

7.2.4. Hierarchical hybrid sparse linear solver for multicore platforms

The solution of large sparse linear systems is a critical operation for many numerical simulations. To cope with the hierarchical design of modern supercomputers, hybrid solvers based on Domain Decomposition Methods (DDM) have been been proposed. Among them, approaches consisting of solving the problem on the interior of the domains with a sparse direct method and the problem on their interface with a preconditioned iterative method applied to the related Schur Complement have shown an attractive potential as they can combine the robustness of direct methods and the low memory footprint of iterative methods. In this report, we consider an additive Schwarz preconditioner for the Schur Complement, which represents a scalable candidate but whose numerical robustness may decrease when the number of domains becomes too large. We thus propose a two-level MPI/thread parallel approach to control the number of domains and hence the numerical behaviour. We illustrate our discussion with large-scale matrices arising from real-life applications and processed on both a modern cluster and a supercomputer. We show that the resulting method can process matrices such as tdr455k for which we previously either ran out of memory on few nodes or failed to converge on a larger number of nodes. Matrices such as Nachos_4M that could not be correctly processed in the past can now be efficiently processed up to a very large number of CPU cores (24 576 cores). The corresponding code has been incorporated into the MaPHyS package.

More details on this work can be found in [44]

7.2.5. Task-based conjugate gradient: from multi-GPU towards heterogeneous architectures

Whereas most parallel High Performance Computing (HPC) numerical libaries have been written as highly tuned and mostly monolithic codes, the increased complexity of modern architectures led the computational science and engineering community to consider more mod- ular programming paradigms such as task-based paradigms to design new generation of parallel simulation code; this enables to delegate part of the work to a third party software such as a runtime system. That latter approach has been shown to be very productive and efficient with compute-intensive algorithms, such as dense linear algebra and sparse direct solvers. In this study, we consider a much more irregular, and synchronizing algorithm, namely the Conjugate Gradient (CG) algorithm. We propose a task-based formulation of the algorithm together with a very ne instrumentation of the runtime system. We show that almost optimum speed up may be reached on a multi-GPU platform (relatively to the mono-GPU case) and, as a very preliminary but promising result, that the approach can be e ectively used to handle heterogenous architectures composed of a multicore chip and multiple GPUs. We expect that these results will pave the way for investigating the design of new advanced, irregular numerical algorithms on top of runtime systems.

More details on this work can be found in [42]

7.2.6. Analysis of rounding error accumulation in conjugate gradients to improve the maximal attainable accuracy of pipelined CG

Pipelined Krylov solvers typically offer better scalability in the strong scaling limit compared to standard Krylov methods. The synchronization bottleneck is mitigated by overlap- ping time-consuming global communications with useful computations in the algorithm. However, to achieve this communication hiding strategy, pipelined methods feature multiple recurrence re- lations on additional auxiliary variables to update the guess for the solution. This paper aims at studying the influence of rounding errors on the convergence of the pipelined Conjugate Gradient method. It is analyzed why rounding effects have a significantly larger impact on the maximal attainable accuracy of the pipelined CG algorithm compared to the traditional CG method. Fur- thermore, an algebraic model for the accumulation of rounding errors throughout the (pipelined) CG algorithm is derived. Based on this rounding error model, we then propose an automated residual replacement strategy to reduce the effect of rounding errors on the final iterative solution. The resulting pipelined CG method with automated residual replacement improves the maximal attainable accuracy of pipelined CG

to a precision comparable to that of standard CG, while maintaining the efficient parallel performance of the pipelined method.

More details on this work can be found in [46].

7.2.7. Nearly optimal fast preconditioning of symmetric positive definite matrices

We consider the hierarchical off-diagonal low-rank preconditioning of symmetric positive definite matrices arising from second order elliptic boundary value problems. When the scale of such problems becomes large combined with possibly complex geometry or unstable of boundary conditions, the representing matrix is large and typically ill-conditioned. Multilevel methods such as the hierarchical matrix approximation are often a necessity to obtain an efficient solution. We propose a novel hierarchical preconditioner that attempts to minimize the condition number of the preconditioned system. The method is based on approximating the low-rank off-diagonal blocks in a norm adapted to the hierarchical structure. Our analysis shows that the new preconditioner effectively maps both small and large eigenvalues of the system approximately to 1. Finally through numerical experiments, we illustrate the effectiveness of the new designed scheme which outperforms more classical techniques based on regular SVD to approximate the off-diagonal blocks and SVD with filtering.

This work is a joint effort between Professor Darve's group at Stanford and the Inria HiePACS team within FASTLA. More details on this work can be found in [41].

7.2.8. Robust coarse spaces for abstract Schwarz preconditioners via generalized eigenproblems

The solution of large sparse linear systems is one of the most important kernels in many numerical simulations. The domain decomposition methods (DDM) community has de- veloped many efficient and robust solvers in the last decades. While many of these solvers fall in Abstract Schwarz (AS) framework, their robustness has often been demonstrated on a case-by-case basis. In this paper, we propose a bound for the condition number of all deflated AS methods pro- vided that the coarse grid consists of the assembly of local components that contain the kernel of some local operators. We show that classical results from the literature on particular instances of AS methods can be retrieved from this bound. We then show that such a coarse grid correction can be explicitly obtained algebraically via generalized eigenproblems, leading to a condition number independent of the number of domains. This result can be readily applied to retrieve the bounds previously obtained via generalized eigenproblems in the particular cases of Neumann-Neumann (NN), additive Schwarz (aS) and optimized Robin but also generalizes them when applied with approximate local solvers. Interestingly, the proposed methodology turns out to be a comparison of the considered particular AS method with generalized versions of both NN and aS for tackling the lower and upper part of the spectrum, respectively. We furthermore show that the application of the considered grid corrections in an additive fashion is robust in the aS case although it is not robust for AS methods in general. In particular, the proposed framework allows for ensuring the robustness of the aS method applied on the Schur complement (aS/S), either with deflation or additively, and with the freedom of relying on an approximate local Schur complement, leading to a new powerful and versatile substructuring method. Numerical experiments illustrate these statements.

More details on this work can be found in [45]

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

7.3.1. Task-based fast multipole method

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. The successful ports of fully-featured numerical libraries on several recent runtime system proposals have shown, indeed, the benefit of task-based parallelism models in terms of performance portability on complex platforms. 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 a 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 as part of 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. We illustrate our discussion with the ScalFMM library, which implements state-of-the-art fast multipole method (FMM) algorithms, that we have deeply re-designed 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. We furthermore propose extensions to the OPENMP 4 standard and show how they can enhance FMM performance. To assess our statement, we have implemented this support within the KLANG-OMP source-to-source compiler that translates OPENMP directives into calls to the StarPU task-based runtime system. This study, [38] shows that we can take advantage of the advanced capabilities of a fully-featured runtime system without resorting to a specific, native runtime port, hence bridging the gap between the OPENMP standard and the very high performance that was so far reserved to expert-only runtime system APIs.

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. In [40] we illustrate our discussion with the ScalFMM FMM library and the StarPU runtime system.

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

7.4.1. Load Balancing for Coupled Simulations

In the field of scientific computing, the load balancing is an important step conditioning the performance of parallel programs. The goal is to distribute the computational load across multiple processors in order to minimize the execution time. This is a well-known problem that is unfortunately NP-hard. The most common approach to solve it is based on graph or hypergraph partitioning method, using mature and efficient software tools such as Metis, Zoltan or Scotch. Nowadays, numerical simulation are becoming more and more complex, mixing several models and codes to represent different physics or scales. Here, the key idea is to reuse available legacy codes through a coupling framework instead of merging them into a standalone application. For instance, the simulation of the earth's climate system typically involves at least 4 codes for atmosphere, ocean, land surface and sea-ice . Combining such different codes are still a challenge to reach high performance and scalability. In this context, one crucial issue is undoubtedly the load balancing of the whole coupled simulation that remains an open question. The goal here is to find the best data distribution for the whole coupled codes and not only for each standalone code, as it is usually done. Indeed, the naive balancing of each code on its own can lead to an important imbalance and to a communication bottleneck during the coupling phase, that can dramatically decrease the overall performance. Therefore, one argues that it is required to model the coupling itself in order to ensure a good scalability, especially when running on tens of thousands of processors. In this work, we develop new algorithms to perform a coupling-aware partitioning of the whole application.

Surprisingly, we observe in our experiments that our proposed algorithms do not highly degrade the global edgecut for either component and thus the internal communication among processors of the same component is still minimized. This is not the case for the *Multiconst* method especially as the number of processors increases. Regarding the coupled simulation for the real application AVTP-AVBP (provided by Cerfacs), we noticed that one may carefully decide the parameters of the co-partitioning algorithms in order not to increase the global edgecut. More precisely, the number of processors assigned in the coupling interface is an important factor that needs to be determined based on the geometry of the problem and the ratio of the coupling interface compared to the entire domain. Again, we remark that our work on co-partitioning is still theoretical and further investigation should be conducted with different geometries and more coupled simulations that are more or less coupling-intensive.

This work corresponds to the PhD of Maria Predari, defended on December 9^{th} 2016.

7.5. Application Domains

7.5.1. Material physics

7.5.1.1. Molecular vibrational spectroscopy

Quantum chemistry eigenvalue problem is a big challenge in recent research. Here we are interested in solving eigenvalue problems coming from the molecular vibrational analysis. These problems are challenging because the size of the vibrational Hamiltonian matrix to be diagonalized is exponentially increasing with the size of the molecule we are studying. So, for molecules bigger than 10 atoms the actual existent algorithms suffer from a curse of dimensionality or computational time.

A new variational algorithm called adaptive vibrational configuration interaction (A-VCI) intended for the resolution of the vibrational Schrödinger equation was developed. The main advantage of this approach is to efficiently reduce the dimension of the active space generated into the configuration interaction (CI) process. Here, we assume that the Hamiltonian writes as a sum of products of operators. This adaptive algorithm was developed with the use of three correlated conditions i.e. a suitable starting space ; a criterion for convergence, and a procedure to expand the approximate space. The velocity of the algorithm was increased with the use of a posteriori error estimator (residue) to select the most relevant direction to increase the space. Two examples have been selected for benchmark. In the case of H_2CO, we mainly study the performance of A-VCI algorithm: comparison with the variation-perturbation method, choice of the initial space, residual contributions. For CH_3CN, we compare the A-VCI results with a computed reference spectrum using the same potential energy surface and for an active space reduced by about 90 %. This work was published in [9].

7.5.1.2. Dislocations

We have focused on the improvements in collision detection in the Optidis Code. Junction formation mechanisms are essential to characterize material behavior such as strain hardening and irradiation effects. Dislocations junctions appear when dislocation segments collide with each other, therefore, reliable collision detection algorithms must be used to detect an handle junction formations. Collision detection is also a very costly operation in dislocation dynamics simulations, and performance must be carefully optimized to allow massive simulations.

During the first year of this PhD thesis, new collision algorithms have been implemented for the Dislocation Dynamics code OptiDis. The aim was to allow fast and accurate collision detection between dislocation segments using hierarchical methods. The complexity to solve the N-body collision problem can be reduced to O(N) using spatial partitioning; computation can be accelerated using fast-reject techniques, and OpenMP parallelism. Finally, new collision handling algorithms for dislocations have been implemented to increase the reliability of the simulation.

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

7.5.2.1. Interior penalty discontinuous Galerkin method for coupled elasto-acoustic media

We introduce a high order interior penalty discontinuous Galerkin scheme for the nu- merical solution of wave propagation in coupled elasto-acoustic media. A displacement formulation is used, which allows for the solution of the acoustic and elastic wave equations within the same framework. Weakly imposing the correct transmission condition is achieved by the derivation of adapted numerical fluxes. This generalization does not weaken the discontinuous Galerkin method, thus hp-non-conforming meshes are supported. Interior penalty discontinuous Galerkin methods were originally developed for scalar equations. Therefore, we propose an optimized formulation for vectorial equations more suited than the straightforward standard transposition. We prove consis- tency and stability of the proposed schemes. To study the numerical accuracy and convergence, we achieve a classic plane wave analysis. Finally, we show the relevance of our method on numerical experiments.

More details on this work can be found in [47].

7.5.2.2. 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 had been carried on in the framework of Fabien Rozar's PhD in collaboration with CEA-IRFM (defended in November 2015) and is continued in the PhD of Nicolas Bouzat funded by IPL C2S@Exa. The scientific objectives of this new work will be 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.

7.5.2.3. 3D aerodynamics for unsteady problems with bodies in relative motion

The first part of our research work concerning the parallel aerodynamic code FLUSEPA has been to design an operational MPI+OpenMP version based on a domain decomposition. We achieved an efficient parallel version up to 400 cores and the temporal adaptive method used without bodies in relative motion has been tested successfully for complex 3D take-off blast wave computations. Moreover, an asynchronous strategy for computing bodies in relative motion and mesh intersections has been developed and has been used for 3D stage separation cases. This first version is the current industrial production version of FLUSEPA for Airbus Safran Launchers.

However, this intermediate version shows synchronization problems for the aerodynamic solver due to the time integration used. To tackle this issue, a task-based version over the runtime system StarPU has been developed and evaluated. Task generation functions have been designed in order to maximize asynchronism during execution while respecting the data pattern access of the code. This led to the re-factorization of the FLUSEPA computation kernels. It's clearly a successful proof of concept as a task-based version is now available for the aerodynamic solver and for both shared and distributed memory. It uses three parallelism levels : MPI processes between sub-domains, StarPU workers in shared memory (for each sub-domain) themselves running OpenMP parallel tasks. This version has been validated for large 3D take-off blast wave computations (80 millions of cells) and is much more efficient than the previous MPI+OpenMP version: we achieve a gain in computation time equal to 70 % for 320 cores and to 50 % for 560 cores. The next step will consist in extending the task-based version to the motion and intersection operations. This work has been carried on in the framework of Jean-Marie Couteyen's PhD (defended in September 2016) in collaboration with Airbus Safran Launchers ([2], [17]).

KERDATA Project-Team

7. New Results

7.1. Convergence of HPC and Big Data

7.1.1. Transactional storage

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Concurrent Big Data applications often require high-performance storage, as well as ACID (Atomicity, Consistency, Isolation, Durability) transaction support. Although blobs (binary large objects) are an increasingly popular model for addressing the storage needs of such applications, state-of-the-art blob storage systems typically offer no transaction semantics. This demands users to coordinate access to data carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior. We argue there is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications.

Týr is the first blob storage system to provide built-in, multi-blob transactions, while retaining sequential consistency and high throughput under heavy access concurrency. Týr offers fine-grained random write access to data and in-place atomic operations.

Large-scale experiments on Microsoft Azure with a production application from CERN LHC show Týr throughput outperforms state-of-the-art solutions by more than 75 %.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.1.2. Big Data on HPC

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Over the last decade, Map-Reduce has stood as the most powerful Big Data processing model. Map-Reduce model is now used by many companies and research labs to facilitate large-scale data analysis. With the growing needs of users and size of data, commodity-based infrastructure (most commonly used as of now) will strain under the heavy weight of Big Data. On the other hand, HPC systems offer a rich set of opportunities for Big Data processing.

As first steps towards Big Data processing on HPC systems, several research efforts have been devoted to understand Map-Reduce performance on these systems. Yet, the impact of the specific features of HPC environments have not been fully investigated, yet.

We conducted an experimental campaign to provide a clearer understanding of Map-Reduce performance on HPC systems. We use Spark, a widely adopted Map-Reduce framework, and representative Big Data workloads on Grid'5000 testbed to evaluate how the latency, contention and file system's configuration can influence the application performance.

7.1.3. Energy vs. performance trade-offs

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

Most large popular web applications, like Facebook and Twitter, have been relying on large amounts of in-memory storage to cache data and provide a low response time. As the 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 still little knowledge regarding the energy efficiency of in-memory storage systems. This is unfortunate, as it is known that main memory is a major energy bottleneck in many computing systems. For instance, DRAM consumes up to 40 % of a server's power.

By means of experimental evaluation, we have studied the performance and energy-efficiency of RAMCloud — a well-known in-memory storage system. We demonstrated that although RAMCloud is scalable for read-only applications, it exhibits non-proportional power consumption. We also found that the current replication scheme implemented in RAMCloud limits the performance and results in high energy consumption. Surprisingly enough, we also showed that replication can even play a negative role in crash-recovery.

Collaboration. *This work was carried out in collaboration with Toni Cortes (BSC, Spain).*

7.2. Efficient I/O and communication for Extreme-scale HPC systems

7.2.1. Adaptive performance-constrained in situ visualisation

Participant: Lokman Rahmani.

While many parallel visualization tools now provide in situ visualization capabilities, the trend has been to feed such tools with large amounts of unprocessed output data and let them render everything at the highest possible resolution. This leads to an increased run time of simulations that still have to complete within a fixed-length job allocation.

We have been working on tackling the challenge of enabling in situ visualization under performance constraints. Our approach shuffles data across processes according to their contents and filters out part of them. Thereby, the visualization pipeline is only fed with a reorganized subset of the data produced by the simulation.

Our framework, as presented in [22], leverages fast, generic evaluation procedures to score blocks of data, using information theory, statistics, and linear algebra. It monitors its own performance and dynamically adapts to achieve appropriate visual fidelity within predefined performance constraints. Experiments on the Blue Waters supercomputer with the CM1 simulation show that our approach enables a 5-time speedup with respect to the initial visualization pipeline, and is able to meet performance constraints.

Collaboration. *This was was carried out with the collaboration of Matthieu Dorier, ANL, USA.*

7.2.2. Dragonfly

Participants: Nathanaël Cheriere, Shadi Ibrahim, Gabriel Antoniu.

High-radix direct network topologies such as Dragonfly have been proposed for Petascale and Exascale supercomputers. It has been shown that they ensure fast interconnections and reduce the cost of the network compared to traditional network topologies. However, current algorithms for communication do not consider the topology and thus waste numerous opportunities of optimization for performance.

In our studies, we exploit the strength of the Dragonfly with topology-aware algorithms for AllGather and Scatter operations. We analyze existing algorithms, then propose derived algorithms, that we evaluate using CODES, an event-driven simulator.

As expected, making AllGather algorithms topology-aware does improve the performance and reduces the link utilization. However, simulations of various Scatter algorithms show surprising results, and point out the important role played by hardware for the efficiency of the algorithms. In particular, the knowledge of the number and size of input-output buffers in routers can be exploited to accelerate the Scatter operation by a factor up to 2 times.

Collaboration. *This work was done in collaboration with Matthieu Dorier and Rob Ross, ANL, USA.*

7.2.3. Interference between HPC jobs

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

As we move toward the Exascale era, performance variability in HPC systems remains a challenge. I/O interference, a major cause of this variability, is becoming more important every day with the growing number of concurrent applications that share larger machines. Earlier research efforts on mitigating I/O interference focus on a single potential cause of interference (e.g., the network). Yet the root causes of I/O interference can be diverse.

In [27], we conducted an extensive experimental campaign to explore the various root causes of I/O interference in HPC storage systems. We used micro-benchmarks on the Grid'5000 testbed to evaluate how I/O interference is influenced by the applications' access pattern, the network components, the file system's configuration, and the backend storage devices.

Our studies revealed that in many situations interference is a result of a bad flow control in the I/O path, rather than being caused by some single bottleneck in one of its components. We further show that interference-free behavior is not necessarily a sign of optimal performance. To the best of our knowledge, our work provides the first deep insight into the role of each of the potential root causes of interference and their interplay. Our findings can help developers and platform owners improve I/O performance and motivate further research addressing the problem across all components of the I/O stack.

Collaboration. *This work was done in collaboration with Matthieu Dorier and Rob Ross, ANL, USA.*

7.3. Workflow on clouds

7.3.1. Managing hot metadata for scientific workflows on multisite clouds

Participants: Luis Eduardo Pineda Morales, Alexandru Costan, Gabriel Antoniu.

Large-scale scientific applications are often expressed as workflows that help defining data dependencies between their different components. Such workflows may incur huge storage and computation requirements, so that they need to be processed in multiple (cloud-federated) datacenters. A major challenge in such multisite clouds is the long latency of the network links between datacenters, that limits the performance of multisite applications. Moreover, it has been shown that poor metadata handling can further impact the efficiency of computing systems. Many efforts have been done to improve metadata management; however, most of them concern only single-site, HPC systems to date.

In [26], we assert that some workflow metadata are more frequently accessed than other, and thus should be handled with higher priority during the workflow's execution. We call them *hot metadata*. We present a hybrid decentralized/distributed model for handling hot metadata in *multisite* architectures. We couple our model with a scientific workflow management system (SWfMS) to validate and tune its applicability to various real-life scientific scenarios. We show that efficient management of hot metadata improves the performance of SWfMS, reducing the workflow execution time up to 50 % for highly parallel jobs by enabling timely data provisioning and avoiding unnecessary *cold* metadata operations.

7.3.2. Probabilistic optimizations for resource provisioning of cloud workflows

Participants: Chi Zhou, Shadi Ibrahim.

In many data-intensive applications, data management routines can be represented as workflows, where tasks are organized according to data and computation dependencies. Recently, the optimal provisioning of resources (e.g., VMs) for workflows running in the cloud has attracted a lot of attention. Most resource provisioning solutions overlook the important factor of cloud dynamics, e.g., the fluctuation of I/O, network performance, and system failures. In our experiments on the Amazon EC2 cloud, these issues significantly impact resource allocation quality. Therefore, we study how cloud dynamics should be incorporated into the resource provisioning process.

Our approach models cloud dynamics as time-dependent random variables (e.g., a probability distribution of workflow execution times) and performs probabilistic optimizations for resource provisioning problems using those random variables as optimization input. This solution yields more effective resource provisioning for cloud workflows. However, it involves heavy computation effort due to the complex structures of workflows and the large number of probability calculations.

To overcome this problem, we develop a three-stage pruning process, which simplifies workflow structure and reduces probability evaluation overhead. We have also implemented our techniques in a runtime library, which allows users to integrate our techniques into their existing resource provisioning methods. Experiments on two common resource provisioning problems show that probabilistic solutions can improve the performance by 51 % —70 % compared with state-of-the-art, static solutions.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.3.3. A taxonomy and survey of scientific computing in the cloud

Participants: Chi Zhou, Shadi Ibrahim.

Cloud computing has evolved as a popular computing infrastructure for many applications. With (big) data acquiring a crucial role in eScience, efforts have been made recently to develop and deploy scientific applications efficiently on the unprecedentedly scalable cloud infrastructures.

In [29], we review recent efforts in developing and deploying scientific computing applications in the cloud. In particular, we introduce a taxonomy specifically designed for scientific computing in the cloud, and further review the taxonomy with four major kinds of science applications, including life sciences, physics sciences, social and humanities sciences, and climate and earth sciences.

Due to the large data size in most scientific applications, the performance of I/O operations can greatly affect the overall performance of the applications. As a consequence, the dynamic I/O performance of the cloud has made resource provisioning an important and complex problem for scientific applications in the cloud.

We present our efforts on improving the resource provisioning efficiency and effectiveness of scientific applications in the cloud. Finally, we present the open problems for developing the next-generation eScience applications and systems in the cloud and give our conclusions.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.4. Fault tolerant data processing

7.4.1. Fast recovery

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop has emerged as a prominent tool for Big Data processing in large-scale clouds. Failures are inevitable in large-scale systems, especially in shared environments. Consequently, Hadoop was designed with hardware failures in mind. In particular, Hadoop handles machine failures by re-executing all the tasks of the failed machine. Unfortunately, the efforts to handle failures are entirely entrusted to the core of Hadoop and hidden from Hadoop schedulers. This may prevent Hadoop schedulers from meeting their objectives (e.g., fairness, job priority, performance) and can significantly impact the performance of the applications.

In our previous work, we addressed this issue through the design and implementation of a new scheduling strategy called Chronos. Chronos is conductive to improving the performance of Map-Reduce applications by enabling an early action upon failure detection. Chronos tries to launch recovery tasks immediately by preempting tasks belonging to low priority jobs, thus avoiding to wait until slots are freed.

In [20], we further investigated the potential benefit of launching local recovery tasks by implementing and evaluating Chronos*. To this end, we slightly changed the smart slot allocation strategy of Chronos into aggressive slot allocation strategy. With Chronos, recovery tasks with higher priority would preempt the selected tasks with less priority. With Chronos*, we also allow recovery tasks to preempt the selected tasks with the same priority (e.g., recovery tasks belonging to the same job with selected tasks). The experimental results indicate that Chronos* results in 100 % locality execution for recovery tasks thanks to its aggressive slot allocation strategy. Moreover, Chronos* improves the completion time of the jobs by up to 17 %.

7.4.2. Dynamic replica placement

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Large-scale applications are ever-increasingly geo-distributed. Maintaining the highest possible *data locality* is crucial to ensure high performance of such applications. Dynamic replication addresses this problem by dynamically creating replicas of frequently accessed data close to the clients. This data is often stored in decentralized storage systems such as Dynamo or Voldemort, which offer support for *mutable data*.

However, existing approaches to dynamic replication for such mutable data remain centralized, thus incompatible with these systems. We introduce a write-enabled dynamic replication scheme that leverages the decentralized architecture of such storage systems. We propose an algorithm enabling clients to locate tentatively the closest data replica without prior request to any metadata node. Large-scale experiments show a read latency decrease of up to 42% compared to other state-of-the-art, caching-based solutions.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.5. Advanced data management on clouds

7.5.1. Benchmarking Spark and Flink

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

Spark and Flink are two Apache-hosted data analytics frameworks that represent the state of the art in modern in-memory Map-Reduce processing. They facilitate the development of multi-step data pipelines using directly acyclic graph (DAG) patterns. In the framework of our BigStorage project, we performed a comparative study [23] which evaluates the performance of Spark versus Flink. The objective is to identify and explain the impact of the different architectural choices and the parameter configurations on the perceived end-to-end performance.

Based on empirical evidences, the study points out that in Big Data processing there is not a single framework for all data types, sizes and job patterns and emphasize a set of design choices that play an important role in the behaviour of a Big Data framework: memory management, pipelined execution, optimizations and parameter configuration easiness. What raises our attention is that a streaming engine (i.e., Flink) delivers in many benchmarks better performance than a batch-based engine (i.e., Spark), showing that a more general Big Data architecture (treating batches as finite sets of streamed data) is plausible and may subsume both streaming and batching use cases.

Collaboration. *This work was done in collaboration with María Pérez, UPM, Spain.*

7.5.2. Geo-distributed graph processing

Participants: Chi Zhou, Shadi Ibrahim.

Graph processing is an emerging model adopted by a wide range of applications to easily parallelize the computations over graph data. Partitioning graph processing workloads to multiple machines is an important task for reducing the communication cost and improving the performance of graph processing jobs. Recently, many real-world applications store their data on multiple geographically distributed datacenters (DCs) to ensure flexible and low-latency services. Due to the limited Wide Area Network (WAN) bandwidths and the network heterogeneity of the geo-distributed DCs, existing graph partitioning methods need to be redesigned to improve the performance of graph processing jobs in geo-distributed DCs.

To address the above challenges, we propose a heterogeneity-aware graph partitioning method named G-Cut, which aims at minimizing the runtime of graph processing jobs in geo-distributed DCs while satisfying the WAN usage budget. G-Cut is a two-stage graph partitioning method. In the traffic-aware graph partitioning stage, we adopt the one-pass edge assignment to place edges into different partitions while minimizing the inter-DC data traffic size. In the network-aware partition refinement stage, we map the partitions obtained in the first stage onto different DCs in order to minimize the inter-DC data transfer time. We evaluate the effectiveness and efficiency of G-Cut using real-world graphs and the evaluation results show that G-Cut can achieve both lower WAN usage and shorter inter-DC data transfer time compared to state-of-the-art graph partitioning methods.

Collaboration. *This work was done in collaboration with Bingsheng He NUS, Singapore.*

7.5.3. Fairness and scheduling

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Recently, Map-Reduce and its open-source implementation Hadoop have emerged as prevalent tools for big data analysis in the cloud. Fair resource allocation in-between jobs and users is an important issue, especially in multi-tenant environments such as clouds. Several scheduling policies have been developed to preserve fairness in multi-tenant Hadoop clusters. At the core of these schedulers, simple (non-) preemptive approaches are employed to free resources for tasks belonging to jobs with less share. For example, Hadoop Fair Scheduler is equipped with two approaches: wait and kill. While wait may introduce a serious violation in fairness, kill may result in a huge waste of resources. Yet, recently some work have introduced preemption approach in shared Hadoop clusters.

To this end, we closely examine three approaches including wait, kill and preemption when Hadoop Fair Scheduler is employed for ensuring fair execution between multiple concurrent jobs. We perform extensive experiments to assess the impact of these approaches on performance and resource utilization while ensuring fairness. Our experimental results bring out the differences between these approaches and illustrate that these approaches are only sub-optimal for different workloads and cluster configurations: the efficiency of achieving fairness and the overall performance varies with the workload composition, resource availability and the cost of the adopted preemption technique.

7.5.4. Stragglers in Map-Reduce

Participants: Tien-Dat Phan, Shadi Ibrahim.

Big Data systems (e.g., Map-Reduce, Hadoop, Spark) rely increasingly on speculative execution to mask slow tasks also known as stragglers because a job's execution time is dominated by the slowest task instance. Big Data systems typically identify stragglers and speculatively run copies of those tasks with the expectation a copy may complete faster to shorten job execution times.

There is a rich body of recent results on straggler mitigation in Map-Reduce. However, the majority of these do not consider the problem of accurately detecting stragglers. Instead, they adopt a particular straggler detection approach and then study its effectiveness in terms of performance, e.g., reduction in job completion time, or its efficiency, e.g., extra resource usage.

In this work, we consider a complete framework for straggler detection and mitigation. We start with a set of metrics that can be used to characterizes and detect stragglers such as Precision, Recall, Detection Latency, Undetected Time and Fake Positive. We then develop an architectural model by which these metrics can be linked to measures of performance including execution time and system energy overheads.

We further conduct a series of experiments to demonstrate which metrics and approaches are more effective in detecting stragglers and are also predictive of effectiveness in terms of performance and energy efficiency. For example, our results indicate that the default Hadoop straggler detector could be made more effective. In certain cases, precision is low and only 65 % of those detected are actual stragglers and recall, i.e., the proportion of stragglers which are actually detected, is also relatively low at 56 %. For the same case, the hierarchical approach (i.e., a green-driven detector based on the default one) achieves a precision of 98 % and a recall of 33 %.

Further, these increases in precision can be used to achieve lower execution time and energy consumption, and thus higher performance and energy efficiency. Compared to the default Hadoop mechanism, energy consumption is reduced by almost 30 %. These results demonstrate how our framework can offer useful insights and be applied in practical settings to characterize and design new straggler detection mechanisms for Map-Reduce systems.

Collaboration. *This work was carried out in collaboration with Guillaume Aupy and Padma Raghavan whilst they were affiliated with Vanderbilt University, USA.*

POLARIS Team

6. New Results

6.1. Asymptotic Models

The analysis of a set of n stochastic entities interacting with each others can be particularly difficult. The *mean field approximation* is a very effective technique to characterize the transient probability distribution or steady-state regime of such systems when the number of entities n grows very large. The idea of mean-field approximation is to replace a complex stochastic system by a simpler deterministic dynamical system. This dynamical system is constructed by assuming that the objects are asymptotically independent. Each object is viewed as interacting with an average of the other objects (the *mean*-field). When each object has a finite or countable state-space, this dynamical system is usually a non-linear ordinary differential equation (ODE). An introduction to these techniques is provided in the book chapter [29].

- Mean-field games model the rational behavior of an infinite number of indistinguishable players in interaction [79]. An important assumption of mean-field games is that, as the number of player is infinite, the decisions of an individual player do not affect the dynamics of the mass. Each player plays against the mass. A mean-field equilibrium corresponds to the case when the optimal decisions of a player coincide with the decisions of the mass. This leads to a simpler computation of the equilibrium.

  It has been shown in [72], [96] that for some games with a finite number of players, the Nash equilibria converge to mean-field equilibria as the number of players tends to infinity. Hence, many authors argue that mean-field games are a good approximation of symmetric stochastic games with a large number of players. The classical argument is that the impact of one player becomes negligible when the number of players goes to infinity. In [17], [36], we show that, in general, this convergence does not hold. We construct an example for which the mean-field limit only describes a sub-set of the limiting equilibria. Each finite-player game has an equilibrium with a good social cost, this is not the case for the limit game.
- 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 [21], [3], [20] we introduce approximations to determine the cache hit probability of two classes of cache replacement algorithms: the recently introduced h-LRU and LRU(m). These approximations only require the requests to be generated according to a general Markovian arrival process (MAP). This includes phase-type renewal processes and the IRM model as special cases. We provide both numerical and theoretical support for the claim that the proposed TTL approximations are asymptotically exact. We further show, by using synthetic and trace-based workloads, that h-LRU and LRU(m) perform alike, while the latter requires less work when a hit/miss occurs.
- In [16], we consider stochastic models in presence of uncertainty, originating from lack of knowledge of parameters or by unpredictable effects of the environment. We focus on population processes, encompassing a large class of systems, from queueing networks to epidemic spreading. We set up a formal framework for imprecise stochastic processes, where some parameters are allowed to vary in time within a given domain, but with no further constraint. We then consider the limit behaviour of these systems as the population size goes to infinity. We prove that this limit is given by a differential inclusion that can be constructed from the (imprecise) drift. We also we discuss different numerical algorithms to compute bounds of the so-obtained differential inclusions. We are currently working on an implementation of these algorithms in a numerical toolbox.

- In [37], we develop a fluid-limit approach to compute the expected absorbing time T_n of a n-dimensional discrete time Markov chain. We show that the random absorbing time T_n is well approximated by a deterministic time t_n that is the first time when a fluid approximation of the chain approaches the absorbing state at a distance $1/n$. 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.

6.2. Simulation

Simgrid is a toolkit providing core functionalities for the simulation of distributed applications in heterogeneous distributed environments. Although it was initially designed to study large distributed computing environments such as grids, we have recently applied it to performance prediction of HPC configurations.

- Finite difference methods are, in general, well suited to execution on parallel machines and are thus commonplace in High Performance Computing. Yet, despite their apparent regularity, they often exhibit load imbalance that damages their efficiency. In [38], we characterize the spatial and temporal load imbalance of Ondes3D, a seismic wave propagation simulator used to conduct regional scale risk assessment. Our analysis reveals that this imbalance originates from the structure of the input data and from low-level CPU optimizations. We then show that the CHARM++ runtime can effectively dynamically rebalance the load by migrating data and computation at the granularity of an MPI rank. We propose a methodology that leverages the capabilities of the SimGrid simulation framework and allows to conduct an experimental study at low computational cost.
- The article [35] summarizes 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.

6.3. Trace and Statistical Analysis

- In [19], we present visual analysis techniques to evaluate the performance of HPC task-based applications on hybrid architectures. Our approach is based on composing modern data analysis tools (pjdump, R, ggplot2, plotly), enabling an agile and flexible scripting framework with minor development cost. We validate our proposal by analyzing traces from the full-fledged implementation of the Cholesky decomposition available in the MORSE library running on a hybrid (CPU/GPU) platform. The analysis compares two different workloads and three different task schedulers from the StarPU runtime system. Our analysis based on composite views allows to identify allocation mistakes, priority problems in scheduling decisions, GPU tasks anomalies causing bad performance, and critical path issues.
- Media events are an area of major concern for the science of territory, with a combination of empirical, methodological and theoretical fields of research. The paper [22] presents three variations of increasing complexity around the questions of the application of the concepts of “territory”, “territoriality” and “territorialization” to the description of media events. Each variation is illustrated by recent results from the research project ANR Geomedia on a corpus of international RSS flows produced by newspapers of French, English and Spanish language located in various countries of the world.

6.4. Electricity Markets

The increased penetration of renewable energy sources in existing power systems has led to necessary developments in electricity market mechanisms. Most importantly, renewable energy generation is increasingly made accountable for deviations between scheduled and actual energy generation. However, there is no mechanism to enforce accountability for the additional costs induced by power fluctuations. These costs are socialized and eventually supported by electricity customers. In [1], we propose some metrics for assessing the contribution of all market participants to power regulation needs, as well as an attribution mechanism for fairly redistributing related power regulation costs. We discuss the effect of various metrics used by the attribution mechanisms, and we illustrate, in a game-theoretical framework, their consequences on the strategic behavior of market participants. We also illustrate, by using the case of Western Denmark, how these mechanisms may affect revenues and the various market participants.

6.5. Power control in random wireless networks

Ever since the early development stages of wireless networks, the importance of radiated power has made power control an essential component of network design. In [13], we analyzed the problem of power control in large, random wireless networks that are obtained by "erasing" a finite fraction of nodes from a regular d-dimensional lattice of N transmit-receive pairs. Drawing on tools and ideas from statistical physics, we showed that this problem can be mapped to the Anderson impurity model for diffusion in random media; in this way, by employing the so-called *coherent potential approximation* (CPA) method, we calculated the average power in the system (and its variance) for 1-D and 2-D networks. In this regard, even though infinitely large systems are always unstable beyond a critical value of the users' SINR target, finite systems remain stable with high probability even beyond this critical SINR threshold. We calculated this probability by analyzing the density of low lying eigenvalues of an associated random Schrödinger operator, and we showed that the network can exceed this critical SINR threshold by a factor of at least $O((\log N)^{-2/d})$ before undergoing a phase transition to the unstable regime.

6.6. Energy efficiency in wireless networks

[6] The recent increase in the use of wireless networks for video transmission has led to the increase in the use of rate-adaptive protocols to maximize the resource utilization and increase the efficiency in the transmission. However, a number of these protocols lead to interactions among the users that are subjective in nature and affect the overall performance. In [6], we analyzed the interplay between the wireless network and video transmission dynamics in the light of subjective perceptions of the end users in their interactions – specifically, the trade-off between maximizing the quality of service (QoS) or quality of experience (QoE) and minimizing the transmission cost. By using methods from game theory, we derived an optimized transmission scheme that allows the efficient use of traditional protocols by taking into account the subjective interactions that occur in practical scenarios.

6.7. Cognitive radio and beyond

In cognitive radio networks, secondary (unlicensed) users (SUs) can access the spectrum opportunistically, whenever they sense an opening by the network's primary (licensed) users (PUs). In [7], we analyzed the minimization of overall power consumption over several orthogonal frequency bands under constraints on the minimum quality of service (QoS) and maximum peak and average interference to the network's PUs. To that end, we proposed a projected sub-gradient algorithm which quickly converges to an optimal configuration if the users' channels are fast fading.

Despite such benefits, the conventional cognitive radio network (CCRN) paradgim is not particularly attractive for opportunistic spectrum access because the network's PUs can recapture SU channels at will, thus interrupting the transmission of the latter. To address this crucial limitation, we proposed in [24] a semi-cognitive radio network (SCRN) paradigm where PUs are constrained to first use any free channels before being allowed to capture channels that are in use by SUs. These constraints slightly degrade the performance of the network's PUs, but *a*) they offer remarkable performance improvements to the network's SUs; and *b*)

they can be compensated by imposing a monetary (or other) penalty to the network's secondary owners. In [24], we provided a game-theoretic analysis of the performance trade-offs involved for both the PUs and the SUs, and we derived both centralized and distributed learning algorithms that allow the system control process to converge to a stable state.

6.8. Online resource allocation in dynamic wireless networks

The vast majority of works on wireless resource allocation (spectrum, power, etc.) has focused on two limit cases: In the *static regime*, the attributes of the network are assumed effectively static and the system's optimality analysis relies on techniques from (static) optimization. On the other hand, in the so-called *stochastic regime*, the network is assumed to evolve randomly following some fixed probability law, and the allocation of wireless resources is optimized using tools from stochastic optimization and control. In practical wireless networks however, both assumptions fail because of factors that introduce an unpredictable variability to the system (such as user mobility, users going arbitrarily on- and off-line, etc.).

The works [15], [27], [28] treat this problem by providing no-regret learning algorithms for single-user rate maximization and power control in multi-carrier cognitive radio and Internet of Things networks. The extension of these works to multi-antenna systems was carried out in [44], where we derived a matrix exponential learning algorithm for dynamic power allocation and control in time-varying MIMO systems. Building on this, we also showed in [8] that regret minimization techniques can also be applied to the much more challenging problem of energy efficiency maximization in dynamic networks – i.e. the maximization of successfully received bits per Watt of transmitted power in environments that fluctuate unpredictably over time. Finally, as was shown in [39], [23], [9], these unilateral performance gains also extend to large networks comprising hundreds (or even thousands) of users: there, the proposed matrix exponential learning algorithm converges to a stable state within a few iterations, even for very large of antennas and subcarriers.

6.9. Adaptive multi-path routing

Routing plays a crucial part in the efficient operation of packet-switched data networks, especially with regard to latency reduction and energy efficiency. However, in addition to being distributed (so as to cope with the prolific size of today's networks), optimized routing schemes must also be able to adapt to changes in the underlying network (e.g. due to variations in traffic demands, link quality, etc.).

First, to address the issue of latency reduction, we provided in [32] an adaptive multi-flow routing algorithm to select end-to-end paths in packet-switched networks. The algorithm is based only on local information, so it is suitable for distributed implementation; furthermore, it provides guarantees that the network configuration converges to a stable state and exhibits several robustness properties that make it suitable for use in dynamic real-life networks (such as robustness to measurement errors, outdated information and update desynchronization).

Concerning energy efficiency, [41] examines the problem of routing in optical networks with the aim of minimizing traffic-driven power consumption. To tackle this, [41] proposed a pricing scheme which, combined with a distributed learning method based on the Boltzmann distribution of statistical mechanics, exhibits remarkable operation properties even under uncertainty. Specifically, the long-term average of the network's power consumption converges quickly to its minimum value (in practice, within a few iterations of the algorithm), and this convergence remains robust in the face of uncertainty of arbitrarily high magnitude.

6.10. Learning in finite games

One of the most widely used algorithms for learning in finite games is the so-called *best response algorithm* (BRA); nonetheless, even though sevaral worst-case bounds are known for its convergence time, the algorithm's performance in typical game-theoretic scenarios seems to be far better than these worst-case bounds suggest. In [26], [18], [25], [31], we computed the average execution time of the BR algorithm using Markov chain coupling techniques that recast the average execution time of this discrete algorithm as the solution of an ordinary differential equation. In so doing, we showed that the worst-case complexity of the BR algorithm

in a potential game with N players and A actions per player is $AN(N-1)$, while its average complexity over random potential games is $O(N)$, independently of A.

In [34], we also studied the convergence rate of the HEDGE algorithm (which, contrary to the BR algorithm, leads to no regret even in adversarial settings). Motivated by applications to data networks where fast convergence is essential, we analyzed the problem of learning in generic N-person games that admit Nash equilibria in pure strategies. Despite the (unbounded) uncertainty in the players' observations, we show that hedging eliminates dominated strategies (a.s.) and, with high probability, it converges locally to pure Nash equilibria at the exponential rate $O(\exp(-c\sum_{j=1}^{t}\gamma_j))$, where γ_j is the algorithm s step size.

These results are strongly related to the long-term rationality properties (elimination of dominated strategies, convergence to pure Nash equilibria and evolutionarily stable states, etc.) of an underlying class of game dynamics based on regularization and Riemannian geometry. Specifically, in [42], we introduced a class of evolutionary game dynamics whose defining element is a state-dependent geometric structure on the set of population states. When this geometric structure satisfies a certain integrability condition, the resulting dynamics preserve many further properties of the replicator and projection dynamics and are equivalent to a class of reinforcement learning dynamics studied in [10]. Finally, as we showed in [2], these properties also hold even in the presence of noise, i.e. when the players only have noisy observations of their payoff vectors.

6.11. Learning in games with continuous action spaces

A key limitation of existing game-theoretic learning algorithms is that they invariably revolve around games with a finite number of actions per players. However, this assumption is often unrealistic (especially in network-based applications of game theory), a factor which severely limits the applicability of learning techniques in real-life problems.

To address this issue, we studied in [14] a class of control problems that can be formulated as potential games with continuous action sets, and we proposed an actor-critic reinforcement learning algorithm that provably converges to equilibrium in said class. The method employed is to analyse the learning process under study through a mean-field dynamical system that evolves in an infinite-dimensional function space (the space of probability distributions over the players' continuous controls). To do so, we extend the theory of finite-dimensional two-timescale stochastic approximation to an infinite-dimensional, Banach space setting, and we proved that the continuous dynamics of the process converge to equilibrium in the case of potential games. These results combine to give a provably-convergent learning algorithm in which players do not need to keep track of the controls selected by the other agents.

Finally, to address cases where mixing over a continuum of actions is unrealistic, we examined in [40] the convergence properties of a class of learning schemes for N-person games with continuous action spaces based on a continuous optimization technique known as "dual averaging". To study this multi-agent, pure-strategy learning process, we introduced the notion of *variational stability* (VS), and we showed that stable equilibria are locally attracting with high probability whereas globally stable states are globally attracting with probability 1. Finally, we examined the scheme's convergence speed and we showed that if the game admits a strict equilibrium and the players' mirror maps are surjective, then, with high probability, the process converges to equilibrium in a finite number of steps, no matter the level of uncertainty in the players' observations (or payoffs).

6.12. Stochastic optimization

A key feature of modern data networks is their distributed nature and the stochasticity surrounding users and their possible actions. To account for these issues in a general optimization context, we proposed in [4] a distributed, asynchronous algorithm for stochastic semidefinite programming which is a stochastic approximation of the continous-time matrix exponential scheme derived in [9]. This algorithm converges almost surely to an ϵ-approximation of an optimal solution requiring only an unbiased estimate of the gradient of the problem's stochastic objective. When applied to throughput maximization in wireless multiple-input and multiple-output (MIMO) systems, the proposed algorithm retains its convergence properties under a wide array

of mobility impediments such as user update asynchronicities, random delays and/or ergodically changing channels.

More generally, in view of solving convex optimization problems with noisy gradient input, we also analyzed in [43] the asymptotic behavior of gradient-like flows that are subject to stochastic disturbances. For concreteness, we focused on the widely studied class of mirror descent methods for constrained convex programming and we examined the dynamics' convergence and concentration properties in the presence of noise. In the small noise limit, we showed that the dynamics converge to the solution set of the underlying problem with probability 1. Otherwise, in the case of persistent noise, we estimated the measure of the dynamics' long-run concentration around interior solutions and their convergence to boundary solutions that are sufficiently "robust". Finally, we showed that a rectified variant of the method with a decreasing sensitivity parameter converges irrespective of the magnitude of the noise or the structure of the underlying convex program, and we derived an explicit estimate for its rate of convergence.

6.13. Benchmarking

In modern High Performance Computing architectures, the memory subsystem is a common performance bottleneck. When optimizing an application, the developer has to study its memory access patterns and adapt accordingly the algorithms and data structures it uses. The objective is twofold: on one hand, it is necessary to avoid missuses of the memory hierarchy such as false sharing of cache lines or contention in a NUMA interconnect. On the other hand, it is essential to take advantage of the various cache levels and the memory hardware prefetcher. Still, most profiling tools focus on CPU metrics. The few of them able to provide an overview of the memory patterns involved by the execution rely on hardware instrumentation mechanisms and have two drawbacks. The first one is that they are based on sampling which precision is limited by hardware capabilities. The second one is that they trace a subset of all the memory accesses, usually the most frequent, without information ab out the other ones. In [30] we present Moca, an efficient tool for the collection of complete spatio-temporal memory traces. Moca is based on a Linux kernel module and provides a coarse grained trace of a superset of all the memory accesses performed by an application over its addressing space during the time of its execution. The overhead of Moca is reasonable when taking into account the fact that it is able to collect complete traces which are also more precise than the ones collected by comparable tools.

Benchmarking has proven to be crucial for the investigation of the behavior and performances of a system. However, the choice of relevant benchmarks still remains a challenge. To help the process of comparing and choosing among benchmarks, in [33] we propose a solution for automatic benchmark profiling. It computes unified benchmark profiles reflecting benchmarks' duration, function repartition, stability, CPU efficiency, parallelization and memory usage. It identifies the needed system information for profile computation, collects it from execution traces and produces profiles through efficient and reproducible trace analysis treatments. The paper presents the design, implementation and the evaluation of the approach.

ROMA Project-Team

7. New Results

7.1. A backward/forward recovery approach for the preconditioned conjugate gradient method

Participants: Massimiliano Fasi [Univ. Mancherster, UK], Julien Langou [UC Denver, USA], Yves Robert, Bora Uçar.

Several recent papers have introduced a periodic verification mechanism to detect silent errors in iterative solvers. Chen [PPoPP'13, pp. 167-176] has shown how to combine such a verification mechanism (a stability test checking the orthogonality of two vectors and recomputing the residual) with checkpointing: the idea is to verify every d iterations, and to checkpoint every $c \times d$ iterations. When a silent error is detected by the verification mechanism, one can rollback to and re-execute from the last checkpoint. In this work, we also propose to combine checkpointing and verification, but we use algorithm-based fault tolerance (ABFT) rather than stability tests. ABFT can be used for error detection, but also for error detection and correction, allowing a forward recovery (and no rollback nor re-execution) when a single error is detected. We introduce an abstract performance model to compute the performance of all schemes, and we instantiate it using the preconditioned conjugate gradient algorithm. Finally, we validate our new approach through a set of simulations.

This work has been accepted for publication in the *Journal of Computational Science* [13].

7.2. High performance parallel algorithms for the tucker decomposition of sparse tensors

Participants: Oguz Kaya, Bora Uçar.

We investigate an efficient parallelization of a class of algorithms for the well-known Tucker decomposition of general N-dimensional sparse tensors. The targeted algorithms are iterative and use the alternating least squares method. At each iteration, for each dimension of an N-dimensional input tensor, the following operations are performed: (i) the tensor is multiplied with $N-1$ matrices (TTMc step); (ii) the product is then converted to a matrix; and (iii) a few leading left singular vectors of the resulting matrix are computed (TRSVD step) to update one of the matrices for the next TTMc step. We propose an efficient parallelization of these algorithms for the current parallel platforms with multicore nodes. We discuss a set of preprocessing steps which takes all computational decisions out of the main iteration of the algorithm and provides an intuitive shared-memory parallelism for the TTM and TRSVD steps. We propose a coarse and a fine-grain parallel algorithm in a distributed memory environment, investigate data dependencies, and identify efficient communication schemes. We demonstrate how the computation of singular vectors in the TRSVD step can be carried out efficiently following the TTMc step. Finally, we develop a hybrid MPI-OpenMP implementation of the overall algorithm and report scalability results on up to 4096 cores on 256 nodes of an IBM BlueGene/Q supercomputer.

This work has been published at *ICPP'16* [28].

7.3. Preconditioning techniques based on the Birkhoff–von Neumann decomposition

Participants: Michele Benzi [Emory University, Atlanta, USA], Bora Uçar.

We introduce a class of preconditioners for general sparse matrices based on the Birkhoff–von Neumann decomposition of doubly stochastic matrices. These preconditioners are aimed primarily at solving challenging linear systems with highly unstructured and indefinite coefficient matrices. We present some theoretical results and numerical experiments on linear systems from a variety of applications.

This work has been accepted for publication in the journal *Computational Methods in Applied Mathematics* [10].

7.4. Parallel CP decomposition of sparse tensors using dimension trees

Participants: Oguz Kaya, Bora Uçar.

Tensor factorization has been increasingly used to address various problems in many fields such as signal processing, data compression, computer vision, and computational data analysis. CANDECOMP/PARAFAC (CP) decomposition of sparse tensors has successfully been applied to many well-known problems in web search, graph analytics, recommender systems, health care data analytics, and many other domains. In these applications, computing the CP decomposition of sparse tensors efficiently is essential in order to be able to process and analyze data of massive scale. For this purpose, we investigate an efficient computation and parallelization of the CP decomposition for sparse tensors. We provide a novel computational scheme for reducing the cost of a core operation in computing the CP decomposition with the traditional alternating least squares (CP-ALS) based algorithm. We then effectively parallelize this computational scheme in the context of CP-ALS in shared and distributed memory environments, and propose data and task distribution models for better scalability. We implement parallel CP-ALS algorithms and compare our implementations with an efficient tensor factorization library, using tensors formed from real-world and synthetic datasets. With our algorithmic contributions and implementations, we report up to 3.95x, 3.47x, and 3.9x speedups in sequential, shared memory parallel, and distributed memory parallel executions over the state of the art, and up to 1466x overall speedup over the sequential execution using 4096 cores on an IBM BlueGene/Q supercomputer.

This work is described in a technical report [49].

7.5. Scheduling series-parallel task graphs to minimize peak memory

Participants: Enver Kayaaslan, Thomas Lambert, Loris Marchal, Bora Uçar.

We consider a variant of the well-known, NP-complete problem of minimum cut linear arrangement for directed acyclic graphs. In this variant, we are given a directed acyclic graph and asked to find a topological ordering such that the maximum number of cut edges at any point in this ordering is minimum. In our main variant the vertices and edges have weights, and the aim is to minimize the maximum weight of cut edges in addition to the weight of the last vertex before the cut. There is a known, polynomial time algorithm [Liu, SIAM J. Algebra. Discr., 1987] for the cases where the input graph is a rooted tree. We focus on the variant where the input graph is a directed series-parallel graph, and propose a polynomial time algorithm. Directed acyclic graphs are used to model scientific applications where the vertices correspond to the tasks of a given application and the edges represent the dependencies between the tasks. In such models, the problem we address reads as minimizing the peak memory requirement in an execution of the application. Our work, combined with Liu's work on rooted trees addresses this practical problem in two important classes of applications.

This work is described in a technical report [50].

7.6. Matrix symmetrization and sparse direct solvers

Participants: Raluca Portase [Cluj Napoca, Romania], Bora Uçar.

We investigate algorithms for finding column permutations of sparse matrices in order to have large diagonal entries and to have many entries symmetrically positioned around the diagonal. The aim is to improve the memory and running time requirements of a certain class of sparse direct solvers. We propose efficient algorithms for this purpose by combining two existing approaches and demonstrate the effect of our findings in practice using a direct solver. In particular, we show improvements in a number of components of the running time of a sparse direct solver with respect to the state of the art on a diverse set of matrices.

This work is described in a technical report [53].

7.7. Robust Memory-Aware Mapping for Parallel Multifrontal Factorizations

Participants: Emmanuel Agullo [HIEPACS project-team], Patrick Amestoy [INPT-IRIT], Alfredo Buttari [CNRS-IRIT], Abdou Guermouche [HIEPACS project-team], Jean-Yves L'Excellent, François-Henry Rouet [Lawrence Berkeley Laboratory, CA, USA].

In this work, we study the memory scalability of the parallel multifrontal factorization of sparse matrices. In particular, we are interested in controlling the active memory specific to the multifrontal factorization. We illustrate why commonly used mapping strategies (e.g., the proportional mapping) cannot provide a high memory efficiency, which means that they tend to let the memory usage of the factorization grow when the number of processes increases. We propose "memory-aware" algorithms that aim at maximizing the granularity of parallelism while respecting memory constraints. These algorithms provide accurate memory estimates prior to the factorization and can significantly enhance the robustness of a multifrontal code. We illustrate our approach with experiments performed on large matrices.

This work has been published in the *SIAM Journal on Scientific Computing* [1].

7.8. Fast 3D frequency-domain full waveform inversion with a parallel Block Low-Rank multifrontal direct solver: application to OBC data from the North Sea

Participants: Patrick Amestoy [INPT-IRIT], Romain Brossier [ISTerre], Alfredo Buttari [CNRS-IRIT], Jean-Yves L'Excellent, Théo Mary [UPS-IRIT], Ludovic Métivier [CNRS-ISTerre-LJK], Alain Miniussi [Geoazur], Stéphane Operto [Geoazur].

Wide-azimuth long-offset OBC/OBN surveys provide a suitable framework to perform computationally-efficient frequency-domain full waveform inversion (FWI) with a few discrete frequencies. Frequency-domain seismic modeling is performed efficiently with moderate computational resources for a large number of sources with a sparse multifrontal direct solver (Gauss-elimination techniques for sparse matrices). Approximate solutions of the time-harmonic wave equation are computed using a Block Low-Rank (BLR) approximation, leading to a significant reduction in the operation count and in the volume of communication during the LU factorization as well as offering a great potential for reduction in the memory demand. Moreover, the sparsity of the seismic source vectors is exploited to speed up the forward elimination step during the computation of the monochromatic wavefields. The relevance and the computational efficiency of the frequency-domain FWI performed in the visco-acoustic VTI approximation is shown with a real 3D OBC case study from the North Sea. The FWI subsurface models show a dramatic resolution improvement relative to the initial model built by reflection traveltime tomography. The amplitude errors introduced in the modeled wavefields by the BLR approximation for different low-rank thresholds have a negligible footprint in the FWI results. With respect to a standard multifrontal sparse direct factorization, and without compromise on the accuracy of the imaging, the BLR approximation can bring a reduction of the LU factor size by a factor up to three. This reduction is not yet exploited to reduce the effective memory usage (ongoing work). The flop reduction can be larger than a factor of 10 and can bring a factor of time reduction of around three. Moreover, this reduction factor tends to increase with frequency, namely with the matrix size. Frequency-domain visco-acoustic VTI FWI can be viewed as an efficient tool to build an initial model for elastic FWI of 4-C OBC data.

This work has been published in the journal *Geophysics* [2].

7.9. Matching-Based Allocation Strategies for Improving Data Locality of Map Tasks in MapReduce

Participant: Loris Marchal.

MapReduce is a well-know framework for distributing data-processing computations on parallel clusters. In MapReduce, a large computation is broken into small tasks that run in parallel on multiple machines, and scales easily to very large clusters of inexpensive commodity computers. Before the Map phase, the original dataset is first split into chunks, that are replicated (a constant number of times, usually 3) and distributed onto the computing nodes. During the Map phase, nodes request tasks and are allocated first tasks associated to local chunks (if any). Communications take place when requesting nodes do not hold any local chunk anymore. In this work, we provide the first complete theoretical data locality analysis of the Map phase of MapReduce, and more generally, for bag-of-tasks applications that behaves like MapReduce. We show that if tasks are homogeneous (in term of processing time), once the chunks have been replicated randomly on resources with a replication factor larger than 2, it is possible to find a priority mechanism for tasks that achieves a quasi-perfect number of communications using a sophisticated matching algorithm. In the more realistic case of heterogeneous processing times, we prove using an actual trace of a MapReduce server that this priority mechanism enables to complete the Map phase with significantly fewer communications, even on realistic distributions of task durations.

This work is described in a technical report [41].

7.10. Minimizing Rental Cost for Multiple Recipe Applications in the Cloud

Participant: Loris Marchal.

Clouds are more and more becoming a credible alternative to parallel dedicated resources. The pay-per-use pricing policy however highlights the real cost of computing applications. This new criterion, the cost, must then be assessed when scheduling an application in addition to more traditional ones as the completion time or the execution flow. In this work, we tackle the problem of optimizing the cost of renting computing instances to execute an application on the cloud while maintaining a desired performance (throughput). The target application is a stream application based on a DAG pattern, i.e., composed of several tasks with dependencies, and instances of the same execution task graph are continuously executed on the instances. We provide some theoretical results on the problem of optimizing the renting cost for a given throughput then propose some heuristics to solve the more complex parts of the problem, and we compare them to optimal solutions found by linear programming.

This work has been published in *IPDPS Workshops* [27].

7.11. Malleable task-graph scheduling with a practical speed-up model

Participants: Loris Marchal, Bertrand Simon, Oliver Sinnen [Univ. Auckland, New Zealand], Frédéric Vivien.

Scientific workloads are often described by Directed Acyclic task Graphs. Indeed, DAGs represent both a model frequently studied in theoretical literature and the structure employed by dynamic runtime schedulers to handle HPC applications. A natural problem is then to compute a makespan-minimizing schedule of a given graph. In this work, we are motivated by task graphs arising from multifrontal factorizations of sparsematrices and therefore work under the following practical model. We focus on malleable tasks (i.e., a single task can be allotted a time-varying number of processors) and specifically on a simple yet realistic speedup model: each task can be perfectly parallelized, but only up to a limited number of processors. We first prove that the associated decision problem of minimizing the makespan is NP-Complete. Then, we study a widely used algorithm, PropScheduling, under this practical model and propose a new strategy GreedyFilling. Even though both strategies are 2-approximations, experiments on real and synthetic data sets show that GreedyFilling achieves significantly lower makespans.

This work is described in a technical report [52].

7.12. Dynamic memory-aware task-tree scheduling

Participant: Loris Marchal.

Factorizing sparse matrices using direct multifrontal methods generates directed tree-shaped task graphs, where edges represent data dependency between tasks. This work revisits the execution of tree-shaped task graphs using multiple processors that share a bounded memory. A task can only be executed if all its input and output data can fit into the memory. The key difficulty is to manage the order of the task executions so that we can achieve high parallelism while staying below the memory bound. In particular, because input data of unprocessed tasks must be kept in memory, a bad scheduling strategy might compromise the termination of the algorithm. In the single processor case, solutions that are guaranteed to be below a memory bound are known. The multi-processor case (when one tries to minimize the total completion time) has been shown to be NP-complete. We designed in this work a novel heuristic solution that has a low complexity and is guaranteed to complete the tree within a given memory bound. We compared our algorithm to state of the art strategies, and observed that on both actual execution trees and synthetic trees, we always performed better than these solutions, with average speedups between 1.25 and 1.45 on actual assembly trees. Moreover, we showed that the overhead of our algorithm is negligible even on deep trees (10^5), and would allow its runtime execution.

This work is described in a technical report [39].

7.13. Optimal resilience patterns to cope with fail-stop and silent errors

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

This work focuses on resilience techniques at extreme scale. Many papers deal with fail-stop errors. Many others deal with silent errors (or silent data corruptions). But very few papers deal with fail-stop and silent errors simultaneously. However, HPC applications will obviously have to cope with both error sources. This work presents a unified framework and optimal algorithmic solutions to this double challenge. Silent errors are handled via verification mechanisms (either partially or fully accurate) and in-memory checkpoints. Fail-stop errors are processed via disk checkpoints. All verification and checkpoint types are combined into computational patterns. We provide a unified model, and a full characterization of the optimal pattern. Our results nicely extend several published solutions and demonstrate how to make use of different techniques to solve the double threat of fail-stop and silent errors. Extensive simulations based on real data confirm the accuracy of the model, and show that patterns that combine all resilience mechanisms are required to provide acceptable overheads.

This work was presented at the *IPDPS'2016* conference [20].

7.14. Two-level checkpointing and partial verifications for linear task graphs

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

Fail-stop and silent errors are unavoidable on large-scale platforms. Efficient resilience techniques must accommodate both error sources. A traditional checkpointing and rollback recovery approach can be used, with added verifications to detect silent errors. A fail-stop error leads to the loss of the whole memory content, hence the obligation to checkpoint on a stable storage (e.g., an external disk). On the contrary, it is possible to use in-memory checkpoints for silent errors, which provide a much smaller checkpoint and recovery overhead. Furthermore, recent detectors offer partial verification mechanisms, which are less costly than guaranteed verifications but do not detect all silent errors. In this work, we show how to combine all these techniques for HPC applications whose dependence graph is a chain of tasks, and provide a sophisticated dynamic programming algorithm returning the optimal solution in polynomial time. Simulations demonstrate that the combined use of multi-level checkpointing and partial verifications further improves performance.

This work was presented at the *17th IEEE International Workshop on Parallel and Distributed Scientific and Engineering Computing (PDSEC 2016)* [21].

7.15. Resilient application co-scheduling with processor redistribution

Participants: Anne Benoit, Loïc Pottier, Yves Robert.

Recently, the benefits of co-scheduling several applications have been demonstrated in a fault-free context, both in terms of performance and energy savings. However, large-scale computer systems are confronted to frequent failures, and resilience techniques must be employed to ensure the completion of large applications. Indeed, failures may create severe imbalance between applications, and significantly degrade performance. In this work, we propose to redistribute the resources assigned to each application upon the striking of failures, in order to minimize the expected completion time of a set of co-scheduled applications. First we introduce a formal model and establish complexity results. When no redistribution is allowed, we can minimize the expected completion time in polynomial time, while the problem becomes NP-complete with redistributions, even in a fault-free context. Therefore, we design polynomial-time heuristics that perform redistributions and account for processor failures. A fault simulator is used to perform extensive simulations that demonstrate the usefulness of redistribution and the performance of the proposed heuristics.

This work was presented at the *ICCP'16* conference [22].

7.16. A different re-execution speed can help

Participants: Anne Benoit, Aurélien Cavelan, Valentin Le Fèvre, Yves Robert, Hongyang Sun.

We consider divisible load scientific applications executing on large-scale platforms subject to silent errors. While the goal is usually to complete the execution as fast as possible in expectation, another major concern is energy consumption. The use of dynamic voltage and frequency scaling (DVFS) can help save energy, but at the price of performance degradation. Consider the execution model where a set of K different speeds is given, and whenever a failure occurs, a different re-execution speed may be used. Can this help? We address the following bi-criteria problem: how to compute the optimal checkpointing period to minimize energy consumption while bounding the degradation in performance. We solve this bi-criteria problem by providing a closed-form solution for the checkpointing period, and demonstrate via a comprehensive set of simulations that a different re-execution speed can indeed help.

This work was presented at the *5th International Workshop on Power-aware Algorithms, Systems, and Architectures* [19].

7.17. Coping with recall and precision of soft error detectors

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

Many methods are available to detect silent errors in high-performance computing (HPC) applications. Each method comes with a cost, a recall (fraction of all errors that are actually detected, i.e., false negatives), and a precision (fraction of true errors amongst all detected errors, i.e., false positives). The main contribution of this work is to characterize the optimal computing pattern for an application: which detector(s) to use, how many detectors of each type to use, together with the length of the work segment that precedes each of them. We first prove that detectors with imperfect precisions offer limited usefulness. Then we focus on detectors with perfect precision, and we conduct a comprehensive complexity analysis of this optimization problem, showing NP-completeness and designing an FPTAS (Fully Polynomial-Time Approximation Scheme). On the practical side, we provide a greedy algorithm, whose performance is shown to be close to the optimal for a realistic set of evaluation scenarios. Extensive simulations illustrate the usefulness of detectors with false negatives, which are available at a lower cost than the guaranteed detectors.

This work was accepted for publication in the *Journal of Parallel and Distributed Computing* [7].

7.18. Checkpointing strategies for scheduling computational workflows

Participants: Anne Benoit, Yves Robert.

We study the scheduling of computational workflows on compute resources that experience exponentially distributed failures. When a failure occurs, rollback and recovery is used to resume the execution from the last checkpointed state. The scheduling problem is to minimize the expected execution time by deciding in which order to execute the tasks in the workflow and deciding for each task whether to checkpoint it or not after it completes. We give a polynomial-time optimal algorithm for fork DAGs (Directed Acyclic Graphs) and show that the problem is NP-complete with join DAGs. We also investigate the complexity of the simple case in which no task is checkpointed. Our main result is a polynomial-time algorithm to compute the expected execution time of a workflow, with a given task execution order and specified to-be-checkpointed tasks. Using this algorithm as a basis, we propose several heuristics for solving the scheduling problem. We evaluate these heuristics for representative workflow configurations.

This work was published in the *International Journal of Networking and Computing* [4].

7.19. Assessing General-Purpose Algorithms to Cope with Fail-Stop and Silent Errors

Participants: Anne Benoit, Aurélien Cavelan, Yves Robert, Hongyang Sun.

We combine the traditional checkpointing and rollback recovery strategies with verification mechanisms to cope with both fail-stop and silent errors. The objective is to minimize makespan and/or energy consumption. For divisible load applications, we use first-order approximations to find the optimal checkpointing period to minimize execution time, with an additional verification mechanism to detect silent errors before each checkpoint, hence extending the classical formula by Young and Daly for fail-stop errors only. We further extend the approach to include intermediate verifications, and to consider a bi-criteria problem involving both time and energy (linear combination of execution time and energy consumption). Then, we focus on application workflows whose dependence graph is a linear chain of tasks. Here, we determine the optimal checkpointing and verification locations, with or without intermediate verifications, for the bi-criteria problem. Rather than using a single speed during the whole execution, we further introduce a new execution scenario, which allows for changing the execution speed via dynamic voltage and frequency scaling (DVFS). In this latter scenario, we determine the optimal checkpointing and verification locations, as well as the optimal speed pairs for each task segment between any two consecutive checkpoints. Finally, we conduct an extensive set of simulations to support the theoretical study, and to assess the performance of each algorithm, showing that the best overall performance is achieved under the most flexible scenario using intermediate verifications and different speeds.

This work was accepted for publication in the journal *ACM Transactions on Parallel Computing* [8].

7.20. A failure detector for HPC platforms

Participant: Yves Robert.

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, able to 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 was presented at the *SC'16* conference [24].

7.21. Optimal multistage algorithm for adjoint computatio

Participant: Yves Robert.

We reexamine the work of Stumm and Walther on multistage algorithms for adjoint computation. We provide an optimal algorithm for this problem when there are two levels of checkpoints, in memory and on disk. Previously, optimal algorithms for adjoint computations were known only for a single level of checkpoints with no writing and reading costs; a well-known example is the binomial checkpointing algorithm of Griewank and Walther. Stumm and Walther extended that binomial checkpointing algorithm to the case of two levels of checkpoints, but they did not provide any optimality results. We bridge the gap by designing the first optimal algorithm in this context. We experimentally compare our optimal algorithm with that of Stumm and Walther to assess the difference in performance.

This work was accepted for publication in the *SIAM Journal on Scientific Computing* [5].

7.22. Assessing the cost of redistribution followed by a computational kernel: Complexity and performance results

Participant: Yves Robert.

The classical redistribution problem aims at optimally scheduling communications when reshuffling from an initial data distribution to a target data distribution. This target data distribution is usually chosen to optimize some objective for the algorithmic kernel under study (good computational balance or low communication volume or cost), and therefore to provide high efficiency for that kernel. However, the choice of a distribution minimizing the target objective is not unique. This leads to generalizing the redistribution problem as follows: find a re-mapping of data items onto processors such that the data redistribution cost is minimal, and the operation remains as efficient. This work studies the complexity of this generalized problem. We compute optimal solutions and evaluate, through simulations, their gain over classical redistribution. We also show the NP-hardness of the problem to find the optimal data partition and processor permutation (defined by new subsets) that minimize the cost of redistribution followed by a simple computational kernel. Finally, experimental validation of the new redistribution algorithms are conducted on a multicore cluster, for both a 1D-stencil kernel and a more compute-intensive dense linear algebra routine.

This work has been published in the Parallel Computing journal [14].

7.23. When Amdahl Meets Young/Daly

Participants: Aurélien Cavelan, Yves Robert.

This work investigates the optimal number of processors to execute a parallel job, whose speedup profile obeys Amdahl's law, on a large-scale platform subject to fail-stop and silent errors. We combine the traditional checkpointing and rollback recovery strategies with verification mechanisms to cope with both error sources. We provide an exact formula to express the execution overhead incurred by a periodic checkpointing pattern of length T and with P processors, and we give first-order approximations for the optimal values T* and P* as a function of the individual processor MTBF. A striking result is that P* is of the order of the fourth root of the individual MTBF if the checkpointing cost grows linearly with the number of processors, and of the order of its third root if the checkpointing cost stays bounded for any P. We conduct an extensive set of simulations to support the theoretical study. The results confirm the accuracy of first-order approximation under a wide range of parameter settings.

This work was presented at the *Cluster'16* conference [26].

7.24. Computing the expected makespan of task graphs in the presence of silent errors

Participants: Julien Herrmann, Yves Robert.

Applications structured as Directed Acyclic Graphs (DAGs) of tasks correspond to a general model of parallel computation that occurs in many domains, including popular scientific workflows. DAG scheduling has received an enormous amount of attention, and several list-scheduling heuristics have been proposed and shown to be effective in practice. Many of these heuristics make scheduling decisions based on path lengths in the DAG. At large scale, however, compute platforms and thus tasks are subject to various types of failures with no longer negligible probabilities of occurrence. Failures that have recently received increasing attention are silent errors, which cause a task to produce incorrect results even though it ran to completion. Tolerating silent errors is done by checking the validity of the results and re-executing the task from scratch in case of an invalid result. The execution time of a task then becomes a random variable, and so are path lengths. Unfortunately, computing the expected makespan of a DAG (and equivalently computing expected path lengths in a DAG) is a computationally difficult problem. Consequently, designing effective scheduling heuristics is preconditioned on computing accurate approximations of the expected makespan. In this work we propose an algorithm that computes a first order approximation of the expected makespan of a DAG when tasks are subject to silent errors. We compare our proposed approximation to previously proposed such approximations for three classes of application graphs from the field of numerical linear algebra. Our evaluations quantify approximation error with respect to a ground truth computed via a brute-force Monte Carlo method. We find that our proposed approximation outperforms previously proposed approaches, leading to large reductions in approximation error for low (and realistic) failure rates, while executing much faster.

This work was presented at the *Ninth Int. Workshop on Parallel Programming Models and Systems Software for High-End Computing (P2S2)* [25].

7.25. Toward an Optimal Online Checkpoint Solution under a Two-Level HPC Checkpoint Model

Participants: Yves Robert, Frédéric Vivien.

The traditional single-level checkpointing method suffers from significant overhead on large-scale platforms. Hence, multilevel checkpointing protocols have been studied extensively in recent years. The multilevel checkpoint approach allows different levels of checkpoints to be set (each with different checkpoint overheads and recovery abilities), in order to further improve the fault tolerance performance of extreme-scale HPC applications. How to optimize the checkpoint intervals for each level, however, is an extremely difficult problem. In this work, we construct an easy-to-use two-level checkpoint model. Checkpoint level 1 deals with errors with low checkpoint/recovery overheads such as transient memory errors, while checkpoint level 2 deals with hardware crashes such as node failures. Compared with previous optimization work, our new optimal checkpoint solution offers two improvements: (1) it is an online solution without requiring knowledge of the job length in advance, and (2) it shows that periodic patterns are optimal and determines the best pattern. We evaluate the proposed solution and compare it with the most up-to-date related approaches on an extreme-scale simulation testbed constructed based on a real HPC application execution. Simulation results show that our proposed solution outperforms other optimized solutions and can improve the performance significantly in some cases. Specifically, with the new solution the wall-clock time can be reduced by up to 25.3% over that of other state-of-the-art approaches. Finally, a brute-force comparison with all possible patterns shows that our solution is always within 1% of the best pattern in the experiments.

This work has been published in IEEE Transactions on Parallel and Distributed Systems [11].

7.26. Cell morphing: from array programs to array-free Horn clauses

Participants: Laure Gonnord, David Monniaux [(CNRS/Verimag)], Julien Braine [(M2 Student)].

Automatically verifying safety properties of programs is hard. Many approaches exist for verifying programs operating on Boolean and integer values (e.g. abstract interpretation, counterexample-guided abstraction refinement using interpolants), but transposing them to array properties has been fraught with difficulties. Our work addresses that issue with a powerful and flexible abstraction that morphes concrete array cells into a finite set of abstract ones. This abstraction is parametric both in precision and in the back-end analysis used. From

our programs with arrays, we generate nonlinear Horn clauses over scalar variables only, in a common format with clear and unambiguous logical semantics, for which there exist several solvers. We thus avoid the use of solvers operating over arrays, which are still very immature. Experiments with our prototype VAPHOR show that this approach can prove automatically and without user annotations the functional correctness of several classical examples, including *selection sort*, *bubble sort*, *insertion sort*, as well as examples from literature on array analysis.

This work has been published in Static Analysis Symposium [30] for the array part. We are currently deseigning an extension to programs with inductive data structures.

7.27. Symbolic Analyses of pointers

Participants: Laure Gonnord, Maroua Maalej, Fernando Pereira [(UFMG, Brasil)], Leonardo Barbosa [(UFMG, Brasil)], Vitor Paisante [(UFMG, Brasil)], Pedro Ramos [(UFMG, Brasil)].

Alias analysis is one of the most fundamental techniques that compilers use to optimize languages with pointers. However, in spite of all the attention that this topic has received, the current state-of-the-art approaches inside compilers still face challenges regarding precision and speed. In particular, pointer arithmetic, a key feature in C and C++, is yet to be handled satisfactorily.

A first work presents a new range-based alias analysis algorithm to solve this problem. The key insight of our approach is to combine alias analysis with symbolic range analysis. This combination lets us disambiguate fields within arrays and structs, effectively achieving more precision than traditional algorithms. To validate our technique, we have implemented it on top of the LLVM compiler. Tests on a vast suite of benchmarks show that we can disambiguate several kinds of C idioms that current state-of-the-art analyses cannot deal with. In particular, we can disambiguate 1.35x more queries than the alias analysis currently available in LLVM. Furthermore, our analysis is very fast: we can go over one million assembly instructions in 10 seconds.

A second work starts 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.

This work has been published in the *International Symposium of Code Generation and Optmization* [31] and at CGO'17 [29].

7.28. High-Level Synthesis of Pipelined FSM from Loop Nests

Participants: Christophe Alias, Fabrice Rastello [(Inria/CORSE)], Alexandru Plesco [(XtremLogic SAS, France)].

Embedded systems raise many challenges in power, space and speed efficiency. The current trend is to build heterogeneous systems on a chip with specialized processors and hardware accelerators. Generating an hardware accelerator from a computational kernel requires a deep reorganization of the code and the data. Typically, parallelism and memory bandwidth are met thanks to fine-grain loop transformations. Unfortunately, the resulting control automaton is often very complex and eventually bound the circuit frequency, which limits the benefits of the optimization. This is a major lock, which strongly limits the power of the code optimizations appliable by high-level synthesis tools.

In this work, we propose an architecture of control automaton and an algorithm of high-level synthesis which translates efficiently the control required by fine-grain loop optimizations. Unlike the previous approaches, our control automaton can be pipelined *at will, without any restriction*. Hence, the frequency of the automaton can be as high as possible. Experimental results on FPGA confirms that our control circuit can reach a high frequency with a reasonable resource consumption.

This work is described in a technical report [36].

7.29. Estimation of Parallel Complexity with Rewriting Techniques

Participants: Christophe Alias, Laure Gonnord, Carsten Fuhs [(Birbeck, UK)].

We show how monotone interpretations - a termination analysis technique for term rewriting systems - can be used to assess the inherent parallelism of recursive programs manipulating inductive data structures. As a side effect, we show how monotone interpretations specify a parallel execution order, and how our approach extends naturally affine scheduling - a powerful analysis used in parallelising compilers - to recursive programs. This preliminary work opens new perspectives in automatic parallelisation.

This work has been published in the *Workshop on Termination,* [15].

STORM Team

6. New Results

6.1. Automatic OpenCL Task Adaptation for Heterogeneous Architectures

OpenCL defines a common parallel programming language for all devices, although writing tasks adapted to the devices, managing communication and load-balancing issues are left to the programmer. In this work [11], we propose a novel automatic compiler and runtime technique to execute single OpenCL kernels on heterogeneous multi-device architectures. Our technique splits computation and data automatically across the computing devices. The technique proposed is completely transparent to the user, does not require off-line training or a performance model. It handles communications and load-balancing issues, resulting from hardware heterogeneity, load imbalance within the kernel itself and load variations between repeated executions of the kernel, in an iterative computation. We present our results on benchmarks and on an N-body application over two platforms, a 12-core CPU with two different GPUs and a 16-core CPU with three homogeneous GPUs.

6.2. Fast Forward Error Correction Codes

Erroc Correction Codes are essential for preserving data integrity in communications. These algorithms find errors due to noise in transmissions and correct these errors with a high probability. Several algorithms are used, with different capacities in term of correction and most of them are implemented in cell phones or satellites as ASICS. The need to handle many different usages, different contexts of use pushes towards software solutions. A larger spectrum of algorithms can be explored, in order to meet the expectations in terms of performance, power consumption and error correcting power. These new algorithms, for the 5G for instance, can then be either implemented in software (for large antenna for instance) or in hardware. In both case, software simulation is necessary in order to evaluate the properties of the new algorithms. We developped in collaboration with IMS new versions of algorithms and a new software, AFF3CT http://aff3ct.github.io/index.html, that allows the exploration of many different algorithmic variants and their evaluation. Two conference papers have been published on these new results [7][6].

6.3. Resource aggregation for task-based Cholesky Factorization

Hybrid computing platforms are now commonplace, featuring a large number of CPU cores and accelerators. This trend makes balancing computations between these heterogeneous resources performance critical. In a recent paper [8] we propose aggregating several CPU cores in order to execute larger parallel tasks and thus improve the load balance between CPUs and accelerators. Additionally, we present our approach to exploit internal parallelism within tasks. This is done by combining two runtime systems: one runtime system to handle the task graph and another one to manage the internal parallelism. We demonstrate the relevance of our approach in the context of the dense Cholesky factorization kernel implemented on top of the StarPU task-based runtime system.We present experimental results showing that our solution outperforms state of the art implementations. In addition, we realized an extended version of this paper submitted for review to the Parallel Computing journal special issue for HCW and HeteroPar 2016 workshops. In this new paper [19] we provide additional details on our contribution and propose a brand new study on the recent Intel Xeon Phi Knights Landing (KNL) where we show that we are able to outperform existing state of the art implementations on this platform thanks to our proposed technique.

6.4. Scheduling of Linear Algebra Kernels on Multiple Heterogeneous Resources

In this paper [5], we consider task-based dense linear algebra applications on a single heterogeneous node which contains regular CPU cores and a set of GPU devices. Efficient scheduling strategies are crucial in this context in order to achieve good and portable performance. HeteroPrio, a resource-centric dynamic scheduling strategy has been introduced in a previous work and evaluated for the special case of nodes with exactly two types of resources. However, this restriction can be limiting, for example on nodes with several types of accelerators, but not only this. Indeed, an interesting approach to increase resource usage is to group several CPU cores together, which allows to use intra-task parallelism. We propose a generalization of HeteroPrio to the case with several classes of heterogeneous workers. We provide extensive evaluation of this algorithm with Cholesky factorization, both through simulation and actual execution, compared with HEFT-based scheduling strategy, the state of the art dynamic scheduling strategy for heterogeneous systems. Experimental evaluation shows that our approach is efficient even for highly heterogeneous configurations and significantly outperforms HEFT-based strategy.

6.5. Analyzing Dynamic Task-Based Applications on Hybrid Platforms: An Agile Scripting Approach

In this paper [10], we present visual analysis techniques to evaluate the performance of HPC task-based applications on hybrid architectures. Our approach is based on composing modern data analysis tools (pjdump, R, ggplot2, plotly), enabling an agile and flexible scripting framework with minor development cost. We validate our proposal by analyzing traces from the full-fledged implementation of the Cholesky decomposition available in the MORSE library running on a hybrid (CPU/GPU) platform. The analysis compares two different workloads and three different task schedulers from the StarPU runtime system. Our analysis based on composite views allows to identify allocation mistakes, priority problems in scheduling decisions, GPU tasks anomalies causing bad performance, and critical path issues.

6.6. Distributed StarPU Scalability on Heterogeneous Platforms

The emergence of accelerators as standard computing resources on supercomput- ers 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 moti- vated its support in the OpenMP 4.0 standard. In this paper, we show 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 [16]. We illustrate our discussion with the task- based tile Cholesky algorithm that we implemented on top of this new runtime system layer. We show that it allows for 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.

6.7. Controlling the Memory Subscription of Distributed Applications with a Task-Based Runtime System

The ever-increasing supercomputer architectural complexity emphasizes the need for high-level parallel programming paradigms. Among such paradigms, task-based programming manages to abstract away much of the archi- tecture complexity while efficiently meeting the performance challenge, even at large scale. Dynamic run-time systems are typically used to execute task-based applications, to schedule computation resource usage and memory allocations. While computation scheduling has been well studied, the dynamic management of memory resource subscription inside such run- times has however been little explored. This paper [12] studies

the cooperation between a task-based distributed application code and a run-time system engine to control the memory subscription levels throughout the execution. We show that the task paradigm allows to control the memory footprint of the application by throttling the task submission flow rate, striking a compromise between the performance benefits of anticipative task submission and the resulting memory consumption. We illustrate the benefits of our contribution on a compressed dense linear algebra distributed application.

6.8. StarPU Interfacing with GASPI/GPI2

A version of the distributed dependence support of StarPU has been ported by Corentin Salingue, under the supervision of Olivier Aumage on the high performance GASPI/GPI2 networking layer developed by the Fraunhofer institute in Germany. The GPI2 framework offers a lightweight communication interface specifically designed for thread enabled HPC applications. This work has been conducted as part of the H2020 INTERTWinE european project.

6.9. A Stencil DSEL for Single Code Accelerated Computing with SYCL

Stencil kernels arise in many scientific codes as the result from dis- cretizing natural, continuous phenomenons. Many research works have designed stencil frameworks to help programmer optimize stencil kernels for performance, and to target CPUs or accelerators. However, existing stencil kernels, either library-based or language-based necessitate to write distinct source codes for accelerated ker- nels and for the core application, or to resort to specific keywords, pragmas or language extensions. SYCL is a C++ based approach designed by the Khronos Group to program the core application as well as the application kernels with a single unified, C++ compliant source code. A SYCL application can then be linked with a CPU-only runtime library or processed by a SYCL-enabled compiler to automatically build an OpenCL accelerated application. Our contribution [13] is a stencil domain specific embedded language (DSEL) which leverage SYCL together with expression template techniques to implement statically optimized stencil applications able to run on platforms equipped with OpenCL devices, while preserving the single source benefits from SYCL.

6.10. Bridging the gap between OpenMP 4.0 and native runtime systems for the fast multipole method

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. The successful ports of fully-featured numerical libraries on several recent runtime system proposals have shown, indeed, the benefit of task-based parallelism models in terms of performance portability on complex platforms. 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 a 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 as part of its revision 4.0. The goal of this paper [15] is to assess the effectiveness and limits of this support for designing a high-performance numerical library. We illustrate our discussion with the ScalFMM library, which implements state-of-the-art fast multipole method (FMM) algorithms, that we have deeply re-designed 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. We furthermore propose extensions to the OpenMP 4 standard and show how they can enhance FMM performance. To assess our statement, we have implemented this support within the Klang-OMP source-to-source compiler that translates OpenMP directives into calls to the StarPU task-based runtime system. This study shows that we can take advantage of the advanced capabilities of a fully-featured runtime system without resorting to a specific, native runtime port, hence bridging the gap between the OpenMP standard and the very high performance that was so far reserved to expert-only runtime system APIs.

6.11. Hierarchical Tasks

Modern computing platforms are heterogeneous and the load balancing is more complex to reach high performance. We decided to deal with the granularity problem in the context of task paralleliism and in a dynamic way through the implementation of hierarchical tasks in StarPU runtime. The idea is to give the runtime the ability to control tasks submission in order to choose the good granularity at the right moment. The application describes a control graph and the runtime generates the computation tasks graph on-the-fly according to the state of the machine (available computing resources, memory consumption, ...). As a consequence the runtime is able to limit the size of the computation tasks graph without loosing parallelism. Some experiments have been done on a Cholesky application and in the qr-mumps software and show that the work of an application programmer can be alleviated and the granularity choice could be easily delegated to the task based runtime.

6.12. Software-Hardware Exploration for Read-Only Data

We have proposed a new way of managing the cache by exploiting the difference of behavior in the memory system between read-only data and read-write data. A division of the existing cache-based memory hierarchy is proposed in order to create a dedicated data path for read-only data. This proposition is similar to the existing separation at the L1-level between instruction and data caches. In order to justify this approach, an analysis performed on a set of benchmarks shows that read-only data count for significant part of the working set and are less reused than read-write data. A transparent solution is proposed based on specific compilation support to separate automatically the memory accesses of read-only data at L1-level. This organization exploits the properties of the different sub- workloads in order to increase the overall data locality and data reuse. Simulated in a multicore environment, the evaluation of the new memory organization shows reduction of L1 misses up to 28.5%. Moreover, the messages issued on the interconnection network can be reduced up to 14.7% without any penalty on the performance.

Besides the reduced miss-rate allows maintaining performance with smaller cache size on the read-write path while the properties of the read- only part can benefit of a simplified cache implementation despite a shared multicore access [1].

TADAAM Team

7. New Results

7.1. Network Modeling

NETLOC (see Section 6.1) is a tool in HWLOC to discover the network topology. Our first work with NETLOC was to redesign it to be more efficient and more adapted to the needs. The code was cleaned and some dependencies were removed. We have added a display tool, that is able to show a network topology in a web browser where a user can interact with. It ran on one of the largest European supercomputer (the TGCC/Genci CURIE machine) and successfully modeled its 5200 nodes and its interconnection network (more than 800 switches).

Moreover, it is now possible to interact with Scotch from netloc. The first feature is to export a network topology, or even the current available topology given by the resource manager, into a SCOTCH architecture. Conversely, we can use SCOTCH tools in NETLOC for building a process mapping based on resources found by NETLOC and a process graph describing communications between processes. Tests conducted on a stencil mini-app have shown that the benefits are real and still needs more work.

7.2. Communication and computation overlap

To amortize the cost of communication in HPC application, programmers want to overlap communications with computation. To do so, they assume non-blocking MPI communications will progress in background. NewMadeleine, our communication library, is actually able to make communication progress in background so as to actually have overlap happen. However, not all MPI implementations are able to overlap communication and computation.

We have proposed [8] a benchmark to measure what really happens when trying to overlap non-blocking point-to-point communications with computation. The benchmark measures how much overlap happen in various cases: sender-side, receiver-side, datatypes likely to be offloaded onto NIC or not, multi-threaded computation, multi-threaded communication or not. We have benchmarked a wide panel of MPI libraries and hardware platforms, and thanks to low-level traces, explained the results.

7.3. Topology Aware Performance Monitoring

A tool has been developed to abstract performance metrics and map them onto the HWLOC (see Section 6.6) topology model of the system. During the year 2016, the tool has been entirely rewritten to release a more meaningful and stable programming abstraction, with off the shelf performance abstraction plugins and raw performance acquisition plugin [16]. A special effort has been carried out on output presentation by extending lstopo tool from hwloc into a library embedded in the monitoring tool to display performance metrics on the system topology. Another backend using R has also been developed for the purpose of post-mortem analysis and model extraction from abstract metrics of the topology.

7.4. Locality Aware Roofline Model

The years 2016 marked the achievement of our extension of the famous Cache Aware Roofline Model(CARM) and the associate tool. The latter model targets deep plateform and application analysis on multicore processors. Its model consist into a two-dimensions plane bound by several machine ceils and representative of scientific application workloads. Our extension validate the use of the CARM on emerging processors with heterogeneous memory subsystem, and extend the CARM methodology to encompass interconnection network, thus, enabling full modeling of shared memory systems [17]. This work is a collaboration with the INESC-ID research center under the NESUS project.

7.5. Performance Analysis of Electromagnetic Field Application on Large SMP Node

In the scope of the COLOC project we worked on understanding scalability issues of the efield application on a large shared memory system. Our analysis with above mentionned tools highlighted a potential bandwidth bottleneck. This problem can usually be tackled by the mean of threads and data mapping on respectively the machine cores and the memories. Unfortunately, those techniques can't be applied with this (closed source) application since the system does not allow to monitor memory accesses and traffic on the system.

7.6. Structural Modeling of Heterogeneous Memory Architectures

HWLOC (see Section 6.6) is the de facto standard tool for gathering information of parallel platform topologies. The advent of new memory architecture, with high-bandwidth and/or non-volatile memories cause the memory management subsytem complexity to increase. Indeed, besides taking care of allocating data buffers locally, developers also have to choose between different local memories with different performance and persistence characteristics. Moreover, the operating systems still cannot expose the full details about these technologies to applications. We modified the HWLOC tool to cope with these new needs in collaboration with Intel. This work led to the design a new structural model for platforms with heteregeneous memories [10].

7.7. Scalable Management of Platform Topologies

HWLOC (see Section 6.6) is used for gathering the topology of thousands of nodes in large clusters. Those nodes are now growing to hundreds of cores, making the overall amount of topology information non-negligible. We designed new ways to compress topologies, either lossless or lossy, for easier transfer between compute nodes and front nodes and more compact storage and manipulation [20]. We also 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 [11].

7.8. MPI One-side operations

MPI one-sided operations, aka Remote Memory Access (RMA), are direct read/write memory access to a remote node. Only one node (the origin) explicitely calls MPI operations, while communication progression is implicit for the other node (the target). These operations assume that the communication library is able to make communication progress in background.

Since MadMPI, the MPI implementation of NewMadeleine (see Section 6.2), extensively uses event-driven mechanism to reach asynchronous progression, we have [24] taken advantage of this property to implement MPI RMA operations in the library. This implementation keeps the overlap properties by asynchronously handle the messages exchanged by the applications. The addition also supports MPI_THREAD_MULTIPLE, for both shared and distributed memory contexts.

7.9. Topology and affinity aware hierarchical and distributed load-balancing

The evolution of massively parallel supercomputers make palpable two issues in particular: the load imbalance and the poor management of data locality in applications. Thus, with the increase of the number of cores and the drastic decrease of amount of memory per core, the large performance needs imply to particularly take care of the load-balancing and as much as possible of the locality of data. One mean to take into account this locality issue relies on the placement of the processing entities and load balancing techniques are relevant in order to improve application performance. With large-scale platforms in mind, we developed a hierarchical and distributed algorithm which aim is to perform a topology-aware load balancing tailored for Charm++ applications. This algorithm is based on both LibTopoMap for the network awareness aspects and on Treematch to determine a relevant placement of the processing entities. We show that the proposed algorithm improves the overall execution time in both the cases of real applications and a synthetic benchmark as well. For this last experiment, we show a scalability up to one millions processing entities [12].

7.10. Topology-Aware Data Aggregation for Intensive I/O on Large-Scale Supercomputers

Reading and writing data efficiently from storage systems is critical for high performance data-centric applications. These I/O systems are being increasingly characterized by complex topologies and deeper memory hierarchies. Effective parallel I/O solutions are needed to scale applications on current and future supercomputers. Data aggregation is an efficient approach consisting of electing some processes in charge of aggregating data from a set of neighbors and writing the aggregated data into storage. Thus, the bandwidth use can be optimized while the contention is reduced. In [13], we have taken into account the network topology for mapping aggregators and we propose an optimized buffering system in order to reduce the aggregation cost. We have validated our approach using micro-benchmarks and the I/O kernel of a large-scale cosmology simulation. We have showed improvements up to 15× faster for I/O operations compared to a standard implementation of MPI I/O.

7.11. Communication monitoring in OpenMPI

Monitoring data exchanges is critical when it comes to optimize process placement in a large scale environment. We participated in adding in Open-MPI, which is one of the major MPI implementation, a fine grain, point-to-point monitoring component that keeps track of message exchanges. Unlike implementations using PMPI operations, the layer in which this monitoring acts allow us to record at a lower level the effective data communications, for example, after the covering tree has been calculated. This component has been enriched with a complete coverage of collectives, point-to-point and one-sided communications. This component also reports informations about message sizes distribution. Monitored informations can be accessed by using MPI_Tools interface, or by dumping data in files.

7.12. Process Placement with TreeMatch

We released TreeMatch ver 0.4 in August. The new feature are: a new API, the handling oversubscribing (being able to map more processes that computing resources), fast exhaustive search (for small cases), K-partitioning in case of large arity of the tree, and a set of extensive tests.

7.13. Topology Aware Resource Management

SLURM is a Resource and Job Management System, a middleware in charge of delivering computing power to applications in HPC systems. Our goal is to take in account in SLURM placement process hardware topology but application communication pattern too. We have a new [9], [19] selection option for the cons_res plugin in SLURM. In this case the usually best_fit algorithm used to choose nodes is replaced by TreeMatch, an algorithm to find the best placement among the free nodes list in light of a given application communication matrix. We plan to release this work in the next release SLURM 17.02.

Fragmentation in cluster is one of the criteria important for administrator. Indeed, the way jobs are allocated impacts the global resource usage. Usually it is observed throught utilization of a cluster for a fixed load rate, but no metrics dedicated to fragmentation exist in litterature. Hence we construct several metrics to measure it. Our goal is to study the impact of our selection algorithm on fragmentation in comparison with other.

7.14. Impact of progress threads placement for MPI Non-Blocking Collectives

MPI Non-Blocking Collectives (NBC) allow communication overlap with computation. A good overlapping ratio is obtained when computation and communication are running in parallel. To achieve this, some implementations use progress threads to manage communication tasks. These threads should be bound on different cores to maximize the overlap. Thus, we elaborate several threads placement algorithms. These algorithms have been implemented within the MPC framework, using the HWLOC software to get a global view of the machine topology. We propose [18] a thread placement algorithm taking into account the NUMA topology of the machine in order to improve the overlapping ratio of non-blocking collective communications.

7.15. Hierarchical Communication Management in MPI

MPI, in its current state provides only a very limited set of functionnalities so as to allow the programmer to effectively leverage the physical characteristics of the underlying hardaware, such as the potentially complex memory hierarchy. The MPI philosophy being to be a hardware-agnostic interface, the challenge is therefore to propose an interface extension that offers the programmer significant control over the hardawre without dwelving too much into hardware details. We seek the right level of abstraction for this interface and the goal is push this proposal to the MPI Forum. This new interface is based on the concept of communicators, expands an already existing function available in the standard and also introduces a couple of helper functions. We have prototyped and drafted our proposal for the 2017 meetings of the forum.

7.16. Fully-abstracted approach for efficient thread binding in task-based model of programming

Task-based models and runtimes are quite popular in the HPC community. They help to implement applications with a high level of abstraction while still applying different types of optimizations. An important optimization target is hardware affinity, which concerns to match application behavior (thread, communication, data) to the architecture topology (cores, caches, memory). In fact, realizing a well adapted placement of threads is a key to achieve performance and scalability, especially on NUMA-SMP machines. However, this type of optimization is difficult: architectures become increasingly complex and application behavior changes with implementations and input parameters, *e.g* problem size and number of thread. Thus, by themselves task based runtimes often deal badly with this optimization and leave a lot of fine-tuning to the user. In this work [21], [25], [14], we propose a fully automatic, abstracted and portable affinity module. It produces and implements an optimized affinity strategy that combines knowledge about application characteristics and the architecture's topology. Implemented in the backend of our task-based runtime ORWL, our approach was used to enhance the performance and the scalability of several unmodified ORWL-coded applications: matrix multiplication, a 2D stencil (Livermore Kernel 23), and a video tracking real world application. On two SGI SMP machines with quite different hardware characteristics, our tests show spectacular performance improvements for this unmodified application code due to a dramatic decrease of cache misses. A comparison to reference implementations using OpenMP confirms this performance gain of almost one order of magnitude.

7.17. Multi-criteria graph partitioning for multi-physics simulations load balancing

A new set of algorithms has been designed to compute multi-criteria static mappings for the load balancing of multi-phisics simulations. The multi-criteria graph partitioning is known to be NP-hard, and there exist very few multi-criteria graph partitioners. Moreover, they focus on the edge-cut minimization instead of enforcing load balance. In practice, this strategy often leads to very unbalanced partitions, which are not useful for multi-physics simulations.

We have designed algorithms that focus on balancing several criteria at the same time to ensure that our results always match all balance criteria. We have implemented a prototype in Python to test these different heuristics. One of them, called PIERE, obtained good results [15], in term of balance as well as communication costs. PIERE uses the classic multilevel framework, but implements a new initial partitioning algorithm, which allows to find a balanced partition of the graph. The partition is then refined by local optimization heuristics that ensure the balance is kept for all criteria. This allow us to return a partition respecting the balance constraints. In [15], we compare against well-known partitioners that are SCOTCH and METIS, and highlight that, for a small mesh, the results exhibit a high discrepancy: each tool lacks of robustness.

PIERE outperformed the existing software METIS in our test cases, but there is room for improvement. We also verified the superiority of the hypergraph model over the graph model used by most partitioners. Meanwhile, we studied the source code of well known partitioners, namely METIS and SCOTCH, and we have identified a lot of algorithmic choices and internal parameters that are not described in their documentations. Carefully analyzing them helps us to clearly understand the differences of the different algorithms.

7.18. Scotch

In order to prepare for the inclusion of multi-criteria graph partitioning algorithms in SCOTCH, in the context of the PhD thesis of Rémi Barat, a new branch has been created in the SCOTCH repository. This new branch, labeled as 6.1, is the basis for the next main release of SCOTCH. The sequential graph structure has been adapted to handle graphs with multiple loads per vertex, and all the related algorithms have been adapted to take into account multiple vertex loads. This resulted in minimal updates in the interface of Scotch, with ful ascending compatibility. All of these modifications have been performed so as not to slow down significantly the algorithms in the most common case of graphs with single vertex loads.

7.19. PaMPA

Parallel remeshing has been improved. PaMPA coupled with Mmg (v5) remeshed a tetrahedral mesh from 43Melements to more than 1Belements on 280 Broadwell processors in 20 minutes. The resulted mesh, used by CERFACS, permitted one of the most finest simulation computed with LES (Large Eddy Simulation) on combustion.

The scalability of PT-SCOTCH scalability has been tested on the Curie cluster and compared to that of PARMETIS. These tests used DARI resources.

7.20. Originality of software works

Most judges have very little, if not none, knowledge on software developement. This results in misconceptions and mistakes regarding the application of copyright/author right (*droit d'auteur*) in court cases related to software. More generally, the concept of originality is misunderstood. While this criterion is meant in theory to separate works of the mind that are personal to an author (e.g., literary works), from creations of form that cannot, by nature, reflect the personality of their creator (e.g. mathematical tables), it is often used to qualify the degree of similarity between two different works, in the context of plagiarism. Also, the distinction between the realm of programs, that is, works of the mind, and that of algorithms, is not mastered. Algorithms belong to the *fonds commun*, a French term that has no equivalent in English and might be translated as "common pool". In order to help judges and lawmakers in understanding these notions, and articulate them, we have proposed a methodology for ruling software disputes. This methodology is solely based on the study of similarities in software code, since author right exclusively pertains to the level of the form [22].

ASCOLA Project-Team

7. New Results

7.1. Software composition and programming languages

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

This year we have published a number of new results in the domains of software composition and programming languages that range from pragmatic ones like modularity issues to formal studies in the domain of dependent type theory via static analysis and formal verification.

7.1.1. Formal Methods, logics and type theory

Concerning verification and formal semantics, we have defined the semantics of our dependent interoperability framework and we propose the notion the partial type equivalences as a key feature. We have also studied proofs in dependent type theory and synthesized call-by-value and call-by-name translations.

7.1.1.1. Verified Dependent Interoperability.

Full-spectrum dependent types promise to enable the development of correct-by-construction software. However, even certified software needs to interact with simply-typed or untyped programs, be it to perform system calls, or to use legacy libraries. Trading static guarantees for runtime checks, the dependent interoperability framework provides a mechanism by which simply-typed values can safely be coerced to dependent types and, conversely, dependently-typed programs can defensively be exported to a simply-typed application. In [22], we give a semantic account of dependent interoperability. Our presentation relies on and is guided by a pervading notion of type equivalence, whose importance has been emphasized in recent works on homotopy type theory. Specifically, we develop the notion of partial type equivalences as a key foundation for dependent interoperability. Our framework is developed in Coq; it is thus constructive and verified in the strictest sense of the terms. Using our library, users can specify domain-specific partial equivalences between data structures. Our library then takes care of the (sometimes, heavy) lifting that leads to interoperable programs. It thus becomes possible, as we shall illustrate, to internalize and hand-tune the extraction of dependently-typed programs to interoperable OCaml programs within Coq itself.

7.1.1.2. Forcing in Type Theory.

In [26], we study forcing translations of proofs in dependent type theory, through the Curry-Howard correspondence. Based on a call-by-push-value decomposition, we synthesize two simply-typed translations: i) one call-by-value, corresponding to the translation derived from the presheaf construction as studied in a previous paper; ii) one call-by-name, whose intuitions already appear in Krivine and Miquel's work. Focusing on the call-by-name translation, we adapt it to the dependent case and prove that it is compatible with the definitional equality of our system, thus avoiding coherence problems. This allows us to use any category as forcing conditions, which is out of reach with the call-by-value translation. Our construction also exploits the notion of storage operators in order to interpret dependent elimination for inductive types. This is a novel example of a dependent theory with side-effects, clarifying how dependent elimination for inductive types must be restricted in a non-pure setting. Being implemented as a Coq plugin, this work gives the possibility to formalize easily consistency results, for instance the consistency of the negation of Voevodsky's univalence axiom.

7.1.2. Programming languages

In the domain of programming languages we have presented new results on constraint programming, development of correct programs by construction and better controls for computational effects and modularity for JavaScript.

7.1.2.1. Constraint programming

Constraint programming (CP) relies on filtering algorithms in order to deal with combinatorial problems. Global constraints offer efficient algorithms for complex constraints. In particular a large family of global constraints can be expressed as constraints of finite state automata with counters. We have generalized these automata constraints in order to compose them as transducers [16]. We have also extended these results with different techniques [20]. First, we have improved the automaton synthesis to generate automata with fewer accumulators. Second, we have shown how to decompose a constraint specified by an automaton with accumulators into a conjunction of linear inequalities, for use by a MIP (Mixed-Integer Programming) solver. Third, we have generalized the implied constraint generation to cover the entire family of time-series constraints. The newly synthesized automata for time-series constraints outperform the old ones, for both the CP and MIP decompositions, and the generated implied constraints boost the inference, again for both the CP and MIP decompositions.

7.1.2.2. Program correctness

Most IDEs provide refactoring tools to assist programmers when they modify the structure of their software. However the refactoring facilities of many popular tools (Eclipse, Visual Studio, IntelliJ, etc.) are currently not reliable : they occasionally change the program semantics in unexpected ways, and, as a result, the programmers systematically have to re-test the resulting code. We have build a refactoring tool for C programs which core operation is proved correct by construction [21]. To do that, we build an AST transformation with Coq (based on the CompCert C implementation) and we prove that this transformation preserves the external behavior of programs. The code of the transformation is then extracted to OCaml and is then embedded in a traditional parse/transform/pretty-print setting to provide a working prototype.

7.1.2.3. Effect Capabilities

Computational effects complicate the tasks of reasoning about and maintaining software, due to the many kinds of interferences that can occur. While different proposals have been formulated to alleviate the fragility and burden of dealing with specific effects, such as state or exceptions, there is no prevalent robust mechanism that addresses the general interference issue. Building upon the idea of capability-based security, we propose in [18] effect capabilities as an effective and flexible manner to control monadic effects and their interferences. Capabilities can be selectively shared between modules to establish secure effect-centric coordination. We further refine capabilities with type-based permission lattices to allow fine-grained decomposition of authority. We provide an implementation of effect capabilities in Haskell, using type classes to establish a way to statically share capabilities between modules, as well as to check proper access permissions to effects at compile time. We first exemplify how to tame effect interferences using effect capabilities by treating state and exceptions. Then we focus on taming I/O by proposing a fine-grained lattice of I/O permissions based on the current classification of its operations. Finally, we show that integrating effect capabilities with modern tag-based monadic mechanisms provides a practical, modular and safe mechanism for monadic programming in Haskell.

7.1.2.4. Extensible JavaScript Modules

As part of the SecCloud project, we have studied how to modularly extend JavaScript interpreters with dynamic security analyses in particular information flow analyses. This has led us to study ways to improve on the standard JavaScript module pattern. This pattern is commonly used to encapsulate definitions by using closures. However, closures prevent module definitions from being extended at runtime. We have proposed a simple pattern that not only opens the module, but allows one to extend the module definitions in layers [39]. The pattern leverages the with construct and the prototype delegation mechanism of JavaScript to mimic a form of dynamic binding, while minimizing the changes made to the module code.

Florent Marchand's PhD thesis [13] details the proposal further and shows its application to the modular extension of Narcissus, a full-blown JavaScript interpreter, with several dynamic analyses, including the information flow of Austin and Flanagan based on multiple facets. A comparison with a previous ad hoc implementation of the analysis illustrates the benefits of the proposal.

7.1.3. Software Security and Privacy

In the area of security we have focused on expressing advanced security concerns with abstract and formal languages and the study of policy monitoring and the detection of conflicts.

7.1.3.1. Runtime verification of advanced logical security properties.

Monitoring or runtime verification means to observe the system execution and to check if it deviates or not from a predefined contract. Our contract is a formula written in AAL (Abstract Accountability Language) expressing the expected behavior of a system, the audit steps as well as punishment and compensation. We choose to use the rewriting approach with the three valued logic as many other existing approaches. The monitoring problem raised a validity question, if we start with a formula neither true nor false are we sure to conclude? The response is no and this is a completeness problem and all published solutions are incomplete. For LTL, mixing the standard semantics, the rewriting principle and coinduction we are able to define a complete monitoring mechanism. A first implementation has been done into our AccLab tool support and sketched in [38]. We are investigating the extension of our LTL rewriting mechanism to cope with the first-order case.

7.1.3.2. Specification of advanced security and privacy properties.

Security and privacy requirements in ubiquitous systems need a sophisticated policy language with features to express access restrictions and obligations. Ubiquitous systems involve multiple actors owning sensitive data concerning aspects such as location, discrete and continuous time, multiple roles that can be shared among actors or evolve over time. Conflict management is an important problem in security policy frameworks. In [31] we present an abstract language (AAL) dedicated to accountability. We show how to specify most of these security and privacy features and compare it with the XACML approach. We also classified the existing conflict detection for XACML like approaches in dynamic, testing, or static detection. A thorough analysis of these mechanisms reveals that they have several weaknesses and they are not applicable in our context. We advocate for a classic approach using the notion of logical consistency to detect conflicts in AAL.

7.1.3.3. Composition of privacy-enhancing and security mechanisms.

As part of his PhD thesis [11], Ronan Cherrueau's has defined a language for the composition of three privacy-enhancing and security mechanisms: symmetric key encryption, database fragmentation and on-client computations. The language allows the expression of distributed programs that protect data by applying compositions of the three mechanisms to them. The language ensures basic privacy and security properties by a type system based on dependent types. This type system ensures, for example, that data that has been encrypted and stored in a database fragment cannot be accessed in plain form and from another location than that fragment. Furthermore, the language comes equiped with four major additional results. First, a calculus that allows for the semi-automatic derivation of distributed privacy-preserving and secure programs from an original non-distributed one. Second, a transformation from the language to the π-calculus. Third, a transformation into an input specification to the Proverif model checker for security properties. Fourth, two implementations on the basis of, respectively, the Scala and Idris languages that harness their corresponding dependent type systems.

7.2. Distributed programming and the Cloud

Participants: Frederico Alvares, Bastien Confais, Simon Dupont, Md Sabbir Hasan, Adrien Lebre, Thomas Ledoux, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Rémy Pottier, Anthony Simonet, Mario Südholt.

7.2.1. Cloud applications and infrastructures

Complex event processing. We presented this year the evolution of SensorScript towards a language for complex event processing dedicated to sensor networks. While the model mainly relies on previous works, we highlighted how the new language builds on the multitree in order to provide complex event processing mechanisms. We are able to balance the syntactic concision of the language with a real-time complex event processor for sensor networks. By providing flexible selections over the nodes, with the possibility to filter

them on complex conditions, possibly over a time window, we offer a strong alternative to traditional SQL used in the literature. Moreover, SensorScript does not focus only on data access. In fact it provides the possibility to widen the scope of the methods accessible on nodes to other features than sensors monitoring, including but not limited to addressing actuators functions. Finally we showed that SensorScript is able to address examples proposed in the literature, with simpler results than SQL, while highlighting its limitations, especially on history management. [24]

Secure cloud storage. The increasing number of cloud storage services like Dropbox or Google Drive allows users to store more and more data on the Internet. However, these services do not give users enough guarantees in protecting the privacy of their data. In order to limit the risk that the storage service scans user documents for commercial purposes, we propose a storage service that stores data on several cloud providers while preventing these providers to read user documents. TrustyDrive is 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. [30]

7.2.1.1. Service-level agreement for the Cloud.

Quality-of-service and SLA guarantees are among the major challenges of cloud-based services. In [19], we first present a new cloud model called SLAaaS — SLA aware Service. SLAaaS considers QoS levels and SLA as first class citizens of cloud-based services. This model is orthogonal to other SaaS, PaaS, and IaaS cloud models, and may apply to any of them. More specifically, we make three contributions: (i) we provide a domain-specific language that allows to define SLA constraints in cloud services; (ii) we present a general control-theoretic approach for managing cloud service SLA; (iii) we apply our approach to MapReduce, locking, and e-commerce services.

7.2.1.2. Cloud Capacity Planning and Elasticity.

Capacity management is a process used to manage the capacity of IT services and the IT infrastructure. Its primary goal is to ensure that IT resources (services, infrastructure) are right-sized to meet current and future requirements in a cost-effective and timely manner. In [34], we present a comprehensive overview of capacity planning and management for cloud computing. First, we state the problem of capacity management in the context of cloud computing from the point of view of several service providers. Second, we provide a brief discussion about *when* capacity planning should take place. Finally, we survey a number of methods for capacity planning and management proposed by both people from industry and researchers.

In his PhD [12], Simon Dupont proposes to extend the concept of elasticity to higher layers of the cloud, and more precisely to the SaaS level. He presents the new concept of *software elasticity* by defining the ability of the software to adapt, ideally in an autonomous way, to cope with workload changes and/or limitations of IaaS elasticity. This brings the consideration of Cloud elasticity in a multi-layer way through the adaptation of all kind of Cloud resources (software, virtual machines, physical machines). In [23], we introduce ElaScript, a DSL that 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 Grid'5000 infrastructure testbed.

7.2.1.3. Infrastructure.

Academic and industry experts are now advocating for going from large-centralized Cloud Computing infrastructures to smaller ones massively distributed at the edge of the network (aka., Fog and Edge Computing solutions). Among the obstacles to the adoption of this model is the development of a convenient and powerful IaaS system capable of managing a significant number of remote data-centers in a unified way.

In 2016, we achieved three major results in this context.

The first result is related to the economical viability of Fog/Edge Computing infrastructures that is often debated w-r-t large cloud computing data centers operated by US giants such as Amazon, Google .... To answer such a question, we conducted a specific study that goes beyond the state of the art of the current cost model of Distributed Cloud infrastructures. First, we provided a classification of the different ways of deploying Distributed Cloud platforms. Then, we proposed a versatile cost model that can help new actors evaluate the viability of deploying a Fog/Edge Computing offer. We illustrated the relevance of our proposal by instantiating it over three use-cases and comparing them according to similar computation capabilities provided by the Amazon solution. Such a study clearly showed that deploying a Distributed Cloud infrastructure makes sense for telcos as well as new actors willing to enter the game [29].

The second result is related to the preliminary revisions we made in OpenStack. The OpenStack software suite has become the de facto open-source solution to operate, supervise and use a Cloud Computing infrastructure. Our objective is to study to what extent current OpenStack mechanisms can handle massively distributed cloud infrastructures and to propose revisions/extensions of internal mechanisms when appropriate. The work we conducted this year focused on the Nova service of OpenStack.More precisely, we modified the code base in order to use a distributed key/value store instead of the centralized SQL backend. We conducted several experiments that validate the correct behavior and gives performance trends of our prototype through an emulation of several data-centers using Grid'5000 testbed. In addition to paving the way to the first large-scale and Internet-wide IaaS manager, we expect this work will attract a community of specialists from both distributed system and network areas to address the Fog/Edge Computing challenges within the OpenStack ecosystem [36], [27]. These and additional corresponding results have been presented in a more detailed manner as part of Jonathan Pastor's PhD thesis [14].

The third result is related to the data management in Fog/Edge Computing infrastructures. Our ultimate goal is to propose an Amazon-S3 like system, *i.e.*, a blob storage service, that can take into account Fog/Edge specifics. The study we achieved this year is preliminary. We first identified a list of properties a storage system should meet in this context. Second, we evaluated through performance analysis three "off-the-shelf" object store solutions, namely Rados, Cassandra and InterPlanetary File System (IPFS). In particular, we focused (i) on access times to push and get objects under different scenarios and (ii) on the amount of network traffic that is exchanged between the different sites during such operations. We also evaluated how the network latencies influence the access times and how the systems behave in case of network partitioning. Experiments have been conducted using the Yahoo Cloud System Benchmark (YCSB) on top of the Grid'5000 testbed. We showed that among the three tested solutions IPFS fills most of the criteria expected for a Fog/Edge computing infrastructure. [33], [32]

7.2.2. Renewable energy

With the emergence of the Future Internet and the dawning of new IT models such as cloud computing, the usage of data centers (DC), and consequently their power consumption, increase dramatically. Besides the ecological impact, the energy consumption is a predominant criterion for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale distributed systems. We have design the EpoCloud prototype, from hardware to middleware layers. This prototype aims at optimizing the energy consumption of mono-site Cloud DCs connected to the regular electrical grid and to renewable-energy sources. [17]

7.2.2.1. Green Energy awareness in SaaS Application.

With the proliferation of Cloud computing, data centers have to urgently face energy consumption issues. Although recent efforts such as the integration of renewable energy to data centers or energy efficient techniques in (virtual) machines contribute to the reduction of carbon footprint, creating green energy awareness around *Interactive Cloud Applications* by smartly using the presence of green energy has not been yet addressed. By *awareness*, we mean the inherited capability of SaaS applications to dynamically adapt with the availability of green energy and to reduce energy consumption while green energy is scarce or absent. In [25], we present two application controllers based on different metrics (e.g., availability of green energy, response time, user experience level). Based on extensive experiments with a real application benchmark and

workloads in Grid'5000, results suggest that providers revenue can be increased as high as 64%, while 13% brown energy can be reduced without deprovisioning any physical or virtual resources at IaaS layer and 17 fold increment of performance can be guaranteed.

DIVERSE Project-Team

7. New Results

7.1. Results on Variability modeling and management

7.1.1. Feature Model Synthesis: Algorithms and Empirical Studies

We attack the problem of synthesising feature models by considering both configuration semantics and ontological semantics of a feature model. We define a generic synthesis procedure that computes the likely siblings or parent candidates for a given feature. We develop six heuristics for clustering and weighting the logical, syntactical and semantical relationships between feature names. We then perform an empirical evaluation on hundreds of feature models, coming from the SPLOT repository and Wikipedia. We provide evidence that a fully automated synthesis (i.e., without any user intervention) is likely to produce models far from the ground truths. As the role of the user is crucial, we empirically analyze the strengths and weaknesses of heuristics for computing ranking lists and different kinds of clusters. We show that a hybrid approach mixing logical and ontological techniques outperforms state-of-the-art solutions.

Numerous synthesis techniques and tools have been proposed, but only a few consider both configuration and ontological semantics of a feature model. We also boil down several feature model management operations to a synthesis problem. Our approach, the FAMILIAR environment, and empirical results support researchers and practitioners working on feature models. The synthesis problem is a core issue when reverse engineering, merging, slicing, or refactoring feature models. An article has been published in 2016 at Empirical Software Engineering journal, a major avenue for software engineering research [19].

7.1.2. Product Comparison Matrix

Product Comparison Matrices (PCMs) are widely used for documenting or comparing a set of products. PCMs are simple tabular data in which products are usually organized as rows, features as columns, while each cell define how a product implements the corresponding feature. We develop metamodeling and feature modeling techniques for formalizing PCMs. We perform numerous empirical experiments with users, tools, and data for validating our proposal. We also develop automated techniques to extract PCMs out of informal product descriptions, written in natural language. We establish a connection between PCMs and variability modeling formalism, which is of interest for the product line community. OpenCompare is a direct output of this research and is an important step towards the creation of a community around PCMs. We mined millions of Wikipedia tabular data together with end-users and developers to cross-validate our model-based approach [19]. We also mined data from BestBuy [17].

7.1.3. Machine Learning and Variability Testing

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. We validate our approach on a product-line video generator, using a simple computer vision algorithm as an oracle. We show that an interesting set of cross-tree constraint can be generated, with reasonable precision and recall. Our learning-based testing technique complements our initial effort in engineering an industrial video generator. The use of learning allows to significantly narrow the configuration space and discover complex constraints, hard to discover even for experts. We conduct a series of work in the computer vision domain to generate variants of videos, investigating the usefulness and effectiveness of variability techniques in novel areas. Our approach is novel and general: the same principles can be applied to other configurable systems [55].

7.1.4. Enumeration of All Feature Model Configurations

Feature models are widely used to encode the configurations of a software product line in terms of mandatory, optional and exclusive features as well as propositional constraints over the features. Numerous computationally expensive procedures have been developed to model check, test, configure, debug, or compute relevant information of feature models. We explore the possible improvement of relying on the enumeration of all configurations when performing automated analysis operations. We tackle the challenge of how to scale the existing enumeration techniques by relying on distributed computing. We show that the use of distributed computing techniques might offer practical solutions to previously unsolvable problems and opens new perspectives for the automated analysis of software product lines [40].

7.1.5. Software Unbundling

Unbundling is a phenomenon that consists of dividing an existing software artifact into smaller ones. It can happen for different reasons, one of them is the fact that applications tend to grow in functionalities and sometimes this can negatively influence the user experience. It can be seen as a way to produce different variants of an application. For example, mobile applications from well-known companies are being divided into simpler and more focused new ones. Despite its current importance, little is known or studied about unbundling or about how it relates to existing software engineering approaches, such as modularization. Consequently, recent cases point out that it has been performed unsystematically and arbitrarily. Our main goal is to present this novel and relevant concept and its underlying challenges in the light of software engineering, also exemplifying it with recent cases. We relate unbundling to standard software modularization, presenting the new motivations behind it, the resulting problems, and drawing perspectives for future support in the area [23].

7.1.6. Featured Model Types

By analogy with software product reuse, the ability to reuse (meta)models and model transformations is key to achieve better quality and productivity. To this end, various opportunistic reuse techniques have been developed, such as higher-order transformations, metamodel adaptation, and model types. However, in contrast to software product development that has moved to systematic reuse by adopting (model-driven) software product lines, we are not quite there yet for modelling languages, missing economies of scope and automation opportunities. Our vision is to transpose the product line paradigm at the metamodel level, where reusable assets are formed by metamodel and transformation fragments and "products" are reusable language building blocks (model types). We introduce featured model types to concisely model variability amongst metamodelling elements, enabling configuration, automated analysis, and derivation of tailored model types [53].

7.1.7. A Formal Modeling and Analysis Framework for SPL of Pre-emptive Real-time Systems

We present a formal analysis framework to analyze a family of platform products w.r.t. real-time properties. First, we propose an extension of the widely-used feature model, called Property Feature Model (PFM), that distinguishes features and properties explicitly Second, we present formal behavioral models of components of a real-time scheduling unit such that all real-time scheduling units implied by a PFM are automatically composed to be analyzed against the properties given by the PFM. We apply our approach to the verification of the schedulability of a family of scheduling units using the symbolic and statistical model checkers of Uppaal [44].

7.1.8. Exploration of Architectural Variants

In systems engineering, practitioners shall explore numerous architectural alternatives until choosing the most adequate variant. The decision-making process is most of the time a manual, time-consuming, and error-prone activity. The exploration and justification of architectural solutions is ad-hoc and mainly consists in a series of tries and errors on the modeling assets.

We report on an industrial case study in which we apply variability modeling techniques to automate the assessment and comparison of several candidate architectures (variants). We first describe how we can use a model-based approach such as the Common Variability Language (CVL) to specify the architectural variability. We show that the selection of an architectural variant is a multi-criteria decision problem in which there are numerous interactions (veto, favor, complementary) between criteria. We present a tooled process for exploring architectural variants integrating both CVL and the MYRIAD method for assessing and comparing variants based on an explicit preference model coming from the elicitation of stakeholders' concerns. This solution allows understanding differences among variants and their satisfactions with respect to criteria. Beyond variant selection automation improvement, this experiment results highlight that the approach improves rationality in the assessment and provides decision arguments when selecting the preferred variants. It is a joint work and collaboration with Thales [47].

7.1.9. A Complexity Tale: Web Configurators

Online configurators are basically everywhere. From physical goods (cars, clothes) to services (cloud solutions, insurances, etc.) such configurators have pervaded many areas of everyday life, in order to provide the customers products tailored to their needs. Being sometimes the only interfaces between product suppliers and consumers, much care has been devoted to the HCI aspects of configurators, aiming at offering an enjoyable buying experience. However, at the backend, the management of numerous and complex configuration options results from ad-hoc process rather than a systematic variability-aware engineering approach. We present our experience in analysing web configurators and formalising configuration options in terms of feature models or product configuration matrices. We also consider behavioural issues and perspectives on their architectural design [32].

7.2. Results on Software Language Engineering

7.2.1. Safe Model Polymorphism for Flexible Modeling

Domain-Specific Languages (DSLs) are increasingly used by domain experts to handle various concerns in systems and software development. To support this trend, the Model-Driven Engineering (MDE) community has developed advanced techniques for designing new DSLs. However, the widespread use of independently developed, and constantly evolving DSLs is hampered by the rigidity imposed to the language users by the DSLs and their tooling, e.g., for manipulating a model through various similar DSLs or successive versions of a given DSL. In [24] we propose a disciplined approach that leverages type groups' polymorphism to provide an advanced type system for manipulating models, in a polymorphic way, through different DSL interfaces. A DSL interface, aka. model type, specifies a set of features, or services, available on the model it types, and subtyping relations among these model types define the safe substitutions. This type system complements the Melange language workbench and is seamlessly integrated into the Eclipse Modeling Framework (EMF), hence providing structural interoperability and compatibility of models between EMF-based tools. We illustrate the validity and practicability of our approach by bridging safe interoperability between different semantic and syntactic variation points of a finite-state machine (FSM) language, as well as between successive versions of the Unified Modeling Language (UML).

7.2.2. Execution Framework for Model Debugging

The development and evolution of an advanced modeling environment for a Domain-Specific Modeling Language (DSML) is a tedious task, which becomes recurrent with the increasing number of DSMLs involved in the development and management of complex software-intensive systems. Recent efforts in language workbenches result in advanced frameworks that automatically provide syntactic tooling such as advanced editors. However, defining the execution semantics of languages and their tooling remains mostly hand crafted. Similarly to editors that share code completion or syntax highlighting, the development of advanced debuggers, animators, and others execution analysis tools shares common facilities, which should be reused among various DSMLs. In [37] we present the execution framework offered by the GEMOC studio, an Eclipse-based language and modeling workbench. The framework provides a generic interface to plug in different execution

engines associated to their specific metalanguages used to define the discrete-event operational semantics of DSMLs. It also integrates generic runtime services that are shared among the approaches used to implement the execution semantics, such as graphical animation or omniscient debugging.

7.2.3. Variability Management in Language Families

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 [29], we report on an effort for organizing the literature on language product line engineering. More precisely, we propose a definition for the life-cycle of language product lines, and we use it to analyze the capabilities of current approaches. In addition, we provide a mapping between each approach and the technological space it supports.

Still, the identification and definition of language modules are complex and error-prone activities, thus hindering the reuse exploitation when developing DSLs. In [50], [51], we propose a computer-aided approach to i) identify potential reuse in a set of legacy DSLs; and ii) capitalize such potential reuse by extracting a set of reusable language modules with well defined interfaces that facilitate their assembly. We validate our approach by using realistic DSLs coming out from industrial case studies and obtained from public GitHub repositories. We also developed a publicly available tool, namely Puzzle, that uses static analysis to facilitate the detection of specification clones in DSLs implemented under the executable metamodeling paradigm. Puzzle also enables the extraction specification clones as reusable language modules that can be later used to build up new DSLs.

7.2.4. A Tool-Supported Approach for Concurrent Execution of Heterogeneous Models

In the software and systems modeling community, research on domain-specific modeling languages (DSMLs) is focused on providing technologies for developing languages and tools that allow domain experts to develop system solutions efficiently. Unfortunately, the current lack of support for explicitly relating concepts expressed in different DSMLs makes it very difficult for software and system engineers to reason about information spread across models describing different system aspects. As a particular challenge, we investigate in [38] relationships between, possibly heterogeneous, behavioral models to support their concurrent execution. This is achieved by following a modular executable metamodeling approach for behavioral semantics understanding, reuse, variability and composability. This approach supports an explicit model of concurrency (MoCC) and domain-specific actions (DSA) with a well-defined protocol between them (incl., mapping, feedback and callback) reified through explicit domain-specific events (DSE). The protocol is then used to infer a relevant behavioral language interface for specifying coordination patterns to be applied on conforming executable models. All the tooling of the approach is gathered in the GEMOC studio, and outlined in the next section. Currently, the approach is experienced on a systems engineering language provided by Thales, named Capella.

7.2.5. Various Dimensions of Reuse

Reuse, enabled by modularity and interfaces, is one of the most important concepts in software engineering. This is evidenced by an increasingly large number of reusable artifacts, ranging from small units such as classes to larger, more sophisticated units such as components, services, frameworks, software product lines, and concerns. We give evidence in [43] that a canonical set of reuse interfaces has emerged over time: the variation, customization, and usage interfaces (VCU). A reusable artifact that provides all three interfaces reaches the highest potential of reuse, as it explicitly exposes how the artifact can be manipulated during the reuse process along these three dimensions. We demonstrate the wide applicability of the VCU interfaces along two axes: across abstraction layers of a system specification and across existing reuse techniques. The former is shown with the help of a comprehensive case study including reusable requirements, software, and hardware models for the authorization domain. The latter is shown with a discussion on how the VCU interfaces relate to existing reuse techniques.

7.2.6. Modeling for Sustainability

The complex problems that computational science addresses are more and more benefiting from the progress of computing facilities (e.g., simulators, librairies, accessible languages). Nevertheless, the actual solutions call for several improvements. Among those, we address the needs for leveraging on knowledge and expertise by focusing on Domain-Specific Modeling Languages application. In this work we explored, through concrete experiments, how the last DSML research help getting closer the problem and implementation spaces.

Various disciplines use models for different purposes. While engineering models, including software engineering models, are often developed to guide the construction of a non- existent system, scientific models, in contrast, are created to better understand a natural phenomenon (i.e., an already existing system). An engineering model may incorporate scientific models to build a system. Both engineering and scientific models have been used to support sustainability, but largely in a loosely-coupled fashion, independently developed and maintained from each other. Due to the inherent complex nature of sustainability that must balance trade-offs between social, environmental, and economic concerns, modeling challenges abound for both the scientific and engineering disciplines. In [39] we propose a vision that synergistically combines engineering and scientific models to enable broader engagement of society for addressing sustainability concerns, informed decision-making based on more accessible scientific models and data, and automated feed-back to the engineering models to support dynamic adaptation of sustainability systems. To support this vision, we identify a number of challenges to be addressed with particular emphasis on the socio-technical benefits of modeling.

As first experiments, we presented at the EclipseCon France, Europe and North America 2016, an approach to develop smart cyber physical systems in charge of managing the production, distribution and consumption of energies (e.g., water, electricity). The main objective is to enable a broader engagement of society, while supporting a more informed decision-making, possibly automatically, on the development and run-time adaptation of sustainability systems (e.g., smart grid, home automation, smart cities). We illustrate this approach through a system that allows farmers to simulate and optimize their water consumption by combining the model of a farming system together with agronomical models (e.g., vegetable and animal lifecycle) and open data (e.g., climate series). To do so, we use Model Driven Engineering (MDE) and Domain Specific Languages (DSL) to develop such systems driven by scientific models that define the context (e.g., environment, social and economy), and model experiencing environments to engage general public and policy makers.

7.2.7. Formal Specification of a Packet Filtering Language Using the K Framework

Many project-specific languages, including in particular filtering languages, are defined using non-formal specifications written in natural languages. This leads to ambiguities and errors in the specification of those languages. In [46] we report on an industrial experiment on using a tool-supported language specification framework (K) for the formal specification of the syntax and semantics of a filtering language having a complexity similar to those of real-life projects. This experimentation aims at estimating, in a specific industrial setting, the difficulty and benefits of formally specifying a packet filtering language using a tool-supported formal approach.

7.2.8. Correct-by-construction model driven engineering composition operators

Model composition is a crucial activity in Model Driven Engineering both to reuse validated and verified model elements and to handle separately the various aspects in a complex system and then weave them while preserving their properties. Many research activities target this compositional validation and verification (V & V) strategy: allow the independent assessment of components and minimize the residual V & V activities at assembly time. However, there is a continuous and increasing need for the definition of new composition operators that allow the reconciliation of existing models to build new systems according to various requirements. These ones are usually built from scratch and must be systematically verified to assess that they preserve the properties of the assembled elements. This verification is usually tedious but is mandatory to avoid verifying the composite system for each use of the operators. Our work addresses these issues, we first target the use of proof assistants for specifying and verifying compositional verification frameworks relying on formal verification techniques instead of testing and proofreading. Then, using a divide

and conquer approach, we focus on the development of elementary composition operators that are easy to verify and can be used to further define complex composition operators. In our approach [27], proofs for the complex operators are then obtained by assembling the proofs of the basic operators. To illustrate our proposal, we use the Coq proof assistant to formalize the language-independent elementary composition operators Union and Substitution and the proof that the conformance of models with respect to metamodels is preserved during composition. We show that more sophisticated composition operators that share parts of the implementation and have several properties in common (especially: aspect oriented modeling composition approach, invasive software composition, and package merge) can then be built from the basic ones, and that the proof of conformance preservation can also be built from the proofs of basic operators.

7.2.9. Engineering Modeling Languages

The DiverSE project-team is deeply involved in transferring research knowledge into education. In particular, one book in English have been published in 2016 as a textbook [59]. The book cover the broad scope of MDE, and are based on the experience of the project-team members.

7.3. Results on Heterogeneous and dynamic software architectures

We have selected three main contributions : two are in the field of runtime management, while the third one is in the field of non-functionnal software testing.

7.3.1. Precise and efficient resource management using models@runtime

Contribution. We have developed an efficient monitoring framework to quickly spot an abnormal resource consumption within a complex application. In these papers [25], we have proposed an optimistic adaptive monitoring system to determine the faulty components of an application. Suspected components are finely analyzed by the monitoring system, but only when required. Unsuspected components are left untouched and execute normally.

Originality. Current solutions that perform permanent and extensive monitoring to detect anomalies induce high overhead on the system, and can, by themselves, make the system unstable. Our system performs localized just-in-time monitoring that decreases the accumulated overhead of the monitoring system. Through our evaluation, we show that our technique correctly detects faulty components, while reducing overhead by 92.98 on average%.

Impact. Beyond the scientific originality of this work, the main impacts of this novel approach approach to monitor software component performance 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 (HEADS and B-com PhD thesis). This work has been integrated within the Kevoree platform.

7.3.2. Dynamic web application using models@runtime

Contribution. We have developed a component-based platform supporting the development of dynamically adaptable single Web page applications. An important part of this contribution lies in the possibility to dynamically move code from the server to the client side allowing a great flexibility in the performance management. This contribution [56] is based on a models@runtime approach and has been implemented in our open source KevoreeJS platform.

Originality. Current solutions to create single Web page application are limited to a static code repartition between clients and server, thus limiting the flexibility at runtime.

Impact. Beyond the scientific originality of this work, the main impacts of this novel approach to monitor software component performance has been to (i) reinforce DIVERSE's visibility in the open-source community, (ii) to start several research tracks that are currently explored in different projects of the team (HEADS, STAMP, GRevis). This platforms is modular, one of the component has a monthly download count greater than 100k [0]).

[0] https://www.npmjs.com/package/npmi

7.3.3. Testing non-functional behavior of compiler and code generator

Contribution. We have developed NOTICE [36], [35], a component-based framework for non-functional testing of compilers through the monitoring of generated code in a controlled sand-boxing environment. In this work, we have proposed an automatic way of testing non-functional properties of compilers, while optimizing the generated application with respect to a set of specific non-functional properties (CPU, memory usage, energy consumption, *etc.*).

Originality. Compiler users generally apply different optimizations to generate efficient code with respect to specific non-functional properties such as energy consumption, execution time, etc. However, due to the huge number of optimizations provided by modern compilers, finding the best optimization sequence for a specific objective and a given program is more and more challenging.

Impact. Beyond the scientific originality of this work, the main impact of this novel approach is to enable the auto-tuning of compilers according to user requirements and to construct optimizations that yield to performance results that are better than standard optimization levels.

7.3.4. Automatic Microbenchmark Generation to Prevent Dead Code Elimination and Constant Folding

Contribution. Microbenchmarking consists of evaluating, in isolation, the performance of small code segments that play a critical role in large applications. The accuracy of a microbenchmark depends on two critical tasks: wrap the code segment into a payload that faithfully recreates the execution conditions that occur in the large application; build a scaffold that runs the payload a large number of times to get a statistical estimate of the execution time. While recent frameworks such as the Java Microbenchmark Harness (JMH) take care of the scaffold challenge, developers have very limited support to build a correct payload. This year, we focus on the automatic generation of pay-loads, starting from a code segment selected in a large application [54]. In particular, we aim at preventing two of the most common mistakes made in microbenchmarks: dead code elimination and constant folding. Since a microbench-mark is such a small program, if not designed carefully, it will be *over-optimized* by the JIT and result in distorted time measures. Our technique hence automatically extracts the segment into a compilable payload and generates additional code to prevent the risks of *over-optimization*. The whole approach is embedded in a tool called AutoJMH, which generates payloads for JMH scaffolds. We validate the capabilities AutoJMH, showing that the tool is able to process a large percentage of segments in real programs. We also show that AutoJMH can match the quality of payloads handwritten by performance experts and outperform those written by professional Java developers without experience un microbenchmarking.

7.3.5. 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 [49]. Tellu relies on Kevoree and KevoreeJS to build their health management systems. They will be also a 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 [52].

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 through multiple sensors, run eternally and are open to other services and to users. Here we present our latest observational and technical results about (i) new approaches to increase diversity in software systems, and (ii) software testing to assess the validity of software.

7.4.1. Software diversification

A main achievement in our investigations of software diversity, is a large scale analysis of browser fingerprints [45]. Browser fingerprinting consists in collecting information about a user's browser and its execution environment. A distinctive feature of these fingerprints is that they are unique and can be used to track users. We show that innovations in HTML5 provide access to highly discriminating attributes, notably with the use of the Canvas API which relies on multiple layers of the user's system. In addition, we show that browser fingerprinting is as effective on mobile devices as it is on desktops and laptops, albeit for radically different reasons due to their more constrained hardware and software environments. We also evaluate how browser fingerprinting could stop being a threat to user privacy if some technological evolutions continue (e.g., disappearance of plugins) or are embraced by browser vendors (e.g., standard HTTP headers).

As for automatic diversification of programs, we have had a strong focus on runtime transformations. Online Genetic Improvement embeds the ability to evolve and adapt inside a target software system enabling it to improve at runtime without any external dependencies or human intervention. We recently developed a general purpose tool enabling Online Genetic Improvement in software systems running on the java virtual machine. This tool, dubbed ECSELR, is embedded inside extant software systems at runtime, enabling such systems to autonomously generate diverse variants [31]. We have also worked on diversification against just-in-Time (JIT) Spraying: a technique that embeds return-oriented programming (ROP) gadgets in arithmetic or logical instructions as immediate offsets. We introduce libmask, a JIT compiler extension that transforms constants into global variables and marks the memory area for these global variables as read only. Hence, any constant is referred to by a memory address making exploitation of arithmetic and logical instructions more difficult. Then, these memory addresses are randomized to further harden the security [42].

7.4.2. Software testing

Our work in the area of software testing focuses on tailoring the testing tools (analysis, generation, oracle, etc.) to specific domains and purposes. This allows us to consider domain specific knowledge (e.g., architectural patterns for GUI implementation) in order to increase the relevance and the efficiency of testing. The main results of this year are about test case refactoring and testing code generators.

Software developers design test suites to verify that software meets its expected behaviors. Yet, many dynamic analysis techniques are performed on the exploitation of execution traces from test cases. In practice, one test case may imply various behaviors. However, the execution of a test case only yields one trace, which can hide the others. We have developed a new technique of test code refactoring, which splits a test case into small test fragments that cover a simpler part of the control flow to provide better support for dynamic analysis. This technique can effectively improve the execution traces of the test suite: exception contracts are better verified via applying this refactoring to original test suites [30].

Finding the smallest set of valid test configurations that ensure sufficient coverage of the system's feature interactions is essential, especially when the execution of test configurations is costly or time-consuming. However, this problem is NP-hard in general and approximation algorithms have often been used to address it in practice. We explore an approach based on constraint programming to increase the effectiveness of configuration testing while keeping the number of configurations as low as possible. For 79% of 224 feature models, our technique generated up to 60% fewer test configurations than the competitor tools [26].

The intensive use of generative programming techniques provides an elegant engineering solution to deal with the heterogeneity of platforms and technological stacks. Yet, producing correct and efficient code generator is complex and error-prone. We describe a practical approach based on a runtime monitoring infrastructure to automatically check the potential inefficient code generators. We evaluate our approach by analyzing the

performance of Haxe, a popular high-level programming language that involves a set of cross-platform code generators. The results show that our approach is able to detect some performance inconsistencies that reveal real issues in Haxe code generators [36], [35]

Graphical User Interfaces (GUIs) intensively rely on event-driven programming: widgets send GUI events, which capture users' interactions, to dedicated objects called *controllers*. Controllers implement several *GUI listeners* that handle these events to produce GUI commands. We study to what extent the number of GUI commands that a GUI listener can produce has an impact on the code quality. We then identify a new type of design smell, called *Blob listener* that characterizes GUI listeners that can produce more than two GUI commands. We propose a systematic static code analysis procedure that searches for *Blob Listener* instances that we implement in *InspectorGuidget* [48].

FOCUS Project-Team

7. New Results

7.1. Service-oriented computing

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

7.1.1. Microservices

Microservices is an emerging paradigm for the development of distributed systems that, originating from Service-Oriented Architecture, fosters the creation of an ecosystem of reusable components by focusing on the small dimension, the loose coupling, and the dynamic topology of services. Their dynamic nature calls for suitable techniques that support automatic deployment. In [40] we address this problem and we propose JRO (Jolie Redeployment Optimiser), a tool for the automatic and optimised deployment of microservices written in the Jolie language. The flexibility of microservices is their key advantage, yet it poses many security issues. In [39] we classify the most relevant vulnerabilities related to data reliability, integrity, and authenticity, and we investigate directions for their mitigation.

7.1.2. Orchestrations and choreographies

The practice of programming distributed systems is one of the most 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, choreographies can be compiled to obtain projected systems that enjoy freedom from deadlocks and races by construction. In [10] we studied how to make choreographies, and extensions of them that allow one to perform dynamic updates, a suitable tool for real-world programming.

7.2. Models for reliability

Participants: Elena Giachino, Ivan Lanese.

7.2.1. Reversibility

We have continued the study of causal-consistent reversibility started in the past years. In particular, we concentrated on uncontrolled reversibility, where one specifies how a concurrent computation can go back to past states, without giving policies about when to do that. In [25] we thoroughly studied the problem for higher-order pi-calculus. In particular, we studied the causality structures needed to enable reversibility, and we related them with the causal semantics of Boreale and Sangiorgi. In [30] we proposed a modular approach that, given a formal model equipped with both an LTS semantics and an independence relation capturing causality, defines a causal-consistent reversible semantics for it. The approach is very general, capturing models as different as CCS and concurrent X-machines, but it is not fully automatic.

7.3. Cloud Computing and Deployment

Participants: Elena Giachino, Saverio Giallorenzo, Claudio Guidi, Cosimo Laneve, Gianluigi Zavattaro.

7.3.1. Static deployment

We have continued our foundational investigation of the Aeolus component model for the automatic deployment of a component-based application in a cloud environment. In [42] we have refined a previous Turing completeness result for the Aeolus model. In fact, a previous proof of undecidability of the deployment problem assumes the possibility of performing in a synchronized way atomic configuration actions on a set of interdependent components: this feature is usually not supported by actual deployment frameworks. To make the theoretical model used for our undecidability result closer to the real deployment infrastructures, in [42] we have proved that even without synchronized configuration actions the Aeolus component model is still Turing complete.

7.3.2. Dynamic deployment

We have analyzed linguistic mechanisms for expressing and managing dynamic aspects of deployment, in particular the possibility to dynamically modify the architecture of an application.

In [17] we propose a new mechanism for Dynamic Rebinding, a particular kind of Dynamic Software Updating that focuses on modifying the workflow of a program. This mechanism is built upon the model of Concurrent Object Groups, which is adopted in programming languages like Coboxes, Creol or ABS. Using this model, which extends and solves some of the limitations of Active Objects, it becomes possible for an update to wait for the program to reach a local quiescent state and then perform the update without creating any inconsistency in the program's state.

In [34] we show how deployment can be added as a first-class citizen in the object-oriented modeling language ABS. We follow a declarative approach: programmers specify deployment constraints and a solver synthesizes ABS classes exposing methods like deploy (resp. undeploy) that executes (resp. cancels) configuration actions changing the current deployment towards a new one satisfying the programmer's desiderata. Differently from previous works, this novel approach supports the specification of incremental modifications, thus supporting the declarative modeling of elastic applications.

7.4. Probabilistic Systems and Resource Control

Participants: Martin Avanzini, Flavien Breuvart, Alberto Cappai, Raphaelle Crubillé, Ugo Dal Lago, Francesco Gavazzo, Charles Grellois, Simone Martini, Alessandro Rioli, Davide Sangiorgi, Marco Solieri, Valeria Vignudelli.

7.4.1. Probabilistic Systems

7.4.1.1. Behavioural Equivalences and Metrics

Finding effective methodologies to check program equivalence is one of the oldest problems in the theory of programming languages, and has been studied also in the realm of probabilistic programming idioms. One particularly fruitful research direction consists in *characterising* context equivalence, the most natural way to *define* equivalence in higher-order languages, by way of *coinductive* notions of equivalence akin to bisimulation. In 2016, Focus has been involved in defining notions of *environmental* bisimulation for probabilistic lambda-calculi [37], proving them not only adequate, but also fully-abstract. Environmental bisimulation, contrarily to *applicative* bisimulation, is robust enough to be applicable to languages with local store. Moreover, the proof of adequacy of environmental bisimulation turns out to be simpler than that of applicative bisimulation, the latter requiring sophisticated arguments from linear programming. In a probabilistic setting, programs are more naturally compared through metrics rather than through equivalences, due to their intrinsic quantitative nature. Nicely, coinductive methodologies for program equivalence can be generalised to metrics by way of so-called *behavioural metrics*. This year, we have studied behavioural metrics in the context of concurrent processes, and defined enhancements of the proof method based on bisimulation metrics, by extending the theory of up-to techniques to premetrics on discrete probabilistic concurrent processes [32].

7.4.1.2. Programming Languages for Machine Learning

In recent years, higher-order functional programming languages like Church, Anglican, and Venture, have proved to be extremely effective as ways to specify not algorithms but rather bayesian models in the context of machine learning. The operational semantics of these languages, and learning algorithms when applied to programs in these languages, have been so far defined only informally. In 2016, we developed the operational semantics of an untyped probabilistic lambda-calculus with continuous distributions, as a foundation for universal probabilistic programming languages like those cited above [31]. Our first contribution was to adapt the classic operational semantics of lambda-calculus to a continuous setting. Our second contribution was to formalise the implementation technique of trace Markov chain Monte Carlo (MCMC) for our calculus and to show its correctness.

7.4.2. Resource Control

7.4.2.1. Complexity Analysis of Higher-Order Functional Programs

Complexity analysis of higher-order programs have been one of the main research themes inside Focus since its inception. It remains so today, although the emphasis is progressively shifting towards problems related to the *implementation* of complexity analysis methodologies rather than on their foundations. One issue with most existing complexity analysis methodologies is that they are insensitive to the sharing of computations among subprograms. We have studied how the interpretation method and dependency tuples, two prominent complexity analysis techniques can be adapted to graph-rewriting, thus accounting for the possible performance gains due to sharing [38]. We have also collaborated to the development of TCT, the Tyrolean Complexity Tool [29], a state-of-the-art complexity analyzer for term rewrite systems, making it capable to efficiently apply not one but *many* methodologies to the input program. Finally, we studied how the geometry of interaction can provide effective ways to compile higher-order functional programs into circuits, thus guaranteeing space efficiency [21].

7.4.2.2. On the Foundations of Complexity Analysis

One of the main foundational issues in complexity analysis is whether simple time cost models can be proved invariant, i.e., polynomially related to low-level models like those traditionally defined on Turing machines. We have solved a long-standing open problem, and proved that the unitary cost model, namely that attributing unitary cost to each beta-reduction step, is invariant for the pure lambda-calculus when evaluated leftmost-outermost [12]. We have also studied to which extent traditional methodologies like the interpretation method and light logics can be adapted to higher-order languages [16] and processes [20], respectively.

7.5. Verification techniques

Participants: Daniel Hirschkoff, Elena Giachino, Cosimo Laneve, Davide Sangiorgi.

7.5.1. Deadlock detection

In [22] we present a framework for statically detecting deadlocks in a concurrent object-oriented language with asynchronous method calls and cooperative scheduling of method activations. Since this language features recursion and dynamic resource creation, deadlock detection is extremely complex and state-of-the-art solutions either give imprecise answers or do not scale. In order to augment precision and scalability we propose a modular framework that allows several techniques to be combined. The basic component of the framework is a front-end inference algorithm that extracts abstract behavioural descriptions of methods, called contracts, which retain resource dependency information. This component is integrated with a number of possible different back-ends that analyze contracts and derive deadlock information. As a proof-of-concept, we discuss two such back-ends: (i) an evaluator that computes a fixpoint semantics and (ii) an evaluator using abstract model checking.

In [36] we study deadlock detection in an actor model with wait-by-necessity synchronizations, a lightweight technique that synchronizes invocations when the corresponding values are strictly needed. This approach relies on the use of futures that are not given an explicit Future type. The approach we adopt allows for the implicit synchronization on the availability of some value (where the producer of the value might be decided at runtime), whereas previous work allowed only explicit synchronization on the termination of a well-identified request. This way we are able to analyze the data-flow synchronization inherent to languages that feature wait-by-necessity. We provide a type-system and a solver inferring the type of a program so that deadlocks can be identified statically. As a consequence we can automatically verify the absence of deadlocks in actor programs with wait-by-necessity synchronizations.

7.5.2. Service Level Agreement

There is a gap between run-time service behaviours and the contracted quality expectations with the customers that is due to the informal nature of service level agreements. In [41] we explain how to bridge the gap by formalizing service level agreements with metric functions. We therefore discuss an end-to-end analysis flow

that can either statically verify if a service code complies with a metric function or use run-time monitoring systems to report possible misbehaviours. In both cases, our approach provides a feedback loop to fix and improve the metrics and eventually the resource configurations of the service itself.

7.6. Type Systems

Participants: Daniel Hirschkoff, Simone Martini, Davide Sangiorgi.

7.6.1. Surveys

In [27], Martini elaborates the history of type systems, focusing on that fundamental period covering the seventies and the early eighties. It was then that types became the cornerstone of the programming language design, passing first from the abstract data type (ADT) movement and blossoming then into the object-oriented paradigm. The paper also discusses how it has been possible that a concept like ADTs, with its clear mathematical semantics, neat syntax, and straightforward implementation, can have given way to objects, a lot dirtier from any perspective the language theorist may take.

In another paper [45], the same author compares the notion of "type" as found in programming languages with that found in mathematical logic, pointing out also some important historical remarks such as the role of the Curry-Howard isomorphism. It is argued that there are three different characters at play in programming languages, all of them now called types: the technical concept used in language design to guide implementation; the general abstraction mechanism used as a modelling tool; the classifying tool inherited from mathematical logic.

Two further surveys concerns behavioural types. The successful application of behavioural types requires a solid understanding of several practical aspects, from their representation in a concrete programming language, to their integration with other programming constructs such as methods and functions, to design and monitoring methodologies that take behaviours into account. The survey [15] provides an overview of the state of the art of these aspects.

The behavioural type of a software component specifies its expected patterns of interaction using expressive type languages, so that types can be used to determine automatically whether the component interacts correctly with other components. Two related important notions of behavioural types are those of session types and behavioural contracts. The paper [24] surveys the main accomplishments of the last twenty years within these two approaches.

7.6.2. Subtyping and dualities in name-passing concurrency

The fusion calculi are simplifications of the π-calculus in which input and output are symmetric and restriction is the only binder. In [23], Hirschkoff et al. highlight a major difference between these calculi and the π-calculus from the point of view of types, proving some impossibility results for subtyping in fusion calculi. A modification of fusion calculi is then proposed that allows one to import subtype systems, and related results, from the π-calculus, and examine the consequences of such modifications on theory and expressiveness of the languages.

INDES Project-Team

5. New Results

5.1. Web programming

Participants: Cédric Duminy, Vincent Prunet, Bernard Serpette, Manuel Serrano [correspondant], Colin Vidal.

Hop.js [20], [22] is a new platform for web applications, potentially involving interconnected servers. The server-side execution is compatible with Node.js. Programmers then benefit from numerous existing libraries and applications. Hop.js also introduces distinctive programming features that are expressed in the HopScript programming language, a multitier extension of JavaScript. The Hop.js runtime embeds a multi-backends HopScript compiler.

The HopScript language extends JavaScript to consistently define the server and client part of a web application. HopScript supports syntactic forms that help creating HTML elements. It supports services that enable function calls over HTTP. Being at higher level than traditional Ajax programming, Hop.js services avoid the burden and pitfalls of URL management and explicit data marshalling. They combine the benefits of a high level RPC mechanism and low level HTTP compatibility.

Hop.js supports server-side and client-side parallelism. On the server, it first relies on its built-in pipelining architecture that automatically decodes HTTP requests in parallel. It also relies on server-side web workers that programs may explicitly launch to perform background tasks (functions and services). Each worker runs its own system thread. The service invocation and execution API fully integrates with the JavaScript execution flow, allowing synchronous and asynchronous operations on both client and server processes. The asynchronous response API can be combined with the worker API, allowing processing and asynchronous service responses to be delegated between workers. On the browser client-side parallelism relies on standard web workers.

Although Hop.js can be used to develop traditional web servers, it is particularly adapted to the development of web applications embedded into devices, where the server and client part of the application are intimately interoperating with each other. The programing model of Hop.js fosters the joint specification of server and client code, and allows the rapid development of web user interfaces, on the client, controlling the execution of the distributed application. By defining a single data model, providing functions that can run indifferently on both sides, and almost forgetting about client-server protocols, Hop.js seems well suited for agile development of web applications for this class of applications.

As an example, Hop.js has already been successfully used as the core framework to develop embedded and cloud applications for connected robots and IoT devices. In the context of a European industrial collaborative project, it has been used by various categories of programmers (mostly undergraduate internships, robotic experts, and professional engineers familiar with web development techniques) to build complex distributed applications, where various sort of digital equipments (computers, robots, small devices) communicate with each other, discover themselves, and collaborate. In all cases we have observed an easy adoption from everyone. The tons of JavaScript resources and examples available on the web helped internship students to rapidly become productive. Robotic experts were instantly able to start implementing Hop.js applications. Web experts seemed to feel at home with Hop.js as it let them build working applications with Hop.js core features and then extend them with existing JavaScript third party modules, typically npm modules.

In 2016, we first version of Hop.js as been completed and released. It is available from the Web site http://hop.inria.fr.

5.1.1. Web Reactive Programming

Web UI interfaces are specified as HTML documents. When instantiated in a browser these documents are accessible from JavaScript as abstract data structures conforming to the Document Object Model (*aka* the DOM). Modifying these structures, for instance for applying updates, involves fine surgery for isolating the concerned elements and for applying the intended modifications. As these operations are generally triggered after asynchronous events that may come in response to earlier network requests or a user actions, the programming is complex and error prone. Improving on that situation has been the subject of many previous studies that propose alternative models for helping programming Web UI. Our work constitutes yet another contribution to that problem. It differs from the other solutions by the followings.

- It addresses exclusively the problem of programming the Web UI updates.
- It does not introduce a new programming model and it is fully compatible with traditional JavaScript programming.
- On the client, it only requires a very thin implementation layer whose weight is almost unnoticed in a Web browser.
- It does not impact the rest of the execution, leaving the performances unchanged.

Our proposal consists in introducing a zest of reactive programming used only for denoting the parts of the DOM that need updates. For that, we introduce two new constructs: i) reactive values, called *reactors*, that have the appearance of any regular JavaScript value, and ii) *reactive nodes*, which are DOM nodes that are automatically updated upon reactors changes. Reactors and reactive nodes can be used in pure JavaScript programs but that have been designed to complement other facilities Hop.js. To justify their design and to advocate their benefit, we show how they simplify the programming of classical Web patterns. Let us consider a classical example already detailed in the literature, a timer example, which consists in a simple Web page defined by:

```
var elapsedTime = 0;

function doEverySecond() {
   elapsedTime++;
   document.getElementById( "curTime" )
           .innerHTML = elapsedTime; }

<html>
  <script>setInterval( doEverySecond, 1000 )</script>
  <button onclick="elapsedTime = 0">reset</button>
  <div id="curTime"></div>
</html>
```

Although simple and innocuous at first glance, this program suffers from two major problems. First, the lack of modularity. The function `doEverySecond`, that implements the timer, increments the wall clock *and* updates the UI (via `innterHTML` attribute assignment). Hence, it must be aware of all the elements that needs update. This is problematic as a UI may evolve over time with some elements removed and new elements added. Each evolution of the specification will then impact `doEverySecond` implementation. The second problem we address is the plumbing needed for extracting and modifying the `curTime` element. In the pure JavaScript this involves assigning and looking up unique identifiers (`curTime` identifier). The reactors and reactive nodes we propose solve these two problems.

```
<html> ~{
     const T = hop.reactProxy( { elapsedTime: 0 } );
     setInterval( () => { T.elapsedTime++ }, 1000 );
  }
  <button onclick=~{T.elapsedTime=0}>reset</button>
  <div><react>~{T.elapsedTime}</react></div>
```

```
</html>
```

This Hop.js program solves the two problems previously mentioned. It is modular as new reactive elements depending on the `elapsedTime` can be added without modifying existing code. It avoids tedious surgery of the HTML DOM as the `react` node designates the node that need updates and its positioning in the UI.

We have built a first operational prototypes of reactors and reactive nodes. This work will be pursued in 2017. We will complete the implementation in Hop.js by including them in Hop-3.1.0. We will write a scientific paper describing their design and implementation.

5.1.2. Hiphop.js

Modern Web applications are rich in interactions between users and servers. Those interactions are from different nature: search and play music, book train or airplane tickets, query database or use an interactive map. From the programmer point-of-view, those interactions are handled by asynchronous events from multiple sources. Management of those events, which is called orchestration, is done by using event handlers. It is a mechanism that will call a specific function when a specific event raises. This kind of orchestration doesn't scale well since the behavior of the application has to be deduced by the programmer. Synchronous languages like Esterel, which are used in the industrial area, provides syntactic constructs that allow ordering the temporal behavior of the application. Then, reading the program source gives a precise idea of the behavior of the program at runtime.

The HipHop.js contribution is to adapt the reactive constructs of Esterel to the Web. The goal is to design a high-level tool that simplifies the orchestration of Web applications. In the traditional Esterel setting, the reactive program is written in a different source file of the host program. It is compiled independently of the host program. Therefore, the programmer must make explicit bindings between the reactive program and the host program in order to allow both of them to interact. This is inadequate for Web developments. So, HipHop.js adopts a radically different point of view: the reactive program is written in the same source code with the host program and the interaction between the reactive program and the host program is direct, thanks to a JavaScript API which is offered by the compilation output of the reactive program. HipHop.js uses a XML syntax, where each node corresponds to an Esterel instruction. This syntax has pros and cons but we think its advantages dominate. First, it is familiar to all Web developers, which do not have to learn a new syntax. Second, it is overly simple to implement as Hop.js natively supports XML parsing. Third, it gives macros for free as the XML syntax can be mixed with standard JavaScript that can create and return XML objects.

The classical Esterel example of the synchronous community is "ABRO": a program which is waiting for two events in parallel. When both events are raised, the host program is notified (here it pops a window up). At any moment, the reactive program state can be reset, in which case, the reactive program waits again for both events. For the sake of illustration, we show here how to implement ABRO in HipHop.js inside a Web page:

```
<html> ~{
     var abro =
         <hh.module A B R O>
           <hh.loopeach R>
             <hh.parallel>
                 <hh.await A/>
                 <hh.await B/>
             </hh.parallel>
             <hh.atom apply=${function() {alert("ABRO")}}/>
           </hh.loopeach>
         </hh.module>

     var m = new hh.ReactiveMachine(abro);
  }
  <button onclick=~{m.inputAndReact("A")}>A</button>
  <button onclick=~{m.inputAndReact("B")}>B</button>
```

```
  <button onclick=~{m.inputAndReact("R")}>R</button>
</html>
```

Pushing the buttons “A” and “B” triggers the popup message which contains "ABRO" in the browser page. In spite of its simplicity, the ABRO example is representative of a wide class of real programs. For instance, a program behaviorally similar to ABRO can be used to download a file in several parts of different sources, and merge them when all downloads are completed.

The first HipHop.js version has been released this year. It is available at the following URL http://www-sop.inria.fr/members/Colin.Vidal/hiphop/.

5.1.3. Garbage Collection with non ambiguous roots

Hop uses lot of objets with short time life.

Some Hop programs allocate many temporary objects whose lifetimes are very short. These objects are unefficiently handled by this *Mark&Sweep* garbage collector that Hop currently uses. We expect a speed-up by switching from a *Mark&Sweep* garbage collector to a generational *Stop&Copy* one. *Stop&Copy* collectors demand that all roots of the accessibility graph have to be precisely known (non ambiguous root). We have changed the code generation of the compiler in order to maintain a precise map of the pointers living in the stack.

5.1.4. Event calculus

We have studied functions over streams of events (timed values) and more precisely those which have a temporal causality property: at every instant, current outputs only depends on inputs that have already been received [24]. We have found a clear characterization of causal functions and made some proofs with the Coq system [21].

5.2. Privacy

Participant: Nataliia Bielova.

5.2.1. Hybrid Monitoring of Attacker knowledge

Enforcement of noninterference requires proving that an attacker’s knowledge about the initial state remains the same after observing a program’s public output. We have proposed a hybrid monitoring mechanism which dynamically evaluates the knowledge that is contained in program variables [14]. To get a precise estimate of the knowledge, the monitor statically analyses non-executed branches. We show that our knowledge-based monitor can be combined with existing dynamic monitors for non-interference. A distinguishing feature of such a combination is that the combined monitor is provably more permissive than each mechanism taken separately. We demonstrate this by proposing a knowledge-enhanced version of a no-sensitive-upgrade (NSU) monitor. The monitor and its static analysis have been formalized and proved correct within the Coq proof assistant.

5.3. Security

Participants: Nataliia Bielova, Ilaria Castellani, Tamara Rezk, Dolière Francis Some.

5.3.1. Security for multiparty session calculi

In our previous work, we investigated two security properties for multiparty session calculi: *access control* and *information flow security*. We proposed a type system ensuring both these properties. We also defined a monitored semantics inducing a property that is strictly included between typability and information flow security, which we called *information flow safety*.

The article [5] is an extended version of a previous workshop paper, which introduces refined versions of the safety and security properties examined in that paper and provides two additional results: compositionality of the refined safety property, and the proof that this property is ensured by a simplified version of the type system of [4].

In [18], we argue that the security requirements considered in previous work could be overly restrictive in some cases. In particular, a party is not allowed to communicate any kind of public information after receiving a secret information. The aim of [18] is to overcome this restriction, by proposing a new type discipline for a multiparty session calculus, which classifies messages according to their topics and allows unrestricted sequencing of messages on independent topics.

5.3.2. Security for dynamic and adaptable systems

We have started to study security issues in the context of dynamically evolving communicating systems, namely systems which are able to adapt themselves in reaction to particular events, arising in the system itself or in its environment. When focussing on security, examples of such events are security attacks or changes in security policies.

The paper [11] investigates a simple session calculus in which self-adaptation and security concerns may be jointly addressed. In this calculus, security violations occur when processes attempt to read or write messages of inappropriate security level within a session. Such violations trigger adaptation mechanisms that prevent the violations to propagate their effect in the remainder of the session, while allowing the computation to proceed. More specifically, our calculus is equipped with a monitored semantics based on session types, which activates local and global adaptation mechanisms for reacting respectively to soft and hard security violations. We present type soundness results that ensure that the overall protocol is still correctly executed after the application of these mechanisms.

5.3.3. Information Flow Monitoring

The dynamic aspects of JavaScript make the security analysis of web applications very challenging. Purely static analysis is prohibitively restrictive in practice since it must exclude JavaScript dynamic aspects or over-approximate them. In recent years, several dynamic enforcement mechanisms in the form of information flow monitors have been proposed. In order to better evaluate the currently available information flow monitors trade-offs, our contribution is to rigorously compare them [16]. We compare them with respect to two important dimensions according to the runtime monitor literature: soundness and transparency. We analyse five widely explored information flow monitor techniques: no-sensitive-upgrade, permissive-upgrade, hybrid monitors, secure multi execution, and multiple facets. Furthermore, we formally prove that the generalised belief in the equivalence of two of these approaches, secure multi-execution and multiple facets, is false [17].

5.3.4. Quantitative information flow measures

A number of measures for quantifying information leakage of a program have been proposed. Most of these measures evaluate a program *as a whole* by quantifying how much information can be leaked *on average* by different program outputs. While these measures perfectly fit for static program analyses, they cannot be used by dynamic analyses since they do not specify what information an attacker learns through observing one concrete program output.

In this work, we study the existing definitions of quantitative information flow [15]. Our goal is to find the definition of *dynamic leakage* – it should evaluate how much information an attacker learns when she observes *one program output*. Surprisingly, we find out that none of the existing definitions provide a suitable measure for dynamic leakage. We hence open a new research question in quantitative information flow area: which definition of dynamic leakage is suitable?

5.3.5. Access control and capability systems

Motivated by the problem of understanding the difference between practical access control and capability systems formally, we distill the essence of both in a language-based setting [19]. We first prove that access control systems and (object) capabilities are fundamentally different. We further study capabilities as an enforcement mechanism for confused deputy attacks (CDAs), since CDAs may have been the primary motivation for the invention of capabilities. To do this, we develop the first formal characterization of CDA-freedom in a language-based setting and describe its relation to standard information flow integrity. We show that, perhaps suprisingly, capabilities cannot prevent all CDAs. Next, we stipulate restrictions on programs

under which capabilities ensure CDA- freedom and prove that the restrictions are sufficient. To relax those restrictions, we examine provenance semantics as sound CDA-freedom enforcement mechanisms.

PHOENIX Project-Team

7. New Results

7.1. Tablet-Based Activity Schedule in Mainstream Environment for Children with Autism and Children with ID

Including children with autism spectrum disorders (ASD) in mainstream environments creates a need for new interventions whose efficacy must be assessed in situ. This article presents a tablet-based application for activity schedules that has been designed following a participatory design approach involving mainstream teachers, special education teachers, and school aides. This application addresses two domains of activities: classroom routines and verbal communications. We assessed the efficiency of our application with two overlapping user studies in mainstream inclusion, sharing a group of children with ASD. The first experiment involved 10 children with ASD, where five children were equipped with our tabled-based application and five were not equipped. We show that (1) the use of the application is rapidly self-initiated (after 2 months for almost all the participants) and (2) the tablet-supported routines are better performed after 3 months of intervention. The second experiment involved 10 children equipped with our application; it shared the data collected for the five children with ASD and compared them with data collected for five children with intellectual disability (ID). We show that (1) children with ID are not autonomous in the use of the application at the end of the intervention, (2) both groups exhibited the same benefits on classroom routines, and (3) children with ID improve significantly less their performance on verbal communication routines. These results are discussed in relation with our design principles. Importantly, the inclusion of a group with another neurodevelopmental condition provided insights about the applicability of these principles beyond the target population of children with ASD.

7.2. Self Determination-Based Design To Achieve Acceptance of Assisted Living Technologies For Older Adults

Providing technological support to assist older adults in their daily activities is a promising approach to aging in place. However, acceptance is critical when technologies are embedded in the user's life. Recently, Lee et al. established a connection between acceptance and motivation. They approached motivation via the Self-Determination Theory (SDT): the capacity to make choices and to take decisions. This paper leverages SDT to promote a new design style for gerontechnologies that consists of principles and requirements. We applied our approach to develop an assisted living platform, which was used to conduct a six-month field study with 34 older adults. We show that self-determination is a determining factor of technology acceptance. Furthermore, our platform improved the self-determination of equipped participants, compared to the control group, suggesting that our approach is effective. As such, SDT opens up new opportunities for improving the design process of gerontechnologies.

7.3. Frameworks compiled from declarations: a language-independent approach

Programming frameworks are an accepted fixture in the object-oriented world, motivated by the need for code reuse, developer guidance, and restriction. A new trend is emerging where frameworks require domain experts to provide declarations using a domain-specific language (DSL), influencing the structure and behaviour of the resulting application. These mechanisms address concerns such as user privacy. Although many popular open platforms such as Android are based on declaration-driven frameworks, current implementations provide ad hoc and narrow solutions to concerns raised by their openness to non-certified developers. Most widely used frameworks fail to address serious privacy leaks, and provide the user with little insight into application behaviour. To address these shortcomings, we show that declaration-driven frameworks can limit privacy

leaks, as well as guide developers, independently from the underlying programming paradigm. To do so, we identify concepts that underlie declaration-driven frameworks, and apply them systematically to both an object-oriented language, Java, and a dynamic functional language, Racket. The resulting programming framework generators are used to develop a prototype mobile application, illustrating how we mitigate a common class of privacy leaks. Finally, we explore the possible design choices and propose development principles for developing domain-specific language compilers to produce frameworks, applicable across a spectrum of programming paradigms.

7.4. Analysis of How People with Intellectual Disabilities Organize Information Using Computerized Guidance

Access to residential settings for people with intellectual disabilities (ID) contributes to their social participation, but presents particular challenges. Assistive technologies can help people perform activities of daily living. However, the majority of the computerized solutions offered use guidance modes with a fixed, unchanging sequencing that leaves little room for self-determination to emerge. The objective of the project was to develop a flexible guidance mode and to test it with participants, to describe their information organization methods. This research used a descriptive exploratory design and conducted a comparison between five participants with ID and five participants with no ID. The results showed a difference in the information organization methods for both categories of participants. The people with ID used more diversified organization methods (categorical, schematic, action-directed) than the neurotypical participants (visual, action-directed). These organization methods varied depending on the people, but also on the characteristics of the requested task. Furthermore, several people with ID presented difficulties when switching from virtual to real mode. These results demonstrate the importance of developing flexible guidance modes adapted to the users' cognitive strategies, to maximize their benefits. Studies using experimental designs will have to be conducted to determine the impacts of more-flexible guidance modes.

7.5. Leveraging Declarations over the Lifecycle of Large-Scale Sensor Applications

Masses of sensors and actuators are being deployed in our daily environments to provide innovative services for such spaces as parking lots, buildings, and railway networks. Yet, to realize the full potentials of these sensor network infrastructures, services need to be developed. Service development raises a number of challenges due to existing approaches that are often low level and network/hardware-centric. This paper proposes a high-level approach to the development of large-scale orchestrating applications. It revolves around a declaration language that allows to express the sensor-network dimensions of an application (sensor discovery, delivery models, actuation process). These declarations define the behavior of an application with respect to the sensor network infrastructure. We demonstrate the key relevance of these declarations at every stage of an application lifecycle, from design to runtime. In doing so, declarations allow to match the sensor-network behavior of an application to the target infrastructure. Our approach summarizes and puts in perspective our development of industrial case studies and our experience in using a commercially-operated sensor infrastructure.

7.6. Improving the Reliability of Pervasive Computing Applications By Continuous Checking of Sensor Readings

This paper shows that context-aware applications commonly make implicit assumptions about a sensor infrastructure. Because context-awareness critically relies on these assumptions, the developer typically need to ensure their validity by encoding them in the application code, polluting it with non-functional concerns. This defensive programming approach can be avoided by formulating these assumptions aside from the application, thus factorizing them as an explicit model of the sensor infrastructure. This model can be expressed as a set of rules and can be checked automatically and continuously to ensure the reliability of a sensor infrastructure, both at installation time and during normal functioning. The usefulness of our approach

is demonstrated in the domain of assisted living for seniors. We applied it to sensor data collected in the context of a 9-month field study of an assisted living platform, deployed at the home of 24 seniors. We show that several kinds of sensor malfunctions could have been identified upon their occurrence, thanks for our continuous checking, and resolved.

7.7. Designing Parallel Data Processing for Large-Scale Sensor Orchestration

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 paper 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.

RMOD Project-Team

7. New Results

7.1. Practical Validation of Bytecode to Bytecode JIT Compiler Dynamic Deoptimization.

Speculative inlining in just-in-time compilers enables many performance optimizations. However, it also introduces significant complexity. The compiler optimizations themselves, as well as the deoptimization mechanism are complex and error prone. To stabilize our bytecode to bytecode just-in-time compiler, we designed a new approach to validate the correctness of dynamic deoptimization. The approach consists of the symbolic execution of an optimized and an unop-timized bytecode compiled method side by side, deoptimizing the abstract stack at each deoptimization point (where dynamic deoptimization is possible) and comparing the deoptimized and unoptimized abstract stack to detect bugs. The implementation of our approach generated tests for several hundred thousands of methods, which are now available to be run automatically after each commit [13].

7.2. Recording and Replaying System-Specific Conventions.

In other situations, we found that developers sometimes perform sequences of code changes in a systematic way. These sequences consist of small code changes (*e.g.*, create a class, then extract a method to this class), which are applied to groups of related code entities (*e.g.*, some of the methods of a class). We propose to help this task by letting the developer record the sequence of code changes when he first applies it, and then generalize this sequence to apply it in other locations. The evaluation is based on real instances of such sequences that we found in different open source systems. We were able to replay 92% of the examples, which consisted in up to seven code entities modified up to 66 times. We are still working on the approach to allow for (semi-)automatic generalization of the recorded sequence of changes [71], [70].

7.3. Test Case Selection in Industry: an Analysis of Issues Related to Static Approaches

Automatic testing constitutes an important part of everyday development practice. But running all these tests may take hours. This is especially true for large systems involving, for example, the deployment of a web server or communication with a database. For this reason, tests are not launched as often as they should be and are mostly run at night. The company wishes to improve its development and testing process by giving to developers rapid feedback after a change. An interesting solution to give developers rapid feedback after a change is to reduce the number of tests to run by identifying only those exercising the piece of code changed. Two main approaches are proposed in the literature: static and dynamic. We evaluate these approaches on three industrial, closed source, cases to understand the strengths and weaknesses of each solution. We also propose a classification of problems that may arise when trying to identify the tests that cover a method.

TACOMA Team

6. New Results

6.1. RFID for pervasive computing environments

Participants: Nebil Ben Mabrouk, Frédéric Weis, Paul Couderc [contact].

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

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

In the context of RFID system, the resiliency property of such data structures enables new information architecture and autonomous (offline) operation, which is very important for some RFID applications. We previously applied the self-describing objects approach to the waste management domain, which has shown to be a specially challenging situation for RFID. This challenge is found more generally in pervasive computing scenarios involving RFID reading in uncontrolled environments (see section 4.4).

We achieved the following results:

- We showed the importance of diversity in the context of challenging RFID reading. A reconfigurable antenna was designed to support dynamic reading protocols.
- A software approach based on error correcting code was developed to support robust data storage in groups of RFID.
- An innovative RFID testbed for experimenting a large range of RFID situations/applications was operational (minus some features to be completed), supported by a simulation environment and a control environment.
- A patent was filed and some contacts made with RFID companies.

However, the supports for implementing dynamic reading protocols were lacking, both on the software and the radio side. The following further progress were made:

- The four elements diversity antenna designed in first phase was implemented.
- The control software has been greatly improved. A new environment was designed, offering powerful and flexible programming capabilities for easy prototyping of RFID reading scenarios and collecting experiments results. A simulator of the testbed was also developed, allowing off-site developments. This work is supported by the RFID-Lab ADT.
- Motion-induced improvements of RFID reliability were experimented, as shown below in Figure 4 .
- A significant dissemination efforts toward the industry was made, and we have good hope that some of the contacts will lead to perspectives.

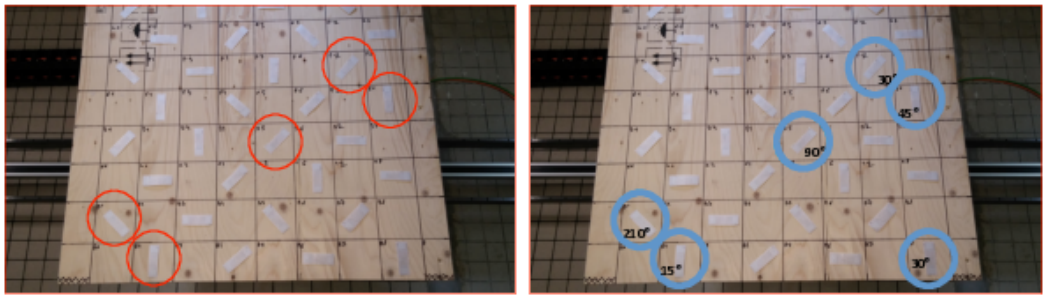

Figure 4. (a) initial read, 20% of the tags are missing (b) After 210 deg of rotation, all the tags are recovered

An example of motion-assisted RFID readings implemented is shown in figure 4 : a matrix of 32 RFID tags are arranged in reduced power conditions, so that the tags are near their sensivity limit. In such conditions, 20% of the tags failed to be read by the reader. By coupling the reading with a rotation of 210 deg, we show that all the missing tags are progressively recovered.

6.2. Building an extensible information sharing mechanism

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

Context aware applications adapt their behavior based on information they can collect on their surrounding environment. Most of these data are provided by third-party software, sensors or computed by the application itself. A striking challenge facing the building of comprehensive pervasive system is the lack of integration between the different services provided by third parties. In this project, we intend to study and to provide mechanisms to enhance information sharing between applications and more specifically to augment information on the surrounding environment. The idea is to endow applications with the capability to increase or augment information on the physical world they are interacting with and to retrieve and share these data seamlessly depending on their location. Such mechanism aims at providing a complementary source of information in order to improve the process of choosing the best service/information provider and to help them keeping additional information on physical resources such as environment specific configuration (e.g., calibration data).

The idea of augmenting information on the physical world is not new. It has been done for centuries in the real world. For instance, the Little Thumb sowed pebbles to find his way just as hikers use cairns so as not to get lost. In daily life, people use sticky notes on pieces of hardware or objects to keep relevant information on their use or capabilities. Applied to IT, such ideas have been pushed by the augmented reality domain where users can access a personalized view of the real world that helps them to carry out their activities. Although this idea has already been implemented in some ad hoc solutions (to exchange ratings for instance), we aim to provide a more generic solution. Our solution must be applicable to nowadays devices and applications with little adjustment 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 tends to think that their âstandardâ are 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.

We are currently implementing these ideas by designing and experimenting two architectures/prototypes:

- **Matriona** is a global 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 comprehensive framework for exposing devices as REST-like resources. Resources functionalities can be extended through the mean of decorators. The system also provides access control mechanisms. The main interest of matriona concerning the information enrichment is its ability to support dynamic extension of resource meta-information by application and to provide means to share these meta-information with others. It implements the concept of group of interest with access control on meta-information. The concepts described in Matriona are in the process to be published.
- **Little Thumb Registry (LTReg)** is an independent resource registry that provides the same enrichment capabilities than Matriona but impose less constraints on the architecture of application. Although the prototype is operational, Matriona does not comply with the principle advocated herebefore: it supposes the use of a pivot technology (REST) and assumes that application developers will develop their application on top of on OSGi based platform. The idea behind LTReg is to decouple the registry from the framework and to provide a registry in the manner of Consul [0] with meta-information enrichment and sharing mechanisms. By focussing only on the discovery mechanism and information sharing, LTReg imposes fewer constraints on application and comply more with the goal of being ready to use in actual application. This is still a work in progress.

6.3. Modeling activities to promote self-consumption of locally produced energy

Participants: Jean-Marie Bonnin, Alexandre Rio, Yoann Maurel [contact].

Traditional electricity distribution schemes decouple the production sources from the consumers so that it is necessary to transport energy over long distances. This type of organization is illustrated by the consumption of region such as Brittany, where 91% of the energy consumed is imported. It induces inherent inefficiencies due to the line losses and the transformation steps and therefore induces a high infrastructure and distribution cost. To face these problems and in order to reduce the environmental impacts associated with the use of energy, recent years have seen the development of initiatives to produce energy locally.

The sources of renewable energies are good candidates for this because they are varied and adapt easily to the different geographical situations. The infrastructures necessary for their implementation also impose fewer constraints in terms of installation and safety. One of the main obstacles to the unique use of these technologies comes from their strong dependence on physical and meteorological characteristics, which makes it more difficult to foresee production capacities. These characteristics vary from one facility to another and from one region to another. The combined use of these technologies therefore appears to be necessary to ensure that there will always be available energy at the lowest possible cost. In this context, OKWind proposes to deploy self-production units directly where the consumption is done and has developed expertise in multi-source energy production (see section 8.1).

In 2016, we started to study a solution favoring maximum autonomy of the instrumented sites from the traditional channels energy production by modeling business processes and using learning algorithms to shift demanding activities according to local production capacities. For example, the system should be able to anticipate a potential consumption of hot water (and thus of the energy needed for its production) in order to produce it at the best time when the renewable energy is available. It should also choose the best storage solution for this energy: hot water could be directly stored by the heat pump for instance. The system must implement policies that will intelligently shift demanding activities according to the predictions of energy production. It thus requires:

- **capabilities to predict the production of energy**. A lot of theoretical work has been done in the literature to predict the production of renewable sources of energy. In addition, in order to

[0] https://www.consul.io/

evaluate the production of energy and its consumption over time, OKWind has developed data retrieval mechanisms on each deployed sites. They produce accurate statistics on production and consumption. Both approaches should be used as inputs of our decision processes and model. One of our goals is to evaluate the precision of the theoretical prediction models against these real-world data to determine which are the most relevant for the implementation of our approach.

- **capabilities to model the consumption on energy**. Numerous works of the literature are interested in similar problems but focuses mainly on building electricity consumption model of machine tools [10]. We propose to focus instead on activity and business processes. In a related domain, modeling work has been conducted on water consumption of farms [7]. The objective was to predict the water consumption of an operating farm by modeling business processes. Our goal is to propose a similar model for electricity targeting a broader scope of economic sectors.
- **capabilities to schedule activities in order to match production and consumption so as to promote self-consumption.** This requires developing control loop that will proactively analyze and predict consumption and take measure to shift demand. This can, in a first approach, be done by assisting the consumers and providing them guidance on when to perform certain tasks. Assisted demand shifting have already been developed for the residential domain [6] but this project focused on uses mainly and little on the modeling of business processes. Ultimately, we would like to develop automated process transparently when possible. The learning algorithms will be developed in collaboration with Ubiant [0], a company specialized in artificial intelligence to smart-buildings.

To validate the approach and to understand business processes, we have started a field study targeting two types of activities (e.g. farm or hotel). We also want to develop tools to simulate a site so that we can quickly evaluate our policies over simulated long periods of time.

6.4. Definition of a Smart Energy Aware architecture

Participant: Jean-Marie Bonnin [contact].

In the past years, energy demand has increased and shifted especially towards electricity as the form of consuming energy. As the number of electric devices constantly grows and energy production must increasingly rely on renewable sources, this leads into noteworthy disparity between electricity production and consumption. Within the ITEA2 12004 Smart Energy Aware Systems (SEAS) project (see section 1), we proposed the "SEAS Reference Architecture Model" (S-RAM). This architecture relies on four distributed services that enable to interconnect any energy actors and give them the opportunity to provide new energy services. The benefits of S-RAM have been studied on a specific use case, which aims to provide a service for estimating local photovoltaic production. It particularly helps energy management systems better plan electric consumption. 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 continue to develop the components of the architecture that will be demonstrated in the final review of the project.

6.5. Context modeling for Smart Spaces

Participants: Yoann Maurel, Frédéric Weis [contact].

To provide services for Smart Building, automation based on pre-set scenarios is ineffective: human behavior is hardly predictable and application should be able to adapt their behavior at runtime depending on the context. We focused on recognizing user's activities to adapt applications behaviors. Our aim is to compute small pieces of context we called *context attributes*. Those context attributes are diverse, for example a presence in a room, the number of people in a room etc. Building efficient and accurate context information using inexpensive and non-invasive sensors was and is still a great challenge. We proved, through the use of dedicated algorithms and a layered architecture that it is achievable when the targeted space (controlled environment) is known - due to the specific and non automated calibration process we used. Among all the available theories, we used the Belief Function Theory (BFT) [8] [9] as it allows to express **uncertainty** and **imprecision**.

[0] https://www.ubiant.com/en/about/

Context is computed by a chain if three a tasks:

- The transition between a raw sensor value and a belief function is made through the use of a belief model which maps a sensor value to a belief function. A belief function represents the degree of belief associated to each possible value of the context attribute.
- Then a set of belief functions (corresponding to a set of sensors) can be combined (fused).
- Finally the system can decide what is the "best" value for the context attribute.

Currently the BFT theories requires a huge calibration process. In 2016, we obtained new results on the semi-automated building of belief functions, that have to be provided by each sensor, using our BFT Java implementation (see section 5.1).

6.6. Towards Metamorphic Housing: the on-demand room

Participants: Frédéric Weis, Michele Dominici [contact].

6.6.1. A concrete example of Metamorphic Housing: the on-demand room

The research activities related to the research program on Metamorphic Housing mainly focused on defining the detailed architecture and functionalities of the selected case study, the on-demand room. We conducted an iterative co-design process, involving the partners of the chair "Habitat Intelligent et Innovation". Valuable input was also obtained by collaborating with Delta Dore, LOUSTIC, Université de Bretagne Occidentale, etc. The result was the identification of the needs of end users, building owners and managers with respect to the on-demand room. To satisfy these requirements, we proposed a system architecture, combining computer and mobile applications with domotic equipments and novel interaction means for end users.

These are inspired by the Pervasive Computing and Interactive Architecture principles, where a continuous and implicit interaction between occupants and the physical world is made possible by augmented architectural structures, which sense the natural actions of people and respond accordingly. In this way, the occupants of the dwellings equipped with on-demand room experience a new form of housing, stimulating social interactions between neighbors and satisfying periodic needs of additional housing surface, as we illustrated in [4]. We submitted our system architecture, novel interaction means and augmented structure designs to the industrial property services of Inria and University of Rennes 1, which are currently evaluating the possibility of establishing patent protection on these inventions.

6.6.2. Experimentation of Metamorphic Housing on social housing

We helped Néotoa, a social landlord, preparing and initiating an experimentation of the on-demand room on one of their residential buildings. For this, we built and coordinated a consortium of partners working on the project: Veolia, CCI Rennes, Cardinal Edifice, Rennes Métropole, Néotoa, Delta Dore, LOUSTIC, Université de Bretagne Occidentale, MobBI platform (University of Rennes 1), Inria, Institut de Gestion de Rennes. We took a user-centered approach to the problem, studying it from several points of view and mobilizing several disciplines: psychology and ergonomics (LOUSTIC), sociology (Université de Bretagne Occidentale), marketing (Institut de Gestion de Rennes). We conducted user interviews, initially leveraging the demonstrator of the on-demand room that we previously built via the Immersia virtual reality platform. Then, we ran on-line inquiries to reach a larger audience. We took into account the lessons that we learned in the design and development of a computing and domotic system, leveraging the expertise of valuable partners (Delta Dore, MobBI platform, Inria), as detailed in section 5.3 .

COATI Project-Team

7. New Results

7.1. Network Design and Management

Participants: Jean-Claude Bermond, Christelle Caillouet, David Coudert, Frédéric Giroire, Nicolas Huin, Joanna Moulierac, Stéphane Pérennes.

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 three topics: firstly Fixed wireless Backhaul Networks, with the objective of achieving a high reliability of the network. Secondly, Software-Defined networks, in which 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. Finally, the third topic investigated is Energy Efficiency within Backbone networks and for content distribution. We focus on Redudancy Elimination, and we use SDN as a tool to turn-off the links in real networks. We validated our algorithms on a real SDN platform [0].

7.1.1. Fault tolerance

7.1.1.1. Survivability in networks with groups of correlated failures

The notion of Shared Risk Link Groups (SRLG) captures survivability issues when a set of links of a network may fail simultaneously. The theory of survivable network design relies on basic combinatorial objects that are rather easy to compute in the classical graph models: shortest paths, minimum cuts, or pairs of disjoint paths. In the SRLG context, the optimization criterion for these objects is no longer the number of edges they use, but the number of SRLGs involved. Unfortunately, computing these combinatorial objects is NP-hard and hard to approximate with this objective in general. Nevertheless some objects can be computed in polynomial time when the SRLGs satisfy certain structural properties of locality which correspond to practical ones, namely the star property (all links affected by a given SRLG are incident to a unique node) and the span 1 property (the links affected by a given SRLG form a connected component of the network). The star property is defined in a multi-colored model where a link can be affected by several SRLGs while the span property is defined only in a mono-colored model where a link can be affected by at most one SRLG. We have extended in [23] these notions to characterize new cases in which these optimization problems can be solved in polynomial time. We have also investigated the computational impact of the transformation from the multi-colored model to the mono-colored one. Reported experimental results validate the proposed algorithms and principles.

7.1.1.2. Reliability of fixed wireless backaul 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. In [48], we provide 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 each leaf in the tree requires a valid routing for the network. Any such tree provides an upper and lower bound on the reliability, and the algorithm improves these bounds by branching

[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.

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. We perform a computational study demonstrating that the proposed methods can calculate reliability of large backhaul networks, and can optimize topology for modest size networks.

7.1.1.3. Fault tolerance of Linear Access Network

In [52], 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, we analyze the probability for a user to experience a disconnection longer than a threshold t_*, leading to a disruption of all on-going communications between the vehicle and the infrastructure network. We provide an approximation formula to estimate this probability for large networks. We then carry out a sensitivity analysis and supply a guide for operators when choosing the parameters of the networks. We focus on two scenarios: an intercity bus and an intercity train. 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. Routing in 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.

7.1.2.1. MINNIE: an SDN World with Few Compressed Forwarding Rules

While SDN brings flexibility in 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. In [34], [54], we investigate 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 entry 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 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 ASICs. Hence, both simulations and experimental results suggest that MINNIE can be safely deployed in real networks, providing compression ratios between 70% and 99%.

7.1.2.2. Energy-Aware Routing in Software-Defined Networks

In [51], we focus on using SDN for energy-aware routing (EAR). Since traffic load has a small influence on power consumption of routers, EAR allows to put unused devices into sleep mode to save energy. SDN can collect traffic matrix and then computes routing solutions satisfying QoS while being minimal in energy consumption. However, prior works on EAR have assumed that the forwarding table of OpenFlow switch can hold an infinite number of rules. In practice, this assumption does not hold since such flow tables are implemented in Ternary Content Addressable Memory (TCAM) which is expensive and power-hungry. We consider the use of wildcard rules to compress the forwarding tables. In this paper, we propose optimization methods to minimize energy consumption for a backbone network while respecting capacity constraints on links and rule space constraints on routers. In details, we present two exact formulations using Integer Linear Program (ILP) and introduce efficient heuristic algorithms. Based on simulations on realistic network topologies, we show that, using this smart rule space allocation, it is possible to save almost as much power consumption as the classical EAR approach

7.1.3. Reducing Networks' Energy Consumption

Due to the increasing impact of ICT (Information and Communication Technology) on power consumption and worldwide gas emissions, energy efficient ways to design and operate backbone networks are becoming a new concern for network operators. Recently, energy-aware routing (EAR) has gained an increasing popularity in the networking research community. The idea is that traffic demands are redirected over a subset of the network devices, allowing other devices to sleep to save energy. We studied variant of this problems.

7.1.3.1. Energy efficient Content Distribution

To optimize energy efficiency in network, operators try to switch off as many network devices as possible. Recently, there is a trend to introduce content caches as an inherent capacity of network equipment, with the objective of improving the efficiency of content distribution and reducing network congestion. In [36], we study the impact of using in-network caches and CDN cooperation on an energy-efficient routing. We formulate this problem as Energy Efficient Content Distribution. The objective is to find a feasible routing, so that the total energy consumption of the network is minimized subject to satisfying all the demands and link capacity. We exhibit the range of parameters (size of caches, popularity of content, demand intensity, etc.) for which caches are useful. Experiment results show that by placing a cache on each backbone router to store the most popular content, along with well choosing the best content provider server for each demand to a CDN, we can save a total up to 23% of power in the backbone, while 16% can be gained solely thanks to caches.

7.1.3.2. Energy-Efficient Service Function Chain Provisioning

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. This is of primary interest to implement energy efficient solutions, which imply to adapt dynamically the resource usage to the demands. In [53], [55], 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.

7.1.4. Other results

7.1.4.1. Well Balanced design for Data placement

The have considered in [17] a problem motivated by data placement, in particular data replication in distributed storage and retrieval systems. We are given a set V of v servers along with b files (data, documents). Each file is replicated on exactly k servers. A placement consists in finding a family of b subsets of V (representing the files) called blocks, each of size k. Each server has some probability to fail and we want to find a placement which minimizes the variance of the number of available files. It was conjectured that there always exists an optimal placement (with variance better than that of any other placement for any value of the probability of failure). We show that the conjecture is true, if there exists a well balanced design, that is a family of blocks, each of size k, such that each j-element subset of V , $1 \leq j \leq k$, belongs to the same or almost the same number of blocks (difference at most one). The existence of well balanced designs is a difficult problem as it contains as a subproblem the existence of Steiner systems. We completely solve the case $k = 2$ and give bounds and constructions for $k = 3$ and some values of v and b.

7.1.4.2. Study of Repair Protocols for Live Video Streaming Distributed Systems

In [33], we study distributed systems for live video streaming. These systems can be of two types: structured and un-structured. In an unstructured system, the diffusion is done opportunistically. The advantage is that it handles churn, that is the arrival and departure of users, which is very high in live streaming systems, in a smooth way. On the opposite, in a structured system, the diffusion of the video is done using explicit diffusion trees. The advantage is that the diffusion is very efficient, but the structure is broken by the churn. In this paper, we propose simple distributed repair protocols to maintain, under churn, the diffusion tree of

a structured streaming system. We study these protocols using formal analysis and simulation. In particular, we provide an estimation of the system metrics, bandwidth usage, delay, or number of interruptions of the streaming. Our work shows that structured streaming systems can be efficient and resistant to churn.

7.1.4.3. Gathering in radio networks

In [16], we consider the problem of gathering information in a gateway in a radio mesh access network. Due to interferences, calls (transmissions) cannot be performed simultaneously. This leads us to define a round as a set of non-interfering calls. Following the work of Klasing, Morales and Pérennes, we model the problem as a Round Weighting Problem (RWP) in which the objective is to minimize the overall period of non-interfering calls activations (total number of rounds) providing enough capacity to satisfy the throughput demand of the nodes. We develop tools to obtain lower and upper bounds for general graphs. Then, more precise results are obtained considering a symmetric interference model based on distance of graphs, called the distance-d interference model (the particular case d = 1 corresponds to the primary node model). We apply the presented tools to get lower bounds for grids with the gateway either in the middle or in the corner. We obtain upper bounds which in most of the cases match the lower bounds, using strategies that either route the demand of a single node or route simul- taneously flow from several source nodes. Therefore, we obtain exact and constructive results for grids, in particular for the case of uniform demands answering a problem asked by Klasing, Morales and Pérennes.

7.2. Graph Algorithms

Participants: Jean-Claude Bermond, Nathann Cohen, David Coudert, Guillaume Ducoffe, Frédéric Giroire, Nicolas Nisse, Stéphane Pérennes.

COATI is also 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. In particular, we try to find the most efficient algorithms to solve the problems, sometimes focusing on specific graph classes from which the problems are polynomial-time solvable. Many results introduced here are presented in detail in the PhD thesis of Guillaume Ducoffe on *Metric properties of large graphs* https://team.inria.fr/coati/phd-defense-of-guillaume-ducoffe/.

7.2.1. Graph decompositions

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.

7.2.1.1. Width parameters of graphs

In [22], we design a Branch and Bound algorithm that computes the exact pathwidth of graphs and a corresponding path-decomposition. Our main contribution consists of several non-trivial techniques to reduce the size of the input graph (pre-processing) and to cut the exploration space during the search phase of the algorithm. We evaluate experimentally our algorithm by comparing it to existing algorithms of the literature. It appears from the simulations that our algorithm offers a significative gain with respect to previous work. In particular, it is able to compute the exact pathwidth of any graph with less than 60 nodes in a reasonable running-time (10 min.). Moreover, our algorithm also achieves good performance when used as a heuristic (i.e., when returning best result found within bounded time-limit). Our algorithm is not restricted to undirected graphs since it actually computes the vertex-separation of digraphs (which coincides with the pathwidth in case of undirected graphs).

Many tree-decomposition-like parameters are related to particular layouts (ordering) of the vertices of the input graph. In [45], we present a new set of constraints for modeling linear ordering problems on graphs using Integer Linear Programming (ILP). These constraints express the membership of a vertex to a prefix rather than the exact position of a vertex in the ordering. We use these constraints to propose new ILP formulations for well-known linear ordering optimization problems, namely the Pathwidth, Cutwidth, Bandwidth, SumCut and Optimal Linear Arrangement problems. Our formulations are not only more compact than previous proposals, but also more efficient as shown by our experimental evaluations on large benchmark instances.

7.2.1.2. Metric properties of graph decompositions

The decomposition of graphs by clique-minimal separators is a common algorithmic tool, first introduced by Tarjan. Since it allows to cut a graph into smaller pieces, it can be applied to pre-process the graphs in the computation of many optimization problems. However, the best known clique-decomposition algorithms have respective $O(nm)$-time and $O(n^{2.69})$-time complexity, that is prohibitive for large graphs. Here we prove that for every graph G, the decomposition can be computed in $O(T(G) + \min\{n^{2.3729}, \omega^2 n\})$-time with $T(G)$ and ω being respectively the time needed to compute a minimal triangulation of G and the clique-number of G. In particular, it implies that every graph can be clique-decomposed in $O(n^{2.3729})$-time. Based on prior work from Kratsch et al., in [46], we prove in addition that computing the clique-decomposition is at least as hard as triangle detection. Therefore, the existence of any $o(n^{2.3729})$-time clique-decomposition algorithm would be a significant breakthrough in the field of algorithmic. Finally, our main result implies that planar graphs, bounded-treewidth graphs and bounded-degree graphs can be clique-decomposed in linear or quasi-linear time.

In [21], we establish general relationships between the topological properties of graphs and their metric properties. For this purpose, we upper-bound the diameter of the *minimal separators* in any graph by a function of their sizes. More precisely, we prove that, in any graph G, the diameter of any minimal separator S in G is at most $\lfloor \frac{\ell(G)}{2} \rfloor \cdot (|S| - 1)$ where $\ell(G)$ is the maximum length of an isometric cycle in G. We refine this bound in the case of graphs admitting a *distance preserving ordering* for which we prove that any minimal separator S has diameter at most $2(|S| - 1)$. Our proofs are mainly based on the property that the minimal separators in a graph G are connected in some power of G. Our result easily implies that the *treelengthtl*(G) of any graph G is at most $\lfloor \frac{\ell(G)}{2} \rfloor$ times its *treewidthtw*(G). In addition, we prove that, for any graph G that excludes an *apex graph*H as a minor, $tw(G) \leq c_H \cdot tl(G)$ for some constant c_H only depending on H. We refine this constant when G has bounded genus. As a consequence, we obtain a very simple $O(\ell(G))$-approximation algorithm for computing the treewidth of n-node m-edge graphs that exclude an apex graph as a minor in $O(nm)$-time.

In [32], [50], we study metric properties of the bags of tree-decompositions of graphs. Roughly, the length and the breadth of a tree-decomposition are the maximum diameter and radius of its bags respectively. The treelength and the treebreadth of a graph are the minimum length and breadth of its tree-decompositions respectively. Pathlength and pathbreadth are defined similarly for path-decompositions. In this paper, we answer open questions of [Dragan and Köhler , Algorithmica 2014] and [Dragan, Köhler and Leitert, SWAT 2014] about the computational complexity of treebreadth, pathbreadth and pathlength. Namely, we prove that computing these graph invariants is NP-hard. We further investigate graphs with treebreadth one, i.e., graphs that admit a tree-decomposition where each bag has a dominating vertex. We show that it is NP-complete to decide whether a graph belongs to this class. We then prove some structural properties of such graphs which allows us to design polynomial-time algorithms to decide whether a bipartite graph, resp., a planar graph, has treebreadth one.

7.2.2. 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.

Topologies for data center networks have been proposed in the literature through various graph classes and operations. A common trait to most existing designs is that they enhance the symmetric properties of the underlying graphs. Indeed, symmetry is a desirable property for interconnection networks because it minimizes congestion problems and it allows each entity to run the same routing protocol. However, despite sharing similarities these topologies all come with their own routing protocol. Recently, generic routing schemes have been introduced which can be implemented for any interconnection networks. The performances of such universal routing schemes are intimately related to the hyperbolicity of the topology. Motivated by the good performances in practice of these new routing schemes, we propose in [19], [29] the first general study of the hyperbolicity of data center interconnection networks. Our findings are disappointingly negative: we prove that the hyperbolicity of most data center topologies scales linearly with their diameter, that it the worst-case possible for hyperbolicity. To obtain these results, we introduce original connection between hyperbolicity and the properties of the endomorphism monoid of a graph. In particular, our results extend to all vertex and edge-transitive graphs. Additional results are obtained for de Bruijn and Kautz graphs, grid-like graphs and networks from the so-called Cayley model.

In [20], we investigate more specifically on the hyperbolicity of bipartite graphs. More precisely, given a bipartite graph $B = (V_0 \cup V_1, E)$ we prove it is enough to consider any one side V_i of the bipartition of B to obtain a close approximate of its hyperbolicity $\delta(B)$ — up to an additive constant 2. We obtain from this result the sharp bounds $\delta(G) - 1 \leq \delta(L(G)) \leq \delta(G) + 1$ and $\delta(G) - 1 \leq \delta(K(G)) \leq \delta(G) + 1$ for every graph G, with $L(G)$ and $K(G)$ being respectively the line graph and the clique graph of G. Finally, promising extensions of our techniques to a broader class of intersection graphs are discussed and illustrated with the case of the biclique graph $BK(G)$, for which we prove $(\delta(G) - 3)/2 \leq \delta(BK(G)) \leq (\delta(G) + 3)/2$.

7.2.3. Combinatorial games on graphs

We study several two-player games on graphs.

7.2.3.1. Games and graph decompositions

Graph Searching is a game where a team of searchers aims at capturing a fugitive in a graph. Graph Searching games have been widely studied because they are an algorithmic interpretation of tree/path-decompositions of graphs.

In [18], we define a new variant of graph searching, where searchers have to capture an invisible fugitive with the constraint that no two searchers can occupy the same node simultaneously. This variant seems promising for designing approximation algorithms for computing the pathwidth of graphs. The main contribution in [18] is the characterization of trees where k searchers are necessary and sufficient to win. Our characterization leads to a polynomial-time algorithm to compute the minimum number of searchers needed in trees.

We also study graph searching in directed graphs. We prove that the graph processing variant is monotone which allows us to show its equivalence with a particular digraph decomposition [25].

7.2.3.2. Distributed computing

We also investigate the games described above in a distributed setting.

Consider a set of mobile robots with minimal capabilities placed over distinct nodes of a discrete anonymous ring. Asynchronously, each robot takes a snapshot of the ring, determining 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. The computation also depends on the required task. In [24], 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 minimal assumptions.

7.2.3.3. Spy games in graphs

In [28], 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 node. 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 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). First, we consider the computational complexity of the problem, showing that it is NP-hard and that it is PSPACE-hard in DAGs. Then, we establish tight tradeoffs between the number k of guards and the required distance d when G is a path or a cycle. Our main result is that there exists $\beta > 0$ such that $\Omega(n^{1+\beta})$ guards are required to win in any $n \times n$ grid.

7.2.4. 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.

7.2.4.1. Bin packing

Motivated by an assignment problem arising in MapReduce computations , we investigate a generalization of the Bin Packing problem which we call Bin Packing with Colocations Problem [41]. Given a set V of items with positive integer weights, an underlying graph G = (V, E), and an integer q, the goal is to pack the items into a minimum number of bins so that (i) the total weight of the items packed in every bin is at most q, and (ii) for each edge (i, j) ∈ E there is at least one bin containing both items i and j. We first show that when the underlying graph is unweighted (i.e., all the items have equal weights), the problem is equivalent to the q-clique problem, and when furthermore the underlying graph is a clique, optimal solutions are obtained from covering designs. We prove that the problem becomes NP-hard even for weighted paths and un-weighted trees and we propose approximation algorithms for particular families of graphs, including: a $(3 + \sqrt{5})$-approximate algorithm for weighted complete graphs (improving a previous 8-approximation), a 2-approximate algorithm for weighted paths, a 5-approximate algorithm for weighted trees, and an (1+)-approximate algorithm for unweighted trees. For general weighted graphs, we propose a 3 + 2mad(G)/2-approximate algorithm, where mad(G) is the maximum average degree of G. Finally, we show how to convert any ρ-approximation algorithm for the Bin Packing (resp. the Densest q-Subgraph problem) into an approximation algorithm for the problem on weighted (resp. unweighted) general graphs.

7.2.4.2. distance preserving ordering

For every connected graph G, a subgraph H of G is isometric if for every two vertices $x, y \in V(H)$ there exists a shortest xy-path of G in H. A distance-preserving elimination ordering of G is a total ordering of its vertex-set $V(G)$, denoted $(v_1, v_2, \cdots, v_n)$, such that any subgraph $G - i = G \setminus \{v_1, v_2, \cdots, v_i\}$ with $1 \leq i < n$ is isometric. This kind of ordering has been introduced by Chepoi in his study on weakly modular graphs. In [47], we prove that it is NP-complete to decide whether such ordering exists for a given graph — even if it has diameter at most 2. Then, we describe a heuristic in order to compute a distance-preserving ordering when it exists one that we compare to an exact exponential algorithm and an ILP formulation for the problem. Lastly, we prove on the positive side that the problem of computing a distance-preserving ordering when it exists one is fixed-parameter-tractable in the treewidth.

7.2.4.3. cycle convexity

Many notions in graph convexity have been defined and studied for various applications, such as geode-tic convexity (generalizing the classical convexity in Euclidean space to graphs), monophonic convexity (to model spreading of rumor or disease in a network), etc. Each of the convexity notions led to the study of important graph invariants such as the hull number (minimum number of vertices whose hull set is the entire graph) or the interval number (minimum number of vertices whose interval is the whole graph). Recently, Araujo et al.

introduced the notion of Cycle Convexity of graphs for its application in Knot Theory. Roughly, the tunnel number of a knot embedded in a plane is equivalent to the hull number of a corresponding planar 4-regular graph in cycle convexity. In [35], we study the interval number of a graph in cycle convexity. Precisely, given a graph G, its interval number in cycle convexity, denoted by incc(G), is the minimum cardinality of a set $S \subseteq V(G)$ such that every vertex $w \in V(G) \setminus S$ has two distinct neighbors $u, v \in S$ such that u and v lie in same connected component of $G[S]$. In this work, first we provide bounds on incc(G) and its relations to other graph convexity parameters, and explore its behavior on grids. Then, we present some hardness results by showing that deciding whether $incc(G) \leq k$ is NP-complete, even if G is a split graph or a bounded-degree planar graph, and that the problem is W[1]-hard in bipartite graphs when k is the parameter. As a consequence, we obtain that it cannot be approximated up to a constant factor in the class of split graphs (unless P = N P). On the positive side, we present polynomial-time algorithms to compute incc(G) for outerplanar graphs, cobipartite graphs and interval graphs. We also present FPT algorithms to compute it for (q, q - 4)-graphs, where q is the parameter and for bounded treewidth graphs.

7.3. Graph theory

Participants: Nathann Cohen, Guillaume Ducoffe, Frédéric Havet, William Lochet, Nicolas Nisse.

Coati also 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 digraphs

7.3.1.1. Arc-disjoint branching flows

The concept of arc-disjoint flows in networks was introduced by Bang-Jensen and Bessy [Theoret. Comput. Science 526, 2014]. This is a very general framework within which many well-known and important problems can be formulated. In particular, the existence of arc-disjoint branching flows, that is, flows which send one unit of flow from a given source s to all other vertices, generalizes the concept of arc-disjoint out-branchings (spanning out-trees) in a digraph. A pair of out-branchings $B^+_{s,1}, B^+_{s,2}$ from a root s in a digraph $D = (V, A)$ on n vertices corresponds to arc-disjoint branching flows x_1, x_2 (the arcs carrying flow in x_i are those used in $B^+_{s,i}$, $i = 1, 2$) in the network that we obtain from D by giving all arcs capacity $n - 1$. It is then a natural question to ask how much we can lower the capacities on the arcs and still have, say, two arc-disjoint branching flows from the given root s. In [15], we prove that for every fixed integer $k \geq 2$ it is

- an NP-complete problem to decide whether a network $\mathcal{N} = (V, A, u)$ where $u_{ij} = k$ for every arc ij has two arc-disjoint branching flows rooted at s.
- a polynomial problem to decide whether a network $\mathcal{N} = (V, A, u)$ on n vertices and $u_{ij} = n - k$ for every arc ij has two arc-disjoint branching flows rooted at s.

The algorithm for the later result generalizes the polynomial-time algorithm, due to Lovász, for deciding whether a given input digraph has two arc-disjoint out-branchings rooted at a given vertex. Finally we prove that under the so-called Exponential Time Hypothesis (ETH), for every $\epsilon > 0$ and for every $k(n)$ with $(\log(n))^{1+\epsilon} \leq k(n) \leq \frac{n}{2}$ (and for every large i we have $k(n) = i$ for some n) there is no polynomial algorithm for deciding whether a given digraph contains two arc-disjoint branching flows from the same root so that no arc carries flow larger than $n - k(n)$.

7.3.1.2. Subdivision of oriented cycles

An *oriented cycle* is an orientation of a undirected cycle. In [43], [27], we first show that for any oriented cycle C, there are digraphs containing no subdivision of C (as a subdigraph) and arbitrarily large chromatic number. In contrast, we show that for any cycle C with two blocks, every strongly connected digraph with sufficiently large chromatic number contains a subdivision of C. This settles a conjecture of Addario-Berry

et al. [J. Combin. Theory B, 97, 2007]. More generally, we conjecture that this result holds for any oriented cycle. As a further evidence, we prove this conjecture for the antidirected cycle on four vertices (in which two vertices have out-degree 2 and two vertices have in-degree 2).

7.3.2. Colourings and partitioning (di)graphs

7.3.2.1. 2-partitions of digraphs

A k-*partition* of a (di)graph D is a partition of $V(D)$ into k disjoint sets. Let $\mathbb{P}_1, \mathbb{P}_2$ be two (di)graph properties, then a $(\mathbb{P}_1, \mathbb{P}_2)$-*partition* of a (di)graph D is a 2-partition (V_1, V_2) where V_1 induces a (di)graph with property $\mathbb{P}_1$ and V_2 a (di)graph with property $\mathbb{P}_2$. In [14], [13] and [38], [37], we give a complete characterization for the complexity of $(\mathbb{P}_1, mathbbP_2)$-partition problems when $\mathbb{P}_1, \mathbb{P}_2$ are one of the following standard properties: acyclic, complete, independent (no arcs), oriented (no directed 2-cycle), semicomplete, tournament, symmetric (if two vertices are adjacent, then they induce a directed 2-cycle), strongly connected, connected, minimum out-degree at least 1, minimum in-degree at least 1, minimum semi-degree at least 1, minimum degree at least 1, having an out-branching, having an in-branching. We also investigate the influence of strong connectivity of the input digraph on this complexity. In particular, we show that some NP-complete probems become polynomial-time solvable when restricted to strongly connected input digraphs.

7.3.2.2. χ-bounded families of oriented graphs

A famous conjecture of Gyárfás and Sumner states for any tree T and integer k, if the chromatic number of a graph is large enough, either the graph contains a clique of size k or it contains T as an induced subgraph. In [57], we discuss some results and open problems about extensions of this conjecture to oriented graphs. We conjecture that for every oriented star S and integer k, if the chromatic number of a digraph is large enough, either the digraph contains a clique of size k or it contains S as an induced subgraph. As an evidence, we prove that for any oriented star S, every oriented graph with sufficiently large chromatic number contains either a transitive tournament of order 3 or S as an induced subdigraph. We then study for which sets $\mathcal{P}$ of orientations of P_4 (the path on four vertices) similar statements hold. We establish some positive and negative results.

7.3.2.3. Locally irregular decompositions of subcubic graphs

A graph G is *locally irregular* if every two adjacent vertices of G have different degrees. A *locally irregular decomposition* of G is a partition $E_1, ..., E_k$ of $E(G)$ such that each $G[E_i]$ is locally irregular. Not all graphs admit locally irregular decompositions, but for those who are decomposable, in that sense, it was conjectured by Baudon, Bensmail, Przybylo and Wozniak that they decompose into at most 3 locally irregular graphs. Towards that conjecture, it was recently proved by Bensmail, Merker and Thomassen that every decomposable graph decomposes into at most 328 locally irregular graphs. In [39], we focus on locally irregular decompositions of subcubic graphs, which form an important family of graphs in this context, as all non-decomposable graphs are subcubic. As a main result, we prove that decomposable subcubic graphs decompose into at most 5 locally irregular graphs, and only 4 when the maximum average degree is less than 12/5. We then consider weaker decompositions, where subgraphs can also include regular connected components, and prove the relaxations of the conjecture above for subcubic graphs.

7.3.2.4. Orientation and edge-weigthing inducing colouring

An orientation of a graph G is *proper* if two adjacent vertices have different indegrees. The *proper-orientation number* of a graph G is the minimum maximum indegree of a proper orientation of G. In a previous paper, we raise the question whether the proper orientation number of a planar graph is bounded. In [12], we prove that every cactus admits a proper orientation with maximum indegree at most 7. We also prove that the bound 7 is tight by showing a cactus having no proper orientation with maximum indegree less than 7. We also prove that any planar claw-free graph has a proper orientation with maximum indegree at most 6 and that this bound can also be attained.

7.3.2.5. 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.

In [58], we study the existence of sum-distinguishing injective $|E(G)|$-edge-weightings. We conjecture that such an edge-weighting always exists (except from K_2). We prove this conjecture for some classes of graphs, such as trees and regular graphs. In addition, for some other classes of graphs, such as 2-degenerate graphs and graphs with maximum average degree at most 3, we prove that, provided we use a constant number of additional edge weights, the desired edge-weighting always exists. Our investigations are strongly related to several aspects of the well-known 1-2-3 Conjecture and the Antimagic Labelling Conjecture.

One of the variants consists in considering total-labelling rather than edge-weighting. A *k-total-weighting* of a graph G is an application from $V(G) \cup E(G)$ into $\{1, \cdots, k\}$. An edge-weigthing is *sum-distinguishing* if for every two adajcent vertices u and v, the sum of of weights of u and the edges incident to u is distinct from the sum of weights of v and the edges incident to v. The 1-2 Conjecture raised by Przybylo, lo and Wozniak in 2010 asserts that every undirected graph admits a 2-total-weighting (both vertices and eedges receives weights) such that the sums of weights "incident" to the vertices yield a proper vertex-colouring. Following several recent works bringing related problems and notions (such as the well-known 1-2-3 Conjecture, and the notion of locally irregular decompositions) to digraphs, we introduce in [40] and study several variants of the 1-2 Conjecture for digraphs. For every such variant, we raise conjectures concerning the number of weights necessary to obtain a desired total-weighting in any digraph. We verify some of these conjectures, while we obtain close results towards the ones that are still open.

7.3.2.6. Colouring game

We wish to motivate the problem of finding decentralized lower-bounds on the complexity of computing a Nash equilibrium in graph games. While the centralized computation of an equilibrium in polynomial time is generally perceived as a positive result, this does not reflect well the reality of some applications where the game serves to implement distributed resource allocation algorithms, or to model the social choices of users with limited memory and computing power. As a case study, we investigate in [31] on the parallel complexity of a game-theoretic variation of graph colouring. These “colouring games" were shown to capture key properties of the more general welfare games and Hedonic games. On the positive side, it can be computed a Nash equilibrium in polynomial-time for any such game with a local search algorithm. However, the algorithm is time-consuming and it requires polynomial space. The latter questions the use of colouring games in the modeling of information-propagation in social networks. We prove that the problem of computing a Nash equi- librium in a given colouring game is PTIME-hard, and so, it is unlikely that one can be computed with an efficient distributed algorithm. The latter brings more insights on the complexity of these games.

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 i G. It *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.

Particular interest was dedicated to grids as many processor networks have a grid topology. There are three types of regular infinite grids in the plane, namely the hexagonal grids, the square grids and the triangular grids. In [26], [42], we study the square grid $\mathcal{S}_k$) with infinite width and bounded height k. We prove that the minimum density of an identifying code in $\mathcal{S}_k$ is at least $\frac{7}{20} + \frac{1}{20k}$ and at most $\frac{7}{20} + \frac{3}{10k}$. We also establish that the minimum density of a code in an infinite square grid of height 3 is $\frac{7}{18}$. In [49], [30], we study the minimum density $d^*(\mathcal{T}_k)$ of the triangular grid $\mathcal{S}_k$) with infinite width and bounded height k. We prove

that $d^*(T_k) = \frac{1}{4} + \frac{1}{4k}$ for every odd k and $\frac{1}{4} + \frac{1}{4k} \leq d^*(T_k) \leq \frac{1}{4} + \frac{1}{2k}$ for every even k. We also prove $d^*(T_2) = \frac{1}{2}$ and $d^*(T_4) = d^*(T_6) = \frac{1}{3}$. All these proofs are made using the discharging method, which seems not have been very rarely used for such problems whereas it applies very well.

DANTE Project-Team

7. New Results

7.1. Graph & Signal Processing

Participants: Sarra Ben Alaya, Éric Fleury, Paulo Gonçalves Andrade.

7.1.1. Isometric graph shift operator]

Following up the PhD work of Benjamin Girault [57], we demonstrated in [26] that the isometric graph shift operator we originally proposed, does have a vertex-domain interpretation as a diffusion operator using a polynomial approximation. We showed that its impulse response exhibits an exponential decay of the energy way from the impulse, demonstrating localisation preservation. Additionally, we formalised several techniques that can be used to study other graph signal operators.

7.2. Performance analysis and networks protocols

Participants: Mohammed Amer, Thomas Begin, Anthony Busson, Éric Fleury, Paulo Gonçalves Andrade, Yannick Léo, Isabelle Guérin Lassous, Philippe Nain, Huu Nghi Nguyen, Laurent Reynaud.

7.2.1. Use of large scale CDR for protocol performance evaluation and modelling

In [11] we use large scale CDR (Call Data Records) coming from a nationwide cellular telecommunication operator during a two month period to validate several DTN approaches for conveying SMS traffic in dense urban areas taking benefits of the density of users and the mobility of the users. We study a mobile dataset including 8 Million users living in large urban area. This gives us a precise estimation of the average transmission time and the global performance of our approach. Our analysis shows that after 30 min, half of the SMS are delivered successfully to destination. In [10], we study the temporal activity of a user and the user movements. At the user scale, the usage is not only defined by the amount of calls but also by the userâs mobility. At a higher level, the base stations have a key role on the quality of service. From a very large Call Detail Records (CDR) we first study call duration and inter-arrival time parameters. Then, we assess user movements between consecutive calls (switching from a station to another one). Our study suggests that user mobility is pretty dependent on user activity. Furthermore, we show properties of the inter-call mobility by making an analysis of the call distribution.

7.2.2. End-to-end delay

Because of the growing complexity of computer networks, a new paradigm has been introduced to ease their design and management, namely, the SDN (Software-defined Networking). In particular, SDN defines a new entity, the controller that is in charge of controlling the devices belonging to the data plane In order to let the controller take its decisions, it must have a global view on the network. This includes the topology of the network and its links capacity, along with other possible performance metrics such delays, loss rates, and available bandwidths. This knowledge can enable a multi-class routing, or help guarantee levels of Quality of Service. In [33], [20], [42], we proposed new algorithms that allow a centralised entity, such as the controller in an SDN network, to accurately estimate the end-to-end delay for a given flow in its network. The proposed methods are passive in the sense that they do not require any additional traffic to be run. Through extensive simulations, we show that these methods are able to accurately estimate the expectation and the standard deviation of end-to-end delays.

In [14] we investigated the traversal time of a file across N communication links subject to stochastic changes in the sending rate of each link. Each link's sending rate is modelled by a finite-state Markov process. Two cases, one where links evolve independently of one another (N mutually independent Markov processes), and the second where their behaviours are dependent (these N Markov processes are not mutually independent) were considered. A particular instance where the above is

7.2.3. Circumventing the complexity of multi-server queues

Many real-life systems can be viewed as instances of multi-server queues. However, when the number of servers is high (say more than 16) and the arrival or/and service process exhibit high variability, current state-of-the-art solutions often become intractable due to the combinatorial growth of the underlying state space of the Markov chain. We proposed two efficient, fast and easy-to-implement approximate solutions to deal with *G/G/c*-like queues in [4], [2]. Our solutions rely the use of an original, though incomplete, state description that heavily breaks the complexity of multi-server queues. We have extensively validated our approximations against discrete-event simulation for several QoS performance metrics such as mean sojourn time and blocking probability with excellent results.

7.2.4. Wi-Fi networks 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 proposed original solutions [23], [22] to this optimization problem based on two contributions. First, we modeled the association optimization problem assuming a realistic share of the medium between APs and stations and among APs when using the 802.11 DCF (Distributed Coordination Function) mode. Then, we introduced a local search algorithm to solve this problem through a suitable neighborhood structure. We show that the classical approaches in the literature, based on a time based fairness scheme, is less efficient than our solution when the number of orthogonal channels is limited. Also, we show through a large set of simulations and scenarios that our models are able to capture the real throughputs of Wi-Fi networks.

7.2.5. Controlled mobility in wireless networks

In this work, 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. In [7], we describe an adapted controlled mobility strategy and detail the design of our Virtual Force Protocol (VFP) which allows a swarm of vehicles to track and follow hornets to their nests, while maintaining connectivity through a wireless multi-hop communication route with a remote ground station used to store applicative data such as hornet trajectory and vehicle telemetry. In [43], we design a physics-based controlled mobility strategy, which we name the extended Virtual Force Protocol (VFPe), allowing self-propelled nodes, and in particular here unmanned aerial vehicles, to fly autonomously and cooperatively. In this way, ground devices scattered on the operation site may establish communications through the wireless multi-hop communication routes formed by the network of aerial nodes. In [28], we design a virtual force-based controlled mobility scheme, named VFPc, and evaluate its ability to be jointly used with a dual packet-forwarding and epidemic routing protocol. In particular, we study the possibility for end-users to achieve synchronous communications at given times of the considered scenarios.

7.3. Modeling of Dynamics of Complex Networks

Participants: Christophe Crespelle, Éric Fleury, Márton Karsai, Yannick Léo, Philippe Nain, Matteo Morini.

7.3.1. Data Driven studies on socioeconomic data and communication networks

The study of correlations between the social network and economic status of individuals is difficult due to the lack of large-scale multimodal data disclosing both the social ties and economic indicators of the same population. Thanks to our collaboration with GranData, we close this gap through the analysis of coupled datasets recording the mobile phone communications and bank transaction history of one million anonymised individuals living in a Latin American country. From this large scale data set based on a representative, society-large population we empirically demonstrate some long-lasting hypotheses on socioeconomic correlations, which potentially lay behind social segregation, and induce differences in human mobility. More precisely, in [12] we show that wealth and debt are unevenly distributed among people in agreement with the Pareto principle; the observed social structure is strongly stratified, with people being better connected to others

of their own socioeconomic class rather than to others of different classes; the social network appears to have assortative socioeconomic correlations and tightly connected ârich clubsâ; and that individuals from the same class live closer to each other but commute further if they are wealthier. In [41], 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. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in recommendation system design. In [36] we provide insight about the effects of marking events on the structure and the dynamics of egocentric networks. More precisely, we study the impact of university admission on the composition and evolution of the egocentric networks of freshmen. In other words, we study whether university helps to build connections between egos from different socioeconomic classes, or new social ties emerge via homophilic effects between students of similar economic status. Finally, in [44],

7.3.2. Generalisation of multilayer and temporal graphs

In [16] we introduce the concept of MultiAspect Graph (MAG) as a graph generalisation that we prove to be isomorphic to a directed graph, and also capable of representing all previous generalisations of multilayer and temporal networks. In our proposal, the set of vertices, layers, time instants, or any other independent features are considered as an aspect of the MAG. For instance, a MAG is able to represent multilayer or time-varying networks, while both concepts can also be combined to represent a multilayer time-varying network and even other higher-order networks. Since the MAG structure admits an arbitrary (finite) number of aspects, it hence introduces a powerful modelling abstraction for networked complex systems. In [17] we develop the algebraic representation and basic algorithms for MultiAspect Graphs (MAGs). In particular, we show that, as a consequence of the properties associated with the MAG structure, a MAG can be represented in matrix form. Moreover, we also show that any possible MAG function (algorithm) can be obtained from this matrix-based representation. This is an important theoretical result since it paves the way for adapting well-known graph algorithms for application in MAGs. We present a set of basic MAG algorithms, constructed from well-known graph algorithms, such as degree computing, Breadth First Search (BFS), and Depth First Search (DFS).

Multilayer networks arise in scenarios when a common set of nodes form multiple networks via different co-existing, and sometimes interdependent means of connectivity. In [6] we studied the threshold on the occupation density in the individual network layers for long-range connectivity to emerge in a large multilayer network. For a multilayer network formed via merging M random instances of a graph G with site-occupation probability q in each layer, we showed that when q exceeds a threshold $q_c(M)$, a giant connected component appears in the M-layer network. We showed that $q_c(M) \lesssim \sqrt{-\ln(1-p_c)}/\sqrt{M}$, where p_c is the bond percolation threshold of G, and $q_c(1) \equiv q_c$ is by definition the site percolation threshold of G. We found $q_c(M)$ exactly for when G is a large random graph with any given node-degree distribution. We calculated $q_c(M)$ numerically for various regular lattices, and obtained an exact lower bound for the kagome lattice. Finally, we established an intriguing close connection between the aforesaid multilayer percolation model and the well-studied problem of site-bond (or, mixed) percolation, in the sense that both models provide a bridge between the traditional independent site and independent bond percolation models. Using this connection, and leveraging some analytical approximations to the site-bond critical region developed in the 1990s, we derived an excellent general approximation to the multilayer threshold $q_c(M)$ for regular lattices, which are not only functions solely of the p_c and q_c of the respective lattices, but also closely match the true values of $q_c(M)$ for a large class of lattices, even for small (single-digit) vales of M.

7.3.3. User-based representation of dynamical multimodal public transportation networks

In this project published as an invited paper [9], we provide a novel user-based representation of public transportation systems, which combines representations, accounting for the presence of multiple lines and reducing the effect of spatial embeddedness, while considering the total travel time, its variability across the schedule, and taking into account the number of transfers necessary. After the adjustment of earlier

techniques to the novel representation framework, we analyse the public transportation systems of several French municipal areas and identify hidden patterns of privileged connections. Furthermore, we study their efficiency as compared to the commuting flow. The proposed representation could help to enhance resilience of local transportation systems to provide better design policies for future developments.

7.3.4. Local cascades induced global contagion

In this paper [8] we analyse and model product adoption dynamics in the world's largest voice over internet service, the social network of Skype. We provide empirical evidence about the heterogeneous distribution of fractional behavioural thresholds, which appears to be independent of the degree of adopting egos. We show that the structure of real-world adoption clusters is radically different from previous theoretical expectations, since vulnerable adoptions induced by a single adopting neighbour appear to be important only locally, while spontaneous adopters arriving at a constant rate and the involvement of unconcerned individuals govern the global emergence of social spreading.

7.3.5. Asymptotic theory of time-varying social networks with heterogeneous activity and tie allocation

In this work [15] we empirically characterise social activity and memory in seven real networks describing temporal human interactions in three different settings: scientific collaborations, Twitter mentions, and mobile phone calls. We find that the individuals' social activity and their strategy in choosing ties where to allocate their social interactions can be quantitatively described and encoded in a simple stochastic network modelling framework. The Master Equation of the model can be solved in the asymptotic limit. The analytical solutions provide an explicit description of both the system dynamic and the dynamical scaling laws characterising crucial aspects about the evolution of the networks. The analytical predictions match with accuracy the empirical observations, thus validating the theoretical approach. Our results provide a rigorous dynamical system framework that can be extended to include other processes shaping social dynamics and to generate data driven predictions for the asymptotic behaviour of social networks.

7.3.6. Link prediction in the Twitter mention network

In this project [35] we analyse a large Twitter data corpus and quantify similarities between people by considering the set of their common friends and the set of their commonly shared hashtags in order to predict mention links among them. We show that these similarity measures are correlated among connected people and that the combination of contextual and local structural features provides better predictions as compared to cases where they are considered separately.

DIANA Project-Team

6. New Results

6.1. Service Transparency

6.1.1. From Network-level Measurements to Expected QoE

Participants: Chadi Barakat, Thierry Spetebroot, Muhammad Jawad Khokhar, Damien Saucez and Nawfal Abbassi Saber.

Internet applications, especially those of multimedia type and in a mobile context, are very sensitive to the delivery service they get from the network. However, the relation between this network service and the quality of these applications as perceived by the end users is often unknown and hard to be quantified. Some of the applications dispose of their own quality estimation techniques such as Skype and Viber. Others leave the users to their own interpretation of the quality they perceive. Linking the quality of Internet applications as perceived by the Internet users to network-level measurements such as bandwidth or delay is more than ever necessary. Such dependence, known in the literature as linking Quality of Experience (QoE) to Quality of Service (QoS) parameters, serves many purposes. On one side it allows the estimation of the quality an Internet user will obtain before launching the application or even before heading to the place where she/he will connect. On the other side, it helps network operators properly dimension their networks so that to anticipate service degradation and optimize the quality they deliver. The correlation of quality measurements among users, or for the same user among different of his/her locations, can help in troubleshooting the reasons of any degraded quality.

Our project, called ACQUA, aims at the estimation of the quality of Internet applications at the access departing from network-level measurements. It leverages measurements done at the network level as done today (bandwidth, delay, loss rate, etc), and applies over them well calibrated models to estimate/predict the quality of experience for main applications even before launching them. ACQUA is an extensible solution in terms of the applications it can track. It allows a fine-grained profiling of the Internet access at the level of application quality. In a recent work, we have proved the feasibility of the approach with the Skype use case. We have integrated into ACQUA a new model based on decision trees for the estimation of Skype QoE. The model has been validated with both local controlled and PlanetLab experiments. In 2016, we focused on the popular YouTube use case. We set up a new experimental setup to automatically stream videos, change network conditions, and write down the corresponding Quality of Experience (modeled as a function of application level Quality of Service metrics). One of the challenges we had to face is the reduction of the complexity of experimentation that we had to solve using sampling techniques. The first results are very promising as we can considerably reduce the complexity of experimentation while reaching high level of accuracy in the prediction of Youtube Quality of Experience. A paper is currently under submission illustrating the methodology and the obtained results. More details on this approach and on our project ACQUA can be found in section 5.1 and on the project web page http://project.inria.fr/acqua/.

6.1.2. Testing for Traffic Differentiation with ChkDiff: The Downstream Case

Participants: Ricardo Ravaioli and Chadi Barakat.

In the past decade it has been found that some Internet operators offer degraded service to selected user traffic by applying various differentiation techniques. If from a legal point of view many countries have discussed and approved laws in favor of Internet neutrality, confirmation with measuring tools for even an experienced user remains hard in practice. In this contribution, we extend and complete our tool ChkDiff, previously presented for the upstream case, by checking for shaping also on the user's downstream traffic. After attempting to localize shapers at the access ISP on upstream traffic, we replay downstream traffic from a measurement server and analyze per-flow one-way delays and losses, while taking into account the possibility of multiple paths between the two endpoints. As opposed to other proposals in the literature, our methodology does not

depend on any specific Internet application a user might want to test and it is robust to evolving differentiation techniques that alter delays or induce losses. In a recent publication [22], we provide a detailed description of the downstream tool and a validation in the wild for wired, WiFi and 3G connections. This work is the result of collaboration with the SIGNET group at I3S in the context of a PhD thesis funded by the UCN@Sophia Labex and defended in 2016.

6.1.3. 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.

6.1.4. How news media use Twitter to attract traffic?

Participants: Arnaud Legout, Maksym Gabielkov.

Online news domains increasingly rely on social media to drive traffic to their website. Yet we know surprisingly little about how social media conversation mentioning an online article actually generates a click to it. Posting behaviors, in contrast, have been fully or partially available and scrutinized over the years. While this has led to to multiple assumptions on the diffusion of information, each were designed or validated while ignoring this important step.

We present in [18] a large scale, validated and reproducible study of social clicks – that is also the first data of its kind – gathering a month of web visits to online resources that are located in 5 leading news domains and that are mentioned in the third largest social media by web referral (Twitter). Our dataset amounts to 2.8 million posts, together responsible for 75 billion potential views on this social media, and 9.6 million actual clicks to 59,088 unique resources. We design a reproducible methodology, carefully corrected its biases, enabling data sharing, future collection and validation. As we prove, properties of clicks and social media Click-Through-Rates (CTR) impact multiple aspects of information diffusion, all previously unknown. Secondary resources, that are not promoted through headlines and are responsible for the long tail of content popularity, generate more clicks both in absolute and relative terms. Social media attention is actually long-lived, in contrast with temporal evolution estimated from posts or impressions. The actual influence of an intermediary or a resource is poorly predicted by their posting behavior, but we show how that prediction can be made more precise.

6.1.5. ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic

Participant: Arnaud Legout.

It is well known that apps running on mobile devices extensively track and leak users' personally identifiable information (PII); however, these users have little visibility into PII leaked through the network traffic generated by their devices, and have poor control over how, when and where that traffic is sent and handled by third parties. In this paper, we present the design, implementation, and evaluation of ReCon: a cross-platform system that reveals PII leaks and gives users control over them without requiring any special privileges or custom OSes. ReCon leverages machine learning to reveal potential PII leaks by inspecting network traffic, and provides a visualization tool to empower users with the ability to control these leaks via blocking or substitution of PII. We evaluate ReCon's effectiveness with measurements from controlled experiments using leaks from the 100 most popular iOS, Android, and Windows Phone apps, and via an Institutional Review Board approved user study with 92 participants. We show that ReCon is accurate, efficient, and identifies a wider range of PII than previous approaches.

6.2. Open Network Architecture

6.2.1. Storage on Wheels: Offloading Popular Contents Through a Vehicular Cloud

Participants: Luigi Vigneri and Chadi Barakat.
The increasing demand for mobile data is overloading the cellular infrastructure. Small cells and edge caching is being explored as an alternative, but installation and maintenance costs for sufficient coverage are significant. In this work, we perform a preliminary study of an alternative architecture based on two main ideas: (i) using vehicles as mobile caches that can be accessed by user devices; compared to small cells, vehicles are more widespread and require lower costs; (ii) combining the mobility of vehicles with delayed content access to increase the number of cache hits (and reduce the load on the infrastructure). Contrary to standard DTN-type approaches, in our system max delays are guaranteed to be kept to a few minutes (beyond this deadline, the content is fetched from the infrastructure). We first propose an analytical framework to compute the optimal number of content replicas that one should cache, in order to minimize the infrastructure load. We then investigate how to optimally refresh these caches to introduce new contents, as well as to react to the temporal variability in content popularity. Simulations suggest that our vehicular cloud considerably reduces the infrastructure load in urban settings, assuming modest penetration rates and tolerable content access delays. This work has been published in [24]. It is the result of collaboration with Thrasyvoulos Spyropoulos from the Mobile Communications Department at Eurecom in the context of a PhD thesis funded by the UCN@Sophia Labex.

In another work, published in [25], and always in the context of the same collaboration with Thrasyvoulos Spyropoulos, we studied the feasibility of the approach using the popular video streaming case. In this work, we assume such a vehicular cloud is in place to provide video streaming to users, and that the operator can decide which content to store in the vehicle caches. Users can then greedily fill their playout buffer with video pieces of the streamed content from encountered vehicles, and turn to the infrastructure immediately when the playout buffer is empty, to ensure uninterrupted streaming. Our main contribution is to model the playout buffer in the user device with a queuing approach, and to provide a mathematical formulation for the idle periods of this buffer, which relate to the bytes downloaded from the cellular infrastructure. We also solve the resulting content allocation problem, and perform trace-based simulations to finally show that up to 50% of the original traffic could be offloaded from the main infrastructure.

6.2.2. SDN for QoE-based network optimization and management

Participants: Vitalii Poliakov, Damien Saucez.
The naive approach of the networking community is to always increase network capacity to absorb the traffic. In this thesis, we take the counterpoint of this approach claiming that it is possible to better use network resources if we take into account the Quality of Experience (QoE) of users while making routing decisions. The idea is that each network service (e.g., video streaming, web, chat) has different requirements in terms of network performances such as bandwidth or delay and that modern networks present high path diversity, particularly 5G. Our work is thus to provide mechanisms to decide how to route traffic in the network, potentially using multiple paths in parallel, based on their real impact on the QoE. For example, if the experience of a user is not negatively impacted if their traffic is diverted on a slow path, we can use it to free resources for traffic that really needs the high speed path. Initial results for this new activities are published in [27] and [21].

6.2.3. Measurements of LISP

Participant: Damien Saucez.

To face the new challenges of the Internet such as the Cloud and mobility the Locator/ID Separation Protocol (LISP) leverages the separation of the identifier and the locator roles of IP addresses. Contrarily to the classical BGP-based routing architecture, LISP relies on a pull model. In particular, routing information is pulled from a new network element, the Mapping System, to provide the association between the identifier (i.e., the address used to identify a host inside a domain) and a list of locators (i.e., the addresses to locate an attachment point) upon an explicit query. We evaluate a LISP Mapping System deployment in the public LISP Beta Network deployment from two aspects: Stability and Consistency. Our measurements show that the mapping information is stable over time and consistent between the different mapping entities and the vantage points. Due to the presence of few cases where the Mapping System is unstable and/or inconsistent, we propose a taxonomy in order to classify such instabilities and/or inconsistencies and investigate them in depth to provide hints on how to improve LISP performance. Results are published in [26].

6.2.4. Rules Placement Problem in OpenFlow Networks

Participants: Xuan Nam Nguyen, Damien Saucez, Chadi Barakat and Thierry Turletti.
Software-Defined Networking (SDN) abstracts low-level network functionalities to simplify network management and reduce costs. The OpenFlow protocol implements the SDN concept by abstracting network communications as flows to be processed by network elements. In OpenFlow, the high-level policies are translated into network primitives called rules that are distributed over the network. While the abstraction offered by OpenFlow allows to potentially implement any policy, it raises the new question of how to define the rules and where to place them in the network while respecting all technical and administrative requirements. We proposed a comprehensive study of the so-called OpenFlow rules placement problem with a survey of the various proposals intending to solve it [17].

6.2.5. 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 have proposed 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 propose in the same paper 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 is used to study the performance of L2BM. A paper on this work is under submission [37].

6.2.6. Towards unifying content level and network level operations

Participants: Amine Loukili, Damien Saucez, Thierry Turletti.
Programmability of the network to provide content level operations is highly desirable. With the advent of virtualization and network function softwarization, the networking world shifts to Software Defined Networking (SDN) and OpenFlow is one of the most suitable candidates to implement the southbound API (the interface allowing the SDN-controller to program network devices). In the meanwhile, the generalization of broadband Internet has led to massive content consumption. However, while content is usually retrieved via layer 7 protocols, OpenFlow operations are performed at lower layers (layer 4 or lower) making the protocol ineffective to deal with contents. To address this issue, we define an abstraction to unify network level and content level operations and present a straw-man logically centralized architecture proposal to support it. Our implementation demonstrates the feasibility of the solution and its advantage over fully centralized approach. This work has been published in the CoNext student workshop [19]. A demonstration was also presented at IEEE SDN/NFV conference [32].

6.2.7. Resiliency in Service Function Chaining

Participants: Ghada Moualla, Damien Saucez, Thierry Turletti.
In the context of the dynamic placement of Virtual Network Functions in the network activity, we have studied the importance of resiliency in service functions chaining. When deploying network service function chains the focus is usually given on metrics such as the cost, the latency, or the energy and it is assumed that the underlying cloud infrastructure provides resiliency mechanisms to handle with the disruptions occurring in the physical infrastructure. In a position paper on this topic published in PROCON 2016 [20], we advocate that while usual performance metrics are essential to decide on the deployment of network service function chains, the notion of resiliency should not be neglected as the choice of virtual-to-physical placement may dramatically improve the ability of the service chains to handle with failures of the infrastructure without requiring complex resiliency mechanisms.

6.2.8. SDN for Public Safety Networks

Participants: Damien Saucez, Xuan Nam Nguyen, Thierry Turletti.
Commercial users of modern communications networks have benefited from a huge progress of the related technologies. However, Public Safety Networks (PSNs) and devices did not follow the same trend. Very often, they still rely on voice or low speed data communications, tempting first responders to use their own private devices when they need to exchange real-time video or geolocation information. Under this consideration, national authorities and specialized organizations have recently initiated the integration of more recent technologies, such as cellular Long Term Evolution (LTE), even though they need further developments to cope with the harsh usages that safety personnel may face. We wrote a report showing the evolution of these networks towards the recent evolution of networking technologies started with Software Defined networking (SDN) and Network Functions Virtualization (NFV). Based on the requirements derived from a standardized earthquake scenario and a study of the main improvements brought by this network softwarization, it analyzes how SDN and NFV can solve part of the issues raised with commercial LTE and enhance PSN communications. The capabilities of these new technologies are applied to a list of characteristics required by mission-critical networks, e.g., rapid deployment, reliability, security or resilience, taking advantage of features such as the separation between control and data planes or the simplified dynamic resources management. The resulting enhancements are then illustrated using example frameworks published in the literature for Cloud Radio Access Networks, resilient backhaul solution, isolated base stations, SDN-based architecture or Service Function Chaining [28].

6.2.9. Standardization Activities

Participant: Damien Saucez.
The Locator/ID Separation Protocol (LISP) aims to improve the Internet routing by leveraging separating the roles of IP addresses. In RFC7834 [36] we studied the impact that the deployment of LISP would have on both the routing infrastructure and the end user if it was largely deployed in today's Internet. In addition, as bringing new protocols to the Internet opens new security questions, in RFC7835 [35] we provide an exhaustive threat analysis of LISP. Both RFCs are used as insights to extend the architecture of LISP to make it more efficient and safer.

Information Centric Networking (ICN) is a radically new way to conceive networks by promoting content information as routing primitives, instead of content location. In RFC7927 [31], we list the research challenges hidden behind this revolutionary approach of networking. This RFC aims to be the baseline for the development of ICN solutions.

6.3. Experimental Evaluation

6.3.1. 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 fine-grained 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 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 submitted to a specialized workshop on Network Localization on Navigation [33].

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation of Call Centers

Participant: Pierre L'Ecuyer.

We develop research activities around the analysis and design of call centers, from a performance perspective. The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty.

One aspect studied in [23] is the development of stochastic models for the daily arrival rate in a call center. Models in which the busyness factors are independent across periods, or in which a common busyness factor applies to all periods, have been studied previously. But they are not sufficiently realistic. We examine alternative models for which the busyness factors have some form of dependence across periods.

We also carry out in [14] large-scale data-based investigation of service times in a call center with many heterogeneous agents and multiple call types to investigate the validity of traditionally used standard Erlang queueing models, based on independent and identically distributed exponential random variables. Our study provides empirical support to the theoretical research that goes beyond standard modelling assumptions in service systems.

In [56], we consider a stochastic staffing problem with uncertain arrival rates. The objective is to minimize the total cost of agents under some chance constraints, defined over the randomness of the service level in a given time period. We present a method that combines simulation, mixed integer programming, and cut generation to solve this problem. In [84], we consider a particular staffing problem with probabilistic constraints in an emergency call center. We propose an algorithm to solve the problem, and validate it with a simulation model based on real data from the 911 emergency call center of Montreal, Canada.

We are also interested in predicting the waiting time of customers upon their arrival in some service system such as a call center or emergency service. In [86], we propose two new predictors that are very simple to implement and can be used in multiskill settings. They are based on the waiting times of previous customers of the same class. In our simulation experiments, these new predictors are very competitive with the optimal ones for a simple queue, and for multiskill centers they perform better than other predictors of comparable simplicity.

7.2. Analytic models

Participants: Gerardo Rubino, Bruno Sericola.

Sojourn times in Markovian models. In [98], we discuss different issues related to the time a Markov chain spends in a part of its state space. This is relevant in many application areas including those interesting Dionysos, namely, in the performance and dependability analysis of complex systems. For instance, in dependability, the reliability of a system subject to failures and repairs of its components, is, in terms of a discrete-space model of it, the probability that it remains in the subset of operational or up states during the whole time interval $[0, t]$. In performance, the occupancy factor of some server is the probability that, in steady state, the model belongs to the subset of states where the server is busy. This book chapter reviews some past work done by the authors on this topic (see our book [111] for a synthesis of these works), and add some new insights on the properties of these sojourn times.

Queuing systems in equilibrium. In the late 70s, Leonard Kleinrock proposed a metric able to capture the tradeoff between the work done by a system and its cost, or, in terms of queueing systems, between throughput and mean response time. The new metric was called *power* and among its properties, it satisfies a nice one informally called "keep the pipe full", specifying that the operation point of many queues that maximizes their power also leads to a mean backlog equal to exactly one customer. Last year [110] we explored what happens with this metric when we consider Jackson queuing networks. After showing that the same property holds for them, we showed that the power metric has some drawbacks, mainly when considering multiserver queues and networks of queues. We then proposed a new metric that we called *effectiveness*, identical to power when there is a single queue with a single server, but different otherwise, that avoids these drawbacks. We analyze it and, in particular, we showed that the same "keep the pipe full" holds for it. In the keynote [34] we presented these ideas together with some new results (for example, the analysis of G-queues from this point of view).

For other analytical-oriented work, see [72] for new applications of queueing theory used at the Markovian level, and [72] for applications of stochastic analysis to general problems where performance and dependability are simultaneously taken into account in the same model.

7.3. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Yann Busnel, Yves Mocquard.

Detection of distributed deny of service attacks. A Deny of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Deny of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term "iceberg" has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs [41]. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modeled as items) prior to sending them to the server, and throw away all the items that appear with a small probability. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the time needed to collect c distinct items among n frequent ones. A thorough analysis of the time needed to collect c distinct items appears in [10].

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

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies. We have proposed a new key grouping technique targeted toward applications working on input streams characterized by a skewed value distribution [80]. Our solution is based on the observation that when the values used to perform the grouping have skewed frequencies, the few most frequent values (the *heavy hitters*) drive the load distribution, while the remaining largest fraction of the values (the *sparse items*) appear so rarely in the stream that the relative impact of each of them on the global load balance is negligible. We have shown, through a theoretical analysis, that our solution provides on average near-optimal mappings using sub-linear spaces in the number of tuples

read from the input stream in the learning phase and the support (value domain) of the tuples. In particular this analysis presents new results regarding the expected error made on the estimation of the frequency of heavy hitters.

Load shedding is a technique employed by stream processing systems to handle unpredictable spikes in the input load whenever available computing resources are not adequately provisioned. A load shedder drops tuples to keep the input load below a critical threshold and thus avoid unbounded queuing and system trashing. In [102] and [79] we propose Load-Aware Shedding (LAS), a novel load shedding solution that, unlike previous works, does not rely neither on a pre-defined cost model nor on any assumption on the tuple execution duration. Leveraging sketches, LAS efficiently builds and maintains at runtime a cost model to estimate the execution duration of each tuple with small error bounds. This estimation enables a proactive load shedding of the input stream at any operator that aims at limiting queuing latencies while dropping as few tuples as possible. We provide a theoretical analysis proving that LAS is an (ε, δ)-approximation of the optimal online load shedder. Furthermore, through an extensive practical evaluation based on simulations and a prototype, we evaluate its impact on stream processing applications, which validate the robustness and accuracy of LAS.

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 tu-ples. 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 [77] 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.

Estimating the frequency of any piece of information in large-scale distributed data streams became of utmost importance in the last decade (*e.g.*, in the context of network monitoring, big data, *etc.*). If some elegant solutions have been proposed recently, their approximation is computed from the inception of the stream. In a runtime distributed context, one would prefer to gather information only about the recent past. This may be led by the need to save resources or by the fact that recent information is more relevant. In [78], we consider the *sliding window* model and propose two different (on-line) algorithms that approximate the items frequency in the active window. More precisely, we determine a (ε, δ)-additive-approximation meaning that the error is greater than ε only with probability δ. These solutions use a very small amount of memory with respect to the size N of the window and the number n of distinct items of the stream, namely, $O(\frac{1}{\varepsilon}\log\frac{1}{\delta}(\log N + \log n))$ and $O(\frac{1}{\tau\varepsilon}\log\frac{1}{\delta}(\log N + \log n))$ bits of space, where τ is a parameter limiting memory usage. We also provide their distributed variant, *i.e.*, considering the *sliding window functional monitoring* model, with a communication cost of $O(\frac{k}{\varepsilon^2}\log\frac{1}{\delta}\log N)$ bits per window (where k is the number of nodes). We compared the proposed algorithms to each other and also to the state of the art through extensive experiments on synthetic traces and real data sets that validate the robustness and accuracy of our algorithms.

Randomized Message-Passing Test-and-Set. In [101], we have presented a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes, returns yes to a unique process and returns no to all the others. Recently, many advances in implementing Test&Set objects have been achieved. However, all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number $p \leq n$ of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two

groups of processes. We propose a message- passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

Throughput Prediction in Cellular Networks Downlink data rates can vary significantly in cellular networks, with a potentially non-negligible effect on the user experience. Content providers address this problem by using different representations (*e.g.*, picture resolution, video resolution and rate) of the same content and switch among these based on measurements collected during the connection. 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 [81].

Population Protocol Model. The computational model of population protocols, introduced by Angluin and his colleagues in 2006, is a formalism that allows the analysis of properties emerging from simple and pairwise interactions among a very large number of anonymous finite-state agents. Significant work has been done so far to determine which problems are solvable in this model and at which cost in terms of states used by the protocols and time needed to converge. The problem tackled in [74] is the population proportion problem: each agent starts independently from each other in one of two states, say A or B, and the objective is for each agent to determine the proportion of agents that initially started in state A, assuming that each agent only uses a finite set of state, and does not know the number n of agents. We propose a solution which guarantees with any high probability that after $O(\log n)$ interactions any agent outputs with a precision given in advance, the proportion of agents that start in state A. The population proportion problem is a generalization of both the majority and counting problems, and thus our solution solves both problems. We show that our solution is optimal in time and space. Simulation results illustrate our theoretical analysis.

The context of [75] 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. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In this paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at time unit, a single node interacts with a randomly chosen neighbor. We perform a thorough study of T_n, the total number of interactions needed for all the n nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow for comparing different protocols, we prove that in a complete graph of size $n \geq 2$, the probability that $T_n > k$ for all $k \geq 1$ is less than $(1 + 2k(n-2)^2/n)(1-2/n)^{k-1}$. We also study the behavior of the complementary distribution of T_n at point $cE(T_n)$ when n tends to infinity, in function of c. This paper received the Best Student Paper Award from the 15th IEEE Symposium on Network Computing and Applications (IEEE NCA 2016).

Bitcoin. Decentralized cryptocurrency systems offer a medium of exchange secured by cryptography, without the need of a centralized banking authority. Among others, Bitcoin is considered as the most mature one. Its popularity lies on the introduction of the concept of the blockchain, a public distributed ledger shared by all participants of the system. Double spending attacks and blockchain forks are two main issues in blockchain-based protocols. The first one refers to the ability of an adversary to use the very same bitcoin more than once, while blockchain forks cause transient inconsistencies in the blockchain. We show in [43], [89], [42] through probabilistic analysis that the reliability of recent solutions that exclusively rely on a particular type of Bitcoin actors, called miners, to guarantee the consistency of Bitcoin operations, drastically decreases with the size of the blockchain.

7.4. Future networks and architectures

Participants: Adlen Ksentini, Bruno Sericola, Yassine Hadjadj-Aoul, Jean-Michel Sanner, Hamza Ben Ammar.

SDN and NFV. Network Function Visualization (NFV) and Software Defined Network (SDN) currently play a key role to transform the network architecture from hardware-based to software-based.

SDN is in the process of revolutionizing the way of managing networks by providing a new way to support current and future services. However, by relocating the control functionality in a remote entity, the measurements' accuracy of the resources' utilization becomes more difficult, which complicates the decision making. Although there are previous works focusing on the problem of network management and measurement in SDN networks, only a few proposed solutions have taken into consideration the trade-off existing between statistics' polling frequency (i.e. generated overhead), and the accuracy of monitoring results (i.e. optimized resourcesâ allocation). In [62], we proposed a new approach calculating accurately the bandwidth utilization while adapting the polling frequency according to ports/switches activity. The emulations' results under Mininet clearly demonstrate the effectiveness of the proposed solution, which proved to be scalable compared to classical approaches. The controllers' placement is another important concern that emerged recently to solve the scalability and the reliability issues of SDN networks. The placement efficiency is influenced by both network operators (NO) strategy and the supported service requirements, which makes more complex the decision-making process. In particular, the need to support QoS-constrained services may lead NO to guide the controllers' placement in a way to ensure services efficiency while optimizing the underlying infrastructure. In [82] and [66], we proposed a model for the placement of network controllers, and we formulated a general optimization problem. To provide more flexibility and to avoid time-prohibitive calculations, we proposed a hierarchical clustering strategy for the controllers' placement allowing to minimize the number of network controllers while reducing the potential disparity of burden between the different controllers. Besides, the algorithms' structure makes it easy to act on other network parameters to improve the reliability of the SDN network. In [107], we proposed an improvement of such algorithms, by considering an evolutionary solution based on a genetic technique with an ad hoc cross-over operator designed to solve a mono-objective controller placement problem.

To connect the VNFs hosted in the same Data Center (DC) or across multiple DCs, virtual switches are required. Besides forwarding functions, virtual switches can be configured to mirror traffics for network management needs. Among the existing virtual switch solutions, Open vSwitch (OVS) is the most known and used. OVS is open source, and included in most of the existing Linux distributions. However, OVS performance in terms of throughput for smaller packets is very smaller than of line rate of the interface. To overcome this limitation, OVS was ported to Data Plane Development Kit (DPDK), namely OVDK. The latter achieves an impressive line rate throughput across physical interfaces. In [83], we presented the result of OVDK performance test when flow and port mirroring are activated, which was not tested so far. The performance test focuses on two parameters, throughput and latency in OVDK, allowing to validate the use of OVDK for flow forwarding and network management in the envisioned virtualized network architecture.

Mobile cloud. To cope with the tremendous growth in mobile data traffic on one hand, and the modest average revenue per user on the other hand, mobile operators have been exploring network virtualization and cloud computing technologies to build cost-efficient and elastic mobile networks and to have them offered as a cloud service. In such cloud-based mobile networks, ensuring service resilience is an important challenge to tackle. Indeed, high availability and service reliability are important requirements of carrier grade, but not necessarily intrinsic features of cloud computing. Building a system that requires the five nines reliability on a platform that may not always grant it is therefore a hurdle. Effectively, in carrier cloud, service resilience can be heavily impacted by a failure of any network function (NF) running on a virtual machine (VM). In [31], we introduce a framework, along with efficient and proactive restoration mechanisms, to ensure service resilience in carrier cloud. As restoration of a NF failure impacts a potential number of users, adequate network overload control mechanisms are also proposed. A mathematical model is developed to evaluate the performance of the proposed mechanisms. The obtained results are encouraging and demonstrate that the proposed mechanisms efficiently achieve their design goals.

Typically, maintaining a static pool of cloud resources to meet peak requirements with good service quality makes the cloud infrastructure costly. To cope with this, [58] proposes an approach that enables a cloud infrastructure to automatically and dynamically scale-up or scale-down resources of a virtualized environment aiming for efficient resource utilization and improved quality of experience (QoE) of the offered services. The QoE-aware approach ensures a truly elastic infrastructure, capable of handling sudden load surges while reducing resource and management costs. The paper also discusses the applicability of the proposed approach within the ETSI NFV MANO framework for cloud-based 5G mobile systems.

Video distribution. Due to the Internet usage evolution over these last years, the current IP-based architecture becomes heavier and less efficient for providing Internet services. In order to face this shortcoming, "Content Centric Networking" has been proposed. One of its important features is the use of in-network caching as a way of improving network performance and service scalability. However, in most of the existing CCN-based approaches several copies of the same content are present in the network, which reduce its efficiency. In [45], we proposed the "CLIque-based cooperative Caching" (CLIC) strategy, which basically consists in detecting cliques within the network topology to allocate more efficiently the content in the network. The main motivation of the proposed solution is to eliminate contents' redundancy between neighboring nodes while promoting the most popular contents. This approach guarantees a sufficient number of copies of popular files within the network while maximizing the number of distinct content items. We evaluate the proposed scheme through simulation. The results show significant improvements in terms of cache management and network performance.

In [59], we make the case for opening the telco CDN infrastructure to content providers by means of network function virtualization (NFV) and cloud technologies. We design and implement a CDN-as-a-Service architecture, where content providers can lease CDN resources on demand at regions where the ISP has presence. Using open northbound RESTful APIs, content providers can express performance requirements and demand specifications, which can be translated to an appropriate service placement on the underlying cloud substrate. To gain insight which can be applied to the design of such service placement mechanisms, we evaluate the capabilities of key enabling virtualization technologies by extensive testbed experiments.

Network design using new dependability metrics. When designing a network taking into account its capabilities face to possible failures to its components, the basic theoretical framework is classical network reliability, where the system under study is represented by a graph with perfect nodes and imperfect links randomly and independently failing. The corresponding connectivity-based metrics must then be evaluated in order to quantify the robustness of the networking architecture. Recently, a new family of metrics, called diameter-constrained, have been proposed and analyzed by Dionysos' collaborators and members. In [53], we developed some elements for a factoring theory associated with these metrics. The paper is focused on the detection of irrelevant components, a key task when evaluating these types of quantities using factorization. The paper also includes a factoring algorithm, which is an up-to-date procedure exploiting all available results for implementing the pivoting idea (proved to be one of the most powerful methods in classical reliability analysis).

In [54], we consider an homogeneous network (identical and independent components). In this context, if p is the probability that each of the components works, then any reliability metric is necessarily a polynomial in p, and computing these metrics can be reduced to counting problems (counting specific classes of paths or of cuts, for instance). In the paper, we quantify, in some sense, the "degree of difficulty" of these counting processes, and we identify the situations where they are "easy". The second contribution of the paper is to propose a fundamental problem from survivable network design, called the Network Utility Problem. The goal is to maximize network utility (defined as the opposite of the level of difficulty minus one), under a minimum edge-connectivity requirement.

Optical network design. Paper [65] presents a fast and accurate mathematical method to evaluate the blocking probability (the probability of a burst loss) in dynamic WDM networks without wavelength conversion (the present used technology). We assume that all links have the same number of wavelengths (the same capacity). The proposed model considers different traffic loads at each network connection (heterogeneous traffic). To take into account the wavelength continuity constraint, the method divides the network into a sequence of

networks where all the links have capacity 1. Every network in the sequence is evaluated separately using an analytical technique. Then, a procedure combines the results of these evaluations in a way that captures the dependencies that occur in the real system due to the competition for bandwidth between the different connections. The method efficiently achieves results very close to those obtained by simulation, but orders of magnitude faster, allowing the evaluation of the blocking probability of all users (connections) for mesh network topologies.

7.5. Network Economics

Participants: Bruno Tuffin, 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 published in 2014 [109].

Network neutrality. Most of our activity has been devoted to the vivid network neutrality debate, going beyond the traditional for or against neutrality. We especially responded to the public consultation on draft BEREC Guidelines on implementation of net neutrality rules held during Summer 2016.

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 provider actually paying those fees. In [20] 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.

In [105], we explain how non neutrality may be pushed by big CPs to their benefits. 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. This fits the network neutrality debate, but 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 [105] a model describing the various actors and their strategies, analyzes it thanks to non-cooperative game theory, 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.

While neutrality is focusing on the behavior of ISPs, we claim that the debate should be generalized. Indeed, the reality of the Internet in the 2010s is that various actors contribute to the delivery of data, with sometimes contradictory objectives. We highlight in [19] the fact that neutrality principles can be bypassed in many ways without violating the rules currently evoked in the debate. For example via Content Delivery Networks (CDNs), which deliver content on behalf of content providers for a fee, or via search engines, which can hinder competition and innovation by affecting the visibility and accessibility of content. We therefore call for an extension of the net neutrality debate to all the actors involved in the Internet delivery chain. We particularly challenge the definition of net neutrality as it is generally discussed. Our goal is to initiate a relevant debate for net neutrality in an increasingly complex Internet ecosystem, and to provide examples of possible neutrality rules for different levels of the delivery chain, this level separation being inspired by the OSI layer model.

The impact of a revenue-oriented CDN is particularly investigated in [104] and [70]. 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 (end-users, the network operators and the content providers) raises concerns about their "neutrality", and therefore the question of their inclusion in the network neutrality debate becomes relevant. We compare the outcome with that of a neutral behavior, and at investigating 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. In addition, using the real traces, we compare the revenue-based policy with policies based on several fairness criteria. The CDN activity being potentially lucrative and not included in the neutrality debate, we analyze in [71] the revenue-optimal strategies and impact of a vertically integrated ISP-CDNs, which can sell those services to content providers. Our approach is based on an economic model of revenues and costs, and a multilevel game-theoretic formulation of the interactions among actors. Our model incorporates the possibility for the vertically-integrated ISP to partially offer CDN services to competitors in order to optimize the trade-off between CDN revenue (if fully offered) and competitive advantage on subscriptions at the ISP level (if not offered to competitors). Our results highlight two counterintuitive phenomena: an ISP may prefer an independent CDN over controlling (integrating) a CDN; and from the user point of view, vertical integration is preferable to an independent CDN or a no- CDN configuration. Hence, a regulator may want to elicit such CDN-ISP vertical integrations rather than prevent them.

Online platforms and search engines. Another set of key actors in the Internet economy is the online platforms and search engines. When a keyword-based search query is received by a search engine, a classified ads website, or an online retailer site, the platform has exponentially many choices in how to sort the search results. Two extreme rules are (a) to use a ranking based on estimated relevance only, which improves customer experience in the long run because of perceived quality, and (b) to use a ranking based only on the expected revenue to be generated immediately, which maximizes short-term revenue. Typically, these two objectives (and the corresponding rankings) differ. A key question then is what middle ground between them should be chosen.We introduce in [16] stochastic models that yield elegant solutions for this situation, and we propose effective solution methods to compute a ranking strategy that optimizes long-term revenues. This strategy has a very simple form and is easy to implement if the necessary data is available. It consists in ordering the output items by decreasing order of a score attributed to each. This score results from evaluating a simple function of the estimated relevance, the expected revenue of the link, and a real-valued parameter. We find the latter via simulation-based optimization, and its optimal value is related to the endogeneity of user activity in the platform as a function of the relevance offered to them.

The impact on other actors of search engines has led to the so-called search neutrality debate, as a parallel to the network neutrality debate. Search engines accused of biasing the ranking of their organic links to provide a competitive advantage to their own content. Based on the optimal ranking policy for a search engine obtained in [16], we investigate in [67] on an example whether non-neutrality impacts innovation. We illustrate that a revenue-oriented search engine may indeed deter innovation at the content level, hence the validity of the argument (without necessarily meaning that search engines should be regulated).

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 [33], [92], 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.

7.6. 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.

Rare event simulation. However, when the events of interest are rare, simulation requires a special attention, to accelerate the occurrence of the event and get unbiased estimators of the event of interest with a sufficiently small relative variance (see our book [108] for a global introduction to the field). This is the main problem in the area. Dionysos' work focuses then on dealing with the rare event situation, with a particular focus on dependability [40].

A non-negligible part of our activity on the application of rare event simulation was about the evaluation of static network reliability models. In a static network reliability model one typically assumes that the failures of the components of the network are independent. This simplifying assumption makes it possible to estimate the network reliability efficiently via specialized Monte Carlo algorithms. Hence, a natural question to consider is whether this independence assumption can be relaxed, while still attaining an elegant and tractable model that permits an efficient Monte Carlo algorithm for unreliability estimation. In [12], we provide one possible answer by considering a static network reliability model with dependent link failures, based on a Marshall-Olkin copula, which models the dependence via shocks that take down subsets of components at exponential times, and propose a collection of adapted versions of permutation Monte Carlo (PMC, a conditional Monte Carlo method), its refinement called the turnip method, and generalized splitting (GS) methods, to estimate very small unreliabilities accurately under this model. The PMC and turnip estimators have bounded relative error when the network topology is fixed while the link failure probabilities converge to zero, whereas GS does not have this property. But when the size of the network (or the number of shocks) increases, PMC and turnip eventually fail, whereas GS works nicely (empirically) for very large networks, with over 5000 shocks in our examples. In [73], we propose a methodology for calibrating a dependent failure model to compute the reliability in a telecommunication network, following a similar starting point (that is, using Marshall-Olkin copulas). In practice, this model is difficult to calibrate because it requires the estimation of a number of parameters that is exponential in the number of links. We formulate an optimization problem for calibrating a Marshall-Olkin copula model to attain given marginal failure probabilities for all links and the correlations between them. Using a geographic failure model, we calibrate various Marshall-Olkin copula models using our methodology, we simulate them, and we benchmark the reliabilities thus obtained. Our experiments show that considering the simultaneous failures of small and connected subsets of links is the key for obtaining a good approximation of reliability, confirming what it is suggested by the telecommunication literature.

A related problem is when links have random capacities and a certain target amount of flow must be carried from some source nodes to some destination nodes is considered in [47]. Each destination node has a fixed demand that must be satisfied and each source node has a given supply. The goal is to estimate the unreliability of the network, defined as the probability that given the realized link capacities, the network cannot carry the required amount of flow to meet the demand at all destination nodes. We adapt GS and PMC to this context. In [55], we explore other methods designed to reduce the variance of the estimators in this context. All of them are adaptations of methods originally developed to make reliability estimations on different network models. These methods are introduced together with a brief review of the algorithms on which they are based.

A new application of our previously designed zero-variance approximation importance sampling method has been developed in [76]: To accurately estimate the reliability of highly reliable rail systems and comply with contractual obligations, rail system suppliers such as ALSTOM require efficient reliability estimation techniques. While in our previous works, the studied graph models were dealing with failing links, we propose an adaptation of the algorithm to evaluate the reliability of real transport systems where nodes are the failing components. This is more representative of railway telecommunication system behavior. Robustness measures of the accuracy of the estimates, bounded or vanishing relative error properties, are discussed and results from a real network (Data Communication System used in automated train control system) showing bounded relative error property, are presented.

Random variable generation. Simulation requires the use of pseudo-random generators. In [18], we examine the requirements and the available methods and software to provide (or imitate) uniform random numbers in parallel computing environments. In this context, for the great majority of applications, independent streams

of random numbers are required, each being computed on a single processing element at a time. Sometimes, thousands or even millions of such streams are needed. We explain how they can be produced and managed. We devote particular attention to multiple streams for GPU devices.

Sampling from the Normal distribution truncated to some finite or semi-infinite interval is of particular interest for certain applications in Bayesian statistics, such as to perform exact posterior simulations for parameter inference. We study and compare in [46] various methods to generate such random variables, with special attention to the situation where the interval is far in the tail. The algorithms are implemented and available in Java, R, and MATLAB, and the software is freely available.

Quasi-Monte Carlo (QMC). Finally, we have continued our work on QMC methods. In [15], we review the Array-RQMC method, its variants, sorting strategies, and convergence results. We are interested in the convergence rate of measures of discrepancy of the states at a given step of the chain, as a function of the sample size, and also the convergence rate of the variance of the sample average of a (cost) function of the state at a given step, viewed as an estimator of the expected cost. We summarize known convergence rate results and show empirical results that suggest much better convergence rates than those that are proved. We also compare different types of multivariate sorts to match the chains with the RQMC points, including a sorting procedure based on a Hilbert curve.

The description of a new software tool and library named Lattice Builder, written in C++, that implements a variety of construction algorithms for good rank-1 lattice rules (a familiy of sequences used in QMC methods) is provided in [17]. The library is extensible, thanks to the decomposition of the algorithms into decoupled components, which makes it easy to implement new types of weights, new search domains, new figures of merit, etc.

7.7. Wireless Networks

Participants: Osama Arouk, Btissam Er-Rahmadi, Adlen Ksentini, Meriem Bouzouita, Pantelis Frangoudis, Yassine Hadjadj-Aoul, César Viho, Quang Pham, Gerardo Rubino.

We are continuing our activities around wireless and mobile networks, by focusing more on leveraging the current mobile and wireless architecture toward building the 5G systems.

Congestion control for M2M applications. Machine-to-Machine (M2M) communications are expected to be one of the major drivers for the future 5G network. It is expected that M2M will come up with substantial revenue growth for Mobile Network Operators (MNO), but they represent at the same time the most important challenge they are facing. For instance, a massive number of Machine-to-Machine (M2M) devices performs simultaneously Random Accesses (RA), which causes severe congestions and reduces the RA success probability. To control the Radio Access Network (RAN) overload and alleviate the congestion between M2M devices, 3GPP developed the Access Class Barring (ACB) procedure that depends on an access probability called the ACB factor. In [48][24], we first presented a simple fluid model of M2M devices' random access. This model is then used to derive an optimal regulator of the ACB factor based on nonlinear non affine control theory. The main advantages of the proposed approach are twofold. First, the proposal is fully compliant with the standard while it reduces significantly the computation and the signaling overheads. Second, it provides an efficient mean to regulate adaptively the ACB factor as it guarantees having an optimal number of M2M devices accessing concurrently to the RAN. The obtained results based on simulations show clearly the robustness of the proposed approach, and its superiority compared to existing proposals. However, such a model assumes a perfect knowledge about the number of M2M attempting the ACB and the RA, which is not possible in realistic use cases. For this reason, we proposed in [50] a system-agnostic controller, which computes the barring factor dynamically based only on the mismatch between the average number of M2M devices succedding in the RA and the optimal number of M2M which should succeed. We base our controlling algorithm in a Propostional Integral Derivative (PID)-based controller. Simulation results show that the algorithm outperforms the existing solutions by improving significantly the access success probability while minimizing radio resources' underutilization.

Different schemes were proposed in the literature to solve the congestion problem by regulating the M2M devices' opportunities of transmission. Nonetheless, as revealed in [51], these schemes turn out to be ineffective in case of heavily congested M2M networks. In fact, in such a condition, the unpredictable and increasingly accumulated number of devices cannot be blocked. This augments the risk of M2M devices' synchronized access, which may result in a congestion collapse. Consequently, we proposed, in [49], a methodology for a better estimation of the number of M2M devices attempting the access. We also proposed a novel implementation of the ACB process, which dynamically computes the ACB factor according to the network's overload conditions and includes a corrective action adapting the controller work, based on the mismatch existing between the computed and the targeted mean load. The simulation results show that the proposed algorithms allow improving considerably the estimation of the number of M2M devices' arrivals, while outperforming existing techniques.

In [32], we proposed a novel approach to deal with massive synchronous access attempts, tailored for both M2M delay-sensitive applications and energy constrained ones. The main idea behind the paper is to leverage crowd sourcing data, transmitted from the devices succeeding in the RACH procedure, to tune the access parameters, without requiring too complex techniques for the estimation of the number of attempts. Simulation results show that the proposed scheme achieves sub-optimal performance in the wireless resources' utilization while reducing significantly both the number of access attempts and the access latency for delay sensitive applications. This allows guaranteeing energy conservation.

In [44], we proposed two optimal solutions that use Device-to-Device (D2D) communications to lightweight the overhead of M2M devices on 5G networks. Each scheme has a specific objective, and aims to manage the communications between devices and eNodeBs to achieve its objective. The proposed solutions nominate the devices that should communicate using D2D communications and those that should directly communicate with eNodeBs. The first solution aims to reduce the energy consumption, whereas the second one aims to reduce the data transfer delay at the eNodeBs. The performance of the proposed schemes is evaluated via simulations and the obtained results demonstrate their feasibility and ability in achieving their design goals.

Network selection and optimization. With the explosion of mobile data traffic, the Fixed and Mobile Converged (FMC) network are being heavily required. Mobile devices have the capability of connecting simultaneously to different access networks in the FMC architecture. Access network selection becomes an issue when mobile devices are under coverage of different access networks, since a bad selection may lead to network congestion and degrade the QoE of users. In order to address this problem, in [91] we modeled and analyzed the interface selection procedure using control theory. Based on our model, we designed a controller which can send to mobile devices a network selection command calculated instantaneously for the access network selection.

Dynamic Adaptive Streaming over HTTP (DASH), with its different proprietary versions, is presently the most widely adopted technology for video delivery over the Internet. DASH offers significant advantages, enabling users to switch dynamically between different available video qualities responding to variations in the current network conditions during video playback. This is particularly interesting in wireless and mobile access networks, which present such variations in a hard to predict manner, but sometimes quite frequently. Moreover, mobile users of these networks share a common radio access link and, thus, a common bottleneck in case of congestion, which may cause user experience to degrade. In this context, the Mobile Edge Computing (MEC) emerging standard gives new opportunities to improve DASH performance, by moving IT and cloud computing capabilities down to the edge of the mobile network. In [69] and [103] we proposed a novel architecture for adaptive HTTP video streaming tailored to a MEC environment. The proposed architecture includes an adaptation algorithm running as a MEC service, aiming to relax network congestion while improving the Quality of Experience (QoE) for mobile users. Our mechanism is standards-compliant and compatible with receiver-driven adaptive video delivery algorithms, with which it cooperates in a transparent manner.

Low-rate wireless personal area networks (WPANs) (and also wireless sensor networks) suffer from many constraints. The IEEE 802.15.4 standard proposes the slotted CSMA/CA as a communication channel access mechanism with collision avoidance that takes into account the constraints of WPANs. In [22], we proposed

to introduce a data fragmentation mechanism into slotted CSMA/CA to improve a bandwidth utilisation. The novelty here is the use of the fragmentation mechanism to replace an acknowledgement frame after the transmission of the fragment and the remaining frame. The beacon frame is used to confirm the success transmission of a data fragment. To evaluate the performance of our proposition, we have developped a three dimension Markov chain which modelises the behaviour of the node using IEEE 802.15.4 with data fragmentation mechanism without using an ACK frame. The analytical results concerning the network throughput and the transmission success delay demonstrate the improvement of the bandwidth occupation.

Mobile networks' improvements. In [85], we introduced the concept of elastic bearer in Evolved Packet System (EPS), which allows the users to enhance on-demand the performance of certain applications and permits the network to efficiently manage the resource allocation taking into account the application type. In particular, the paper introduces a set of mechanisms to trigger and support bearer elasticity in EPS based on the Quality of Experience (QoE) perceived by users or based on feedback from Radio Access Network (RAN). Bearer elasticity can be attained through potential Packet Data Network/Serving Gateway (PDN/S-GW) relocation to eventually improve QoE within and beyond the mobile network operator premises. The paper also introduces a set of methods to identify and cope with a storm of requests for particular applications at densely populated areas.

One important objective of 5G mobile networks is to accommodate a diverse and ever-increasing number of user equipment (UEs). Coping with the massive signaling overhead expected from UEs is an important hurdle to tackle so as to achieve this objective. In [11], we devised an efficient tracking area list management (ETAM) framework that aims for finding optimal distributions of tracking areas (TAs) in the form of TA lists (TALs) and assigning them to UEs, with the objective of minimizing two conflicting metrics, namely paging overhead and tracking area update (TAU) overhead. ETAM incorporates an online part and an and offline one, in order to achieve its design goal. In the online part, two strategies were proposed to assign in real time, TALs to different UEs, while in the offline part, three solutions were proposed to optimally organize TAs into TALs. The performance of ETAM is evaluated via analysis and simulations, and the obtained results demonstrate its feasibility and ability in achieving its design goals, improving the network performance by minimizing the cost associated with paging and TAU.

QoE aware routing in wireless networks. This year we continued our research on QoE-based optimization routing for wireless mesh networks. First, we approximate PSQA models by explicit mathematical forms, which can be used to find the optimal or near to optimal routes. Next, the hardness of the problem is studied and decentralized algorithms are proposed. The quality of the solution, computational complexity of the proposed algorithm, and the fairness are the main concerns of this work. Several centralized approximation algorithms have been proposed in order to address the complexity and the quality of the published solutions. The results can be found in the following papers: [25],[94], [95] and [26]. However, these centralized algorithms are not appropriate in large-scale networks. Thus, a distributed algorithm is necessary as a complement of the existing centralized methods. This is currently studied at the team.

DYOGENE Project-Team

7. New Results

7.1. Fast Weak KAM Integrators for Separable Hamiltonian Systems

In [7], we consider a numerical scheme for Hamilton–Jacobi equations based on a direct discretization of the Lax–Oleinik semi–group. We prove that this method is convergent with respect to the time and space stepsizes provided the solution is Lipschitz, and give an error estimate. Moreover, we prove that the numerical scheme is a *geometric integrator* satisfying a discrete weak–KAM theorem which allows to control its long time behavior. Taking advantage of a fast algorithm for computing min–plus convolutions based on the decomposition of the function into concave and convex parts, we show that the numerical scheme can be implemented in a very efficient way.

7.2. Low Complexity State Space Representation and Algorithms for Closed Queueing Networks Exact Sampling

In [6] we consider exact sampling from the stationary distribution of a closed queueing network with finite capacities. In a recent work a compact representation of sets of states was proposed that enables exact sampling from the stationary distribution without considering all initial conditions in the coupling from the past (CFTP) scheme.This representation reduces the complexity of the one-step transition in the CFTP algorithm to O(KM^2), where K is the number of queues and M the total number of customers; while the cardinality of the state space is exponential in the number of queues. In this paper, we extend these previous results to the multiserver case. The main focus and the contribution of this work is on the algorithmic and the implementation issues. We propose a new representation, that leads to one-step transition complexity of the CFTP algorithm that is in $O(KM)$. We provide a detailed description of our matrix-based implementation. Matlab toolbox Clones (CLOsed queueing Networks Exact Sampling) can be downloaded at http://www.di.ens.fr/~rovetta/Clones

7.3. Queueing Networks with Mobile Servers: The Mean-Field Approach

In [5] we consider queueing networks which are made from servers exchanging their positions on a graph. When two servers exchange their positions, they take their customers with them. Each customer has a fixed destination. Customers use the network to reach their destinations, which is complicated by movements of the servers. We develop the general theory of such networks and establish the convergence of the symmetrized version of such a network to some nonlinear Markov process.

7.4. Distributed Randomized Control for Demand Dispatch

This work, reported in [14], concerns design of control systems for Demand Dispatch to obtain ancillary services to the power grid by harnessing inherent flexibility in many loads. The role of "local intelligence" at the load has been advocated in prior work, randomized local controllers that manifest this intelligence are convenient for loads with a finite number of states. The present work introduces two new design techniques for these randomized controllers: (i) The Individual Perspective Design (IPD) is based on the solution to a one-dimensional family of Markov Decision Processes, whose objective function is formulated from the point of view of a single load. The family of dynamic programming equation appears complex, but it is shown that it is obtained through the solution of a single ordinary differential equation. (ii) The System Perspective Design (SPD) is motivated by a single objective of the grid operator: Passivity of any linearization of the aggregate input-output model. A solution is obtained that can again be computed through the solution of a single ordinary differential equation. Numerical results complement these theoretical results.

7.5. Smart Fridge / Dumb Grid? Demand Dispatch for the Power Grid of 2020

In our previous research [31], it was argued that loads can provide most of the ancillary services required today and in the future. Through load-level and grid-level control design, high-quality ancillary service for the grid is obtained without impacting quality of service delivered to the consumer. This approach to grid regulation is called demand dispatch: loads are providing service continuously and automatically, without consumer interference. In [19] work we investigate what intelligence is required at the grid-level. In particular, does the grid-operator require more than one-way communication to the loads? Our main conclusion: risk is not great in lower frequency ranges, e.g., PJM's RegA or BPA's balancing reserves. In particular, ancillary services from refrigerators and pool-pumps can be obtained successfully with only one-way communication. This requires intelligence at the loads, and much less intelligence at the grid level.

7.6. Efficient Orchestration Mechanisms for Congestion Mitigation in Network Functions Virtualization: Models and Algorithms

Nowadays, telecommunication infrastructures are composed of property hardware operated by a single entity to offer communication services to their final users. While this architecture simplifies the design and optimization of the network equipment for specific tasks, its low degree of flexibility represents the main limitation for the evolution of the network infrastructure. For this reason, network operators and equipment manufacturers have started the standardization process of a plethora of virtualization solutions that have been individually developed in recent years for enabling the sharing of general-purpose resources and increasing the flexibility of their network architectures. Such a process has led to the specification of the Network Functions Virtualization (NFV) technology, which promises to bring about several benefits, such as reduced CAPEX and OPEX (CAPital and OPerational EXpenditure), low time-tomarket for new network services, higher flexibility to scale up and down the services according to users' demand, simple and cheap testing of new services. Nevertheless, the consolidation of the virtualization technology represents one of the main challenging problems for its success and widespread utilization in telecommunication infrastructures, which still consist of a huge set of property hardware appliances and software systems. Indeed, the sharing of the physical infrastructure among multiple virtual operators as well as the simple configuration of network services require the design of complex management mechanisms for the orchestration of the network equipment, with the final goal of dynamically adapting the infrastructure to the resource utilization.

In particular, spatio-temporal correlation of traffic demands and computational loads can result in high congestion and low network performance for virtual operators, thus leading to service level agreement breaches. In [10], we propose novel orchestration mechanisms to optimally control and mitigate the resource congestion of a physical infrastructure based on the NFV paradigm. More specifically, we analyze the congestion resulting from the sharing of the physical infrastructure and propose innovative orchestration mechanisms based on both centralized and distributed approaches, aimed at unleashing the potential of the NFV technology. In particular, we first formulate the network functions composition problem as a non-linear optimization model to accurately capture the congestion of physical resources. To further simplify the network management, we also propose a dynamic pricing strategy of network resources, proving that the resulting system achieves a stable equilibrium in a completely distributed fashion, even when all virtual operators independently select their best network configuration. Numerical results show that the proposed approaches consistently reduce resource congestion. Furthermore, the distributed solution well approaches the performance that can be achieved using a centralized network orchestration system.

7.7. Optimal Planning of Virtual Content Delivery Networks under Uncertain Traffic Demands

Content Delivery Networks (CDNs) have been identified as one of the relevant use cases where the emerging paradigm of Network Functions Virtualization (NFV) will likely be beneficial. In fact, virtualization fosters flexibility, since on-demand resource allocation of virtual CDN nodes can accommodate sudden traffic demand changes. However, there are cases where physical appliances should still be preferred, therefore we envision

a mixed architecture in between these two solutions, capable to exploit the advantages of both of them. Motivated by these reasons, in [13] we formulate a two-stage stochastic planning model that can be used by CDN operators to compute the optimal long-term network planning decision, deploying physical CDN appliances in the network and/or leasing resources for virtual CDN nodes in data centers. Key findings demonstrate that for a large range of pricing options and traffic profiles, NFV can significantly save network costs spent by the operator to provide the content distribution service.

7.8. Distributed Spectrum Management in TV White Space Networks

The radio frequency (RF) spectrum is a scarce resource that has recently become particularly critical with the increased wireless demand. For this reason, the Federal Communications Commission (FCC) has recently allowed for opportunistic access to the unused spectrum in the TV bands (also called "white space"). With opportunistic access, however, there is a need to deploy enhanced channel allocation and power control techniques to mitigate interference, including Adjacent-Channel Interference (ACI). TV White Space (TVWS) spectrum access is often investigated without taking into account ACI between the transmissions of TV Bands Devices (TVBDs) and licensed TV stations. Guard Bands (GBs) can be used to protect data transmissions and mitigate the ACI problem. Therefore, in [9] we consider a spectrum database that is administrated by a database operator, and an opportunistic secondary system, in which every TVBD is equipped with a single antenna that can be tuned to a subset of licensed channels. This can be done, for example, through adaptive channel aggregation or bonding techniques.

We investigate the distributed spectrum management problem in opportunistic TVWS systems using a game theoretical approach that accounts for adjacent channel interference and spatial reuse. TVBDs compete to access idle TV channels and select channel "blocks" that optimize an objective function. This function provides a tradeoff between the achieved rate and a cost factor that depends on the interference between TVBDs. We consider practical cases where contiguous or non-contiguous channels can be accessed by TVBDs, imposing realistic constraints on the maximum frequency span between the aggregated/bonded channels. We show that under general conditions, the proposed TVWS management games admit a potential function. Accordingly, a "best response" strategy allows us to determine the spectrum assignment of all players. This algorithm is shown to converge in a few iterations to a Nash Equilibrium (NE). Furthermore, we propose an effective algorithm based on Imitation dynamics, where a TVBD probabilistically imitates successful selection strategies of other TVBDs in order to improve its objective function. Numerical results show that our game theoretical framework provides a very effective tradeoff (close to optimal, centralized spectrum allocations) between efficient TV spectrum use and reduction of interference between TVBDs.

7.9. Straight: Stochastic Geometry and User History Based Mobility Estimation

5G is envisioned to support scalable networks and improved user experience with virtually zero latency and ultra broad-band service. Supporting unlimited seamless mobility is one of the key issues and also for network resource utilization efficiency. In [16], we focus on mobility management and user equipment (UE) speed class estimation, also known as mobility state estimation (MSE). We propose a method for estimating the UE mobility which is compliant with UE history information specifications by 3GPP (3rd Generation Partnership Project). We also exploit the impact of the environment on the UE trajectory and speed when determining UE mobility state. We evaluate the effectiveness of our algorithm using realistic mobility traces and network topology of the city of Cologne in Germany provided by the Kolntrace project. Results show that the speed classification of UEs can be achieved with much higher accuracy compared to existing legacy 3GPP LTE MSE procedures.

7.10. Mobility State Estimation in LTE

Estimating mobile user speed is a problematic issue which has significant impacts to radio resource management and also to the mobility management of Long Term Evolution (LTE) networks. In [15] introduces two

algorithms that can estimate the speed of mobile user equipments (UE), with low computational requirement, and without modification of neither current user equipment nor 3GPP standard protocol. The proposed methods rely on uplink (UL) sounding reference signal (SRS) power measurements performed at the eNodeB (eNB) and remain efficient with large sampling period (e.g., 40 ms or beyond). We evaluate the effectiveness of our algorithms using realistic LTE system data provided by the eNB Layer1 team of Alcatel-Lucent. Results show that the classification of UE's speed required by LTE can be achieved with high accuracy. In addition, they have minimal impact to the central processing unit (CPU) and the memory of eNB modem. We see that they are very practical to today's LTE networks and would allow a continuous and real-time UE speed estimation

7.11. Cell Planning for Mobility Management in Heterogeneous Cellular Networks

In small cell networks, high mobility of users results in frequent handoff and thus severely restricts the data rate for mobile users. To alleviate this problem, in [25] we propose to use heterogeneous, two-tier network structure where static users are served by both macro and micro base stations, whereas the mobile (i.e., moving) users are served only by macro base stations having larger cells; the idea is to prevent frequent data outage for mobile users due to handoff. We use the classical two-tier Poisson network model with different transmit powers (cf [43]), assume independent Poisson process of static users and doubly stochastic Poisson process of mobile users moving at a constant speed along infinite straight lines generated by a Poisson line process. Using stochastic geometry, we calculate the average downlink data rate of the typical static and mobile (i.e., moving) users, the latter accounted for handoff outage periods. We consider also the average throughput of these two types of users defined as their average data rates divided by the mean total number of users co-served by the same base station. We find that if the density of a homogeneous network and/or the speed of mobile users is high, it is advantageous to let the mobile users connect only to some optimal fraction of BSs to reduce the frequency of handoffs during which the connection is not assured. If a heterogeneous structure of the network is allowed, one can further jointly optimize the mean throughput of mobile and static users by appropriately tuning the powers of micro and macro base stations subject to some aggregate power constraint ensuring unchanged mean data rates of static users via the network equivalence property (see [36]).

7.12. Location Aware Opportunistic Bandwidth Sharing between Static and Mobile Users with Stochastic Learning in Cellular Networks

In [26] we consider location-dependent opportunistic bandwidth sharing between static and mobile downlink users in a cellular network. Each cell has some fixed number of static users. Mobile users enter the cell, move inside the cell for some time and then leave the cell. In order to provide higher data rate to mobile users, we propose to provide higher bandwidth to the mobile users at favourable times and locations, and provide higher bandwidth to the static users in other times. We formulate the problem as a long run average reward Markov decision process (MDP) where the per-step reward is a linear combination of instantaneous data volumes received by static and mobile users, and find the optimal policy. The transition structure of this MDP is not known in general. To alleviate this issue, we propose a learning algorithm based on single timescale stochastic approximation. Also, noting that the unconstrained MDP can be used to solve a constrained problem, we provide a learning algorithm based on multi-timescale stochastic approximation. The results are extended to address the issue of fair bandwidth sharing between the two classes of users. Numerical results demonstrate performance improvement by our scheme, and also the trade-off between performance gain and fairness.

7.13. Gibbsian On-Line Distributed Content Caching Strategy for Cellular Networks

In [27] 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 simple 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.

7.14. Spatial Disparity of QoS Metrics Between Base Stations in Wireless Cellular Networks

This work contributes to the line of research on the development of analytic tools for the QoS evaluation and dimensioning of operator cellular networks which is the subject of long-term collaboration between TREC/DYOGENE and Orange Labs (cf Section 8.1.1). Our focus in [8] is to explicitly characterize the disparity of quality of service (QoS) metrics between base stations in large heterogeneous wireless cellular networks. The considered QoS metrics are cell load, users' number, and user throughput. The spatial disparity of these metrics is due to the irregularity of the cells' geometry. In order to consider these irregularities, we assume a Poisson point process of base station locations, random transmission powers, and log-normal shadowing. The interdependency between the performances of the base stations is characterized by a system of load equations. The typical cell simulation model consists in resolving this system in order to find the loads and then deduce the remaining characteristics for each cell of the network. Using stochastic geometric and queueing theoretic techniques, we define the QoS averages, variances, and distributions. Inspired by the analysis of the typical cell model, several investigations lead us to propose a fully analytic approach, called mean cell model, that approximates the averages, variances, and distributions of these QoS metrics. Numerical experiments show a good agreement between the proposed approximations, simulation results, and real-life network measurements.

7.15. Stronger Wireless Signals Appear More Poisson

This work contributes to the line of research on Poisson convergence in wireless networks with strong shadowing initiated in [37], [35]. More recently, Keeler, Ross and Xia derived in [51] approximation and convergence results, which imply that the point process formed from the signal strengths received by an observer in a wireless network under a general statistical propagation model can be modeled by an inhomogeneous Poisson point process on the positive real line. The basic requirement for the results to apply is that there must be a large number of transmitters with a small proportion having a strong signal. The aim of [12] is to apply some of the main results of [51] in a less general but more easily applicable form, to illustrate how the results can apply to functions of the point process of signal strengths, and to gain intuition on when the Poisson model for transmitter locations is appropriate. A new and useful observation is that it is the stronger signals that behave more Poisson, which supports recent experimental work.

7.16. On Some Diffusion and Spanning Problems in Configuration Model

A number of real-world systems consisting of interacting agents can be usefully modelled by graphs, where the agents are represented by the vertices of the graph and the interactions by the edges. Such systems can be as diverse and complex as social networks (traditional or online), protein-protein interaction networks, internet, transport network and inter-bank loan networks. One important question that arises in the study of these networks is: to what extent, the local statistics of a network determine its global topology. This problem can be approached by constructing a random graph constrained to have some of the same local statistics as those observed in the graph of interest. One such random graph model is configuration model, which is constructed in such a way that a uniformly chosen vertex has a given degree distribution. This is the random graph which provides the underlying framework for the problems considered in the PhD thesis [3]. As our first problem,

we consider propagation of influence on configuration model, where each vertex can be influenced by any of its neighbours but in its turn, it can only influence a random subset of its neighbours. Our (enhanced) model is described by the total degree of the typical vertex and the number of neighbours it is able to influence. We give a tight condition, involving the joint distribution of these two degrees, which allows with high probability the influence to reach an essentially unique non-negligible set of the vertices, called a big influenced component, provided that the source vertex is chosen from a set of good pioneers. We explicitly evaluate the asymptotic relative size of the influenced component as well as of the set of good pioneers, provided it is non-negligible. Our proof uses the joint exploration of the configuration model and the propagation of the influence up to the time when a big influenced component is completed, a technique introduced in Janson and Luczak [48]. Our model can be seen as a generalization of the classical Bond and Node percolation on configuration model, with the difference stemming from the oriented conductivity of edges in our model. We illustrate these results using a few examples which are interesting from either theoretical or real-world perspective. The examples are, in particular, motivated by the viral marketing phenomenon in the context of social networks. Next, we consider the isolated vertices and the longest edge of the minimum spanning tree of a weighted configuration model. Using Stein-Chen method, we compute the asymptotic distribution of the number of vertices which are separated from the rest of the graph by some critical distance, say alpha. This distribution gives the scaling of the length of the longest edge of the nearest neighbour graph with the size of the graph. We then use the results of Fountoulakis [45] on percolation to prove that after removing all the edges of length greater than alpha, the subgraph obtained is connected but for the isolated vertices. This leads us to conclude that the longest edge of the minimal spanning tree and that of the nearest neighbour graph coincide with high probability. Finally, we investigate a more general question, that is, whether some ordering based on local statistics of the graph would lead to an ordering of the global topological properties, so that the bounds for more complex graphs could be obtained from their simplified versions. To this end, we introduce a convex order on random graphs and discuss some implications, particularly how it can lead to the ordering of percolation probabilities in certain situations.

7.17. Inferring Sparsity: Compressed Sensing Using Generalized Restricted Boltzmann Machines

In [23] we consider compressed sensing reconstruction from M measurements of K-sparse structured signals which do not possess a writable correlation model. Assuming that a generative statistical model, such as a Boltzmann machine, can be trained in an unsupervised manner on example signals, we demonstrate how this signal model can be used within a Bayesian framework of signal reconstruction. By deriving a message-passing inference for general distribution restricted Boltzmann machines, we are able to integrate these inferred signal models into approximate message passing for compressed sensing reconstruction. Finally, we show for the MNIST dataset that this approach can be very effective, even for $M < K$.

7.18. Recovering Asymmetric Communities in the Stochastic Block Model

In [22], we consider the sparse stochastic block model in the case where the degrees are uninformative. The case where the two communities have approximately the same size has been extensively studied and we concentrate here on the community detection problem in the case of unbalanced communities. In this setting, spectral algorithms based on the non-backtracking matrix are known to solve the community detection problem (i.e. do strictly better than a random guess) when the signal is sufficiently large namely above the so-called Kesten Stigum threshold. In this regime and when the average degree tends to infinity, we show that if the community of a vanishing fraction of the vertices is revealed, then a local algorithm (belief propagation) is optimal down to Kesten Stigum threshold and we quantify explicitly its performance. Below the Kesten Stigum threshold, we show that, in the large degree limit, there is a second threshold called the spinodal curve below which, the community detection problem is not solvable. The spinodal curve is equal to the Kesten Stigum threshold when the fraction of vertices in the smallest community is above $p^* = \frac{1}{2} - \frac{1}{2\sqrt{3}}$, so that the Kesten Stigum threshold is the threshold for solvability of the community detection in this case. However when the smallest community is smaller than p^*, the spinodal curve only provides a lower bound on the threshold for

solvability. In the regime below the Kesten Stigum bound and above the spinodal curve, we also characterize the performance of best local algorithms as a function of the fraction of revealed vertices. Our proof relies on a careful analysis of the associated reconstruction problem on trees which might be of independent interest. In particular, we show that the spinodal curve corresponds to the reconstruction threshold on the tree.

7.19. A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks

[17] 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 withoverlap (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.

7.20. Impact of Community Structure on Cascades

The threshold model is widely used to study the propagation of opinions and technologies in social networks. In this model individuals adopt the new behavior based on how many neighbors have already chosen it. In [20] we study cascades under the threshold model on sparse random graphs with community structure to see whether the existence of communities affects the number of individuals who finally adopt the new behavior. Specifically, we consider the permanent adoption model where nodes that have adopted the new behavior cannot change their state. When seeding a small number of agents with the new behavior, the community structure has little effect on the final proportion of people that adopt it, i.e., the contagion threshold is the same as if there were just one community. On the other hand, seeding a fraction of population with the new behavior has a significant impact on the cascade with the optimal seeding strategy depending on how strongly the communities are connected. In particular, when the communities are strongly connected, seeding in one community outperforms the symmetric seeding strategy that seeds equally in all communities.

7.21. Clustering from Sparse Pairwise Measurements

In [21] We consider the problem of grouping items into clusters based on few random pairwise comparisons between the items. We introduce three closely related algorithms for this task: a belief propagation algorithm approximating the Bayes optimal solution, and two spectral algorithms based on the non-backtracking and Bethe Hessian operators. For the case of two symmetric clusters, we conjecture that these algorithms are asymptotically optimal in that they detect the clusters as soon as it is information theoretically possible to do so. We substantiate this claim for one of the spectral approaches we introduce.

7.22. Limit Theory for Geometric Statistics of Clustering Point Processes

Let P be a simple, stationary, clustering point process on the d-dimensional Euclidean space, in the sense that its correlation functions factorize up to an additive error decaying exponentially fast with the separation distance. Let P_n be its restriction to a hypercube windows of volume n. We consider statistics of P_n admitting the representation as sums of spatially dependent terms $H_n = \sum_{x \in P_n} \xi(x, P_n)$, where $\xi(x, P_n)$ is a real valued (score) function, representing the interaction of x with P_n. When the score function depends locally on P_n in the sense that its radius of stabilization has an exponential tail, we establish expectation asymptotics, variance asymptotics, and central limit theorems for H_n as the volume n of the window goes to infinity.

This gives the limit theory for non-linear geometric statistics (such as clique counts, the number of Morse critical points, intrinsic volumes of the Boolean model, and total edge length of the k-nearest neighbor graph) of determinantal point processes with fast decreasing kernels, including the α-Ginibre ensembles. It also gives the limit theory for geometric U-statistics of permanental point processes as well as the zero set of Gaussian entire functions. This extends the existing literature treating the limit theory of sums of stabilizing scores of Poisson and binomial input. In the setting of clustering point processes, it also extends the results of Soshnikov [61] as well as work of Nazarov and Sodin [55].

The proof of the central limit theorem relies on a factorial moment expansion originating in Blaszczyszyn [34] to show clustering of mixed moments of the score function. Clustering extends the cumulant method to the setting of purely atomic random measures, yielding the asymptotic normality of H_n.

7.23. The Boolean Model in the Shannon Regime: Three Thresholds and Related Asymptotics

In [4] we consider a family of Boolean models, indexed by integers $n \geq 1$, where the n-th model features a Poisson point process in $\mathbb{R}^n$ of intensity $e^{n\rho_n}$ with $\rho_n \to \rho$ as $n \to \infty$, and balls of independent and identically distributed radii distributed like $\overline{X}_n\sqrt{n}$, with $\overline{X}_n$ satisfying a large deviations principle. It is shown that there exist three deterministic thresholds: τ_d the degree threshold; τ_p the percolation threshold; and τ_v the volume fraction threshold; such that asymptotically as n tends to infinity, in a sense made precise in the paper: (i) for $\rho < \tau_d$, almost every point is isolated, namely its ball intersects no other ball; (ii) for $\tau_d < \rho < \tau_p$, almost every ball intersects an infinite number of balls and nevertheless there is no percolation; (iii) for $\tau_p < \rho < \tau_v$, the volume fraction is 0 and nevertheless percolation occurs; (iv) for $\tau_d < \rho < \tau_v$, almost every ball intersects an infinite number of balls and nevertheless the volume fraction is 0; (v) for $\rho > \tau_v$, the whole space covered. The analysis of this asymptotic regime is motivated by related problems in information theory, and may be of interest in other applications of stochastic geometry.

EVA Project-Team

7. New Results

7.1. Wireless Sensor Networks

7.1.1. Deployment of Wireless Sensor Networks

Participants: Ines Khoufi, Pascale Minet, Anis Laouiti.

In 2016, we studied two types of deployment for wireless sensor networks:

- those ensuring full area coverage and network connectivity;
- those covering some given Points of Interest (PoI) and ensuring network connectivity.

Deployment of sensor nodes to fully cover an area has caught the interest of many researchers. However, some simplifying assumptions are adopted such as the knowledge of obstacles, a centralized algorithm... To cope with these drawbacks, we propose OA-DVFA (Obstacles Avoidance Distributed Virtual Forces Algorithm), a self-deployment algorithm to ensure full area coverage and network connectivity. This fully distributed algorithm is based on virtual forces to move sensor nodes. We show how to avoid the problem of node oscillations and to detect the end of the deployment in a distributed way. We evaluate the impact of the number, shape and position of obstacles on the coverage rate, the distance traveled by all nodes and the number of active nodes. Simulation results show the very good behavior of OA-DVFA. This work done in collaboration with Anis Laouiti has been presented at the CCNC 2016 conference [35].

We also focus on wireless sensor networks deployed to cover some given Points of Interest (PoIs), achieve connectivity with the sink and be robust against link and node failures. The Relay Node Placement problem (RNP) consists in minimizing the number of relays needed and the maximum length of the paths connecting each PoI with the sink. We propose a solution that determines the positions of relay nodes based on the virtual grid computed by the optimal deployment for full area coverage. We compare our solution with two different solutions based respectively on (1) the straight line that builds the shortest path between each PoI and the sink, (2) the Steiner point that connects PoIs together. We then extend these algorithms to achieve k-connectivity. Our solution outperforms the Steiner points solution in terms of maximum path length on the one hand, and the straight line solution in terms of total number of relay nodes deployed on the other hand. We also apply our solution in an area containing obstacles and show that it provides very good performances. This study has been presented at the MASS 2016 conference [34].

7.1.2. Path Planning of Mobile Wireless Nodes Gathering Data

Participants: Ines Khoufi, Pascale Minet, Nadjib Achir.

Mobile wireless nodes in charge of collecting data from static wireless sensor nodes constitute a very attractive solution for wireless sensor networks, WSNs, where the application requirements in terms of node autonomy are very strong unlike the requirement in terms of latency. Mobile nodes allow wireless sensor nodes to save energy.

In 2016 we focused on the path planning problem of mobile wireless nodes gathering data according two different objectives:

- to ensure the monitoring of a given area;
- to visit some given Points of Interest (PoI) in a delay less than a given latency.

For the first objective, we are interested in area monitoring using Unmanned Aerial Vehicles (UAVs). Basically, we propose a path planning approach for area monitoring where UAVs are considered as mobile collectors. The area to be monitored is divided into cells. The goal is to determine the path of each UAV such that each cell is covered by exactly one UAV, fairness is ensured in terms of the number of cells visited by each UAV and the path of each UAV is minimized. To meet our goal, we proceed in two steps. In the first step, we assign to each UAV the cells to visit. In the second step, we optimize the path of each UAV visiting its cells. For the first step, we propose two solutions. The first solution is based on cluster formation, each cluster is made up of the set of cells monitored by a same UAV. The second solution is based on game theory and uses coalition formation to determine the cells to be monitored by each UAV. In the second step and for both solutions, we propose to apply optimization techniques to minimize the path of each UAV that visits all its cells. This study done in collaboration with Nadjib Achir was presented at the PEMWN 2016 conference [32].

For the second objective, we use game theory to model the problem. Game theory is often used to find equilibria where no player can unilaterally increase its own payoff by changing its strategy without changing the strategies of other players. In this paper, we propose to use coalition formation to compute the optimized tours of mobile sinks in charge of collecting data from static wireless sensor nodes. The associated coalition formation problem has a stable solution given by the final partition obtained. However, the order in which the players play has a major impact on the final result. We determine the best order to minimize the number of mobile sinks needed. We evaluate the complexity of this coalition game in terms of the number of rounds and the processing time needed to get convergence, as well as the impact of the number of collect points on the number of mobile sinks needed and on the maximum tour duration of these mobile sinks. In addition, we show how to extend the coalition game to support different latencies for different types of data. Finally, we formalize our problem as a multi-objective optimization problem. We compare the coalition game with a genetic algorithm: for 20 nodes to visit, the coalition game requires a processing time 327 times less than the genetic algorithm. The coalition game provides a scalable solution. These results have been presented at the IPCCC 2016 conference. This work was done in cooperation Mohamed-Amine Koulali and Abdellatif Kobbane [33].

7.1.3. Centralized Scheduling in TSCH-based Wireless Sensor Networks

Participants: Erwan Livolant, Pascale Minet, Thomas Watteyne.

Scheduling in an IEEE802.15.4e TSCH(Time Slotted Channel Hopping 6TiSCH) low-power wireless network can be done in a centralized or distributed way. When using centralized scheduling, a scheduler installs a communication schedule into the network. This can be done in a standards-based way using CoAP. In this study, we compute the number of packets and the latency this takes, on real-world examples. The result is that the cost is very high using today's standards, much higher than when using an ad-hoc solution such as OCARI. We conclude by making recommendations to drastically reduce the number of messages and improve the efficiency of the standardized approach.

7.1.4. Using an IEEE 802.15.4e TSCH network

Participants: Ines Khoufi, Pascale Minet, Erwan Livolant, Thomas Watteyne.

Most wireless sensor networks that are currently deployed use a technology based on the IEEE 802.15.4 standard. However, this standard does not meet all requirements of industrial applications in terms of latency, throughput and robustness. That is why the IEEE 802.15.4e amendment has been designed, including the Time Slotted Channel Hopping (TSCH) mode.

In 2016, we evaluated the time needed for a joining node to detect beacons advertising the TSCH network. This time may be long due to channel hopping in the TSCH network. The beacon advertising policy is left unspecified by the standard. We propose DBA, a Deterministic Beacon Advertising algorithm. DBA ensures a collision-free and regular transmission of beacons on all the frequencies used by the TSCH network. DBA outperforms two solutions already published that are Random Horizontal and Random Vertical. Some results have been presented as a poster at the IPCCC 2016 conference [48].

The medium access in a TSCH network is ruled by a schedule that determines for each pair (slot offset, channel offset) the transmitting node(s) and the receiving node(s). Each node in the TSCH network must have this schedule. The question is how to install it on all nodes. We proposed and evaluated different ways of installing a schedule in a TSCH network, comparing them in terms of the number of messages required. This study has been presented at the AdHocNow 2016 conference [36].

7.1.5. The OCARI Wireless Sensor Network

Participants: Erwan Livolant, Pascale Minet, Mohammed Tahar Hammi.

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, which should authenticate them before they join the network. OCARI is a wireless sensor network technology providing optimized protocols in order to reduce the energy consumption.

To enhance OCARI security and ensure a robust authentication of devices, we propose a strong authentication method based on the One Time Password algorithm and deployed at the MAC layer. This method is specially designed to be implemented on devices with low storage and computing capacities. This work has been done in collaboration with Mohammed Tahar Hammi from Telecom ParisTech and presented at the PEMWN 2016 conference [30].

We also evaluated the performances of the building of an OCARI network. The goal was to identify the most time consuming steps among node association, neighborhood discovery, routing tree building, stabilization of the routing tree and node coloring.

7.1.6. 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. This study was published in [21]. 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.

7.1.7. Massive MIMO Cooperative Communications for Wireless Sensor Networks

Participants: Nadjib Achir, Paul Muhlethaler.

This work is a 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.

7.2. Machine Learning for an efficient and dynamic management of network resources and services

7.2.1. Machine Learning in Networks

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet.

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. The idea developed hereafter consists in using prediction techniques 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.

We consider various prediction methods, called 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 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. This study has been presented at the IWCMC 2016 conference [18].

In the paper presented at the WiMob 2016 conference [18], we apply these prediction methods to caching. We first selected the best experts in charge of predicting the popularity of video contents in real traces of YouTube. We tuned the parameters of the DES expert. We proved that the well-known LFU caching strategy can also be considered as a prediction based strategy on the Basic expert. Simulation results show that the DES prediction-based caching strategy provides similar Hit Ratio to the well-known LFU caching strategy. These results are usually close to the optimal ones that can be achieved only when knowing in advance the popularity of each video content for the next day, which is an unrealistic assumption. The exception occurs when a content whose popularity was very poor becomes suddenly very popular with millions of solicitations. In such a case, the accuracy of prediction methods becomes poor. This opens a research direction where the knowledge of societal events is taken into account to improve the prediction.

7.3. Protocols and Models for Wireless Networks - Application to VANETs

7.3.1. Protocols for VANETs

7.3.1.1. An Infrastructure-Free Slot Assignment Algorithm for Reliable Broadcast of Periodic Messages in VANETs

Participants: Mohamed Elhadad, Paul Muhlethaler, Anis Laouiti.

We have designed a novel **D**istributed **T**DMA based **MAC** protocol, named DTMAC, developed specifically for a highway scenario. DTMAC is designed to provide the efficient delivery of both periodic and event-driven safety messages. The protocol uses the vehicles' location and a new slot reuse concept to ensure that vehicles in adjacent areas have a collision-free schedule. Simulation results and analysis in a highway scenario have been carried out to evaluate the performance of DTMAC and compare it with the VeMAC protocol.

We propose a completely distributed and infrastructure free TDMA scheduling scheme which exploits the linear feature of VANET topologies. The vehicles' movements in a highway environment are linear due to the fact that their movements are constrained by the road topology. Our scheduling mechanism is also based on the assumption that each road is divided into N small fixed areas, denoted by $x_i, i = 1, ..., N$ (see Figure 1). Area IDs can be easily derived using map and GPS Information.

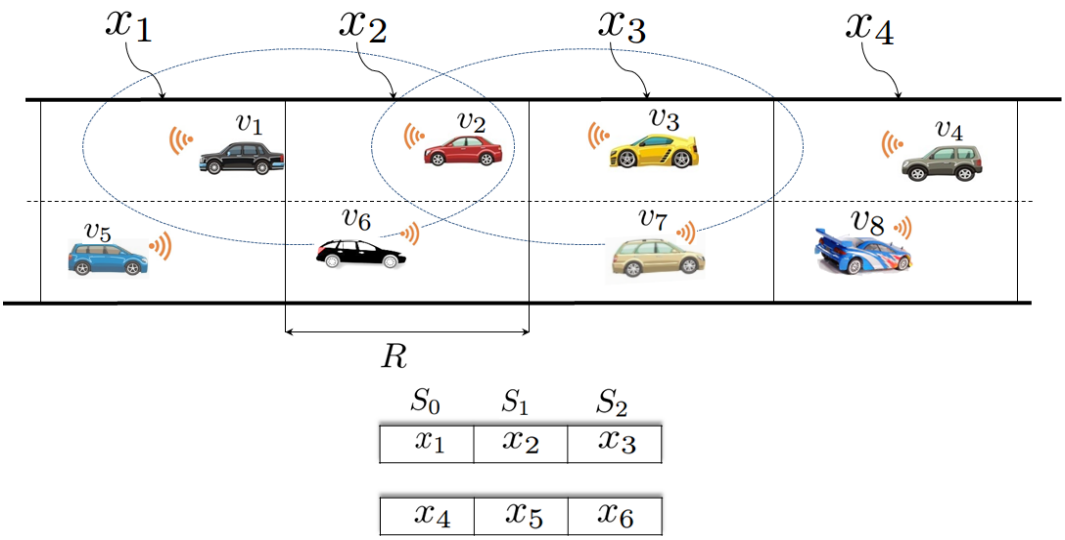

Figure 1. TDMA slots scheduling principle.

The time slots in each TDMA frame are partitioned into three sets S_0, S_1 and S_2 associated with vehicles in three contiguous areas: x_i, x_{i+1} and x_{i+2}, respectively (see Figure 1). Each frame consists of a constant number of time slots, denoted by τ and each time slot is of a fixed time duration, denoted by s. Each vehicle can detect the start time of each frame as well as the start time of a time slot. In the VANET studied, all the vehicles are equipped with a GPS and thus the one-Pulse-Per-Second (1PPS) signal that a GPS receiver gets from GPS satellites can be used for slot synchronization.

To prevent collisions on the transmission channel, our TDMA scheduling mechanism requires that every packet transmitted by any vehicle must contain additional information, called Frame Information (FI). For the frame, this field gives the status of the slot (Idle, Busy, Collision) and the ID of the vehicles accessing each slot with the characteristic of the data sent: periodic or event-driven safety messages.

The simulation results show that, compared to VeMAC which is the reference in terms of TDMA protocols for VANETs, DTMAC provides a lower rate of access and merging collisions, which results in significantly improved broadcast coverage. For further details see [27].

7.3.1.2. 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 2 , 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).

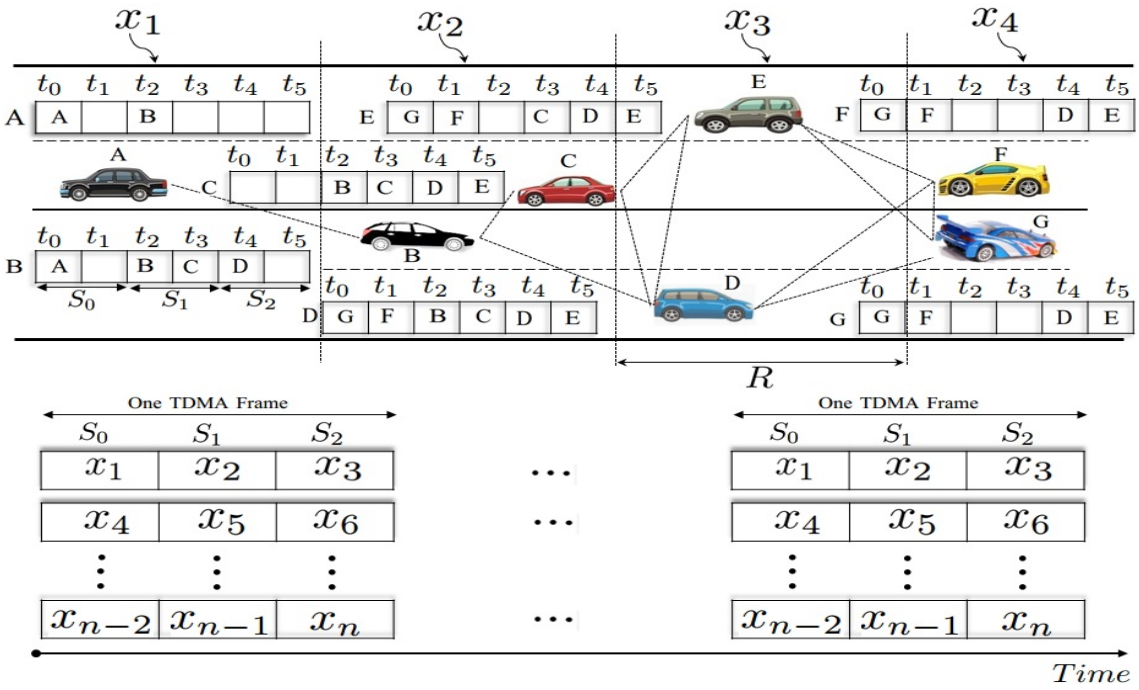

Figure 2. VANET network using DTMAC scheduling scheme.

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 2 . 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 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.*

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.

7.3.1.3. CTMAC: a Centralized TDMA for VANETs

Participants: Mohamed Elhadad Or Hadded, Paul Muhlethaler, Anis Laouiti.

We have designed an infrastructure-based TDMA scheduling scheme which exploits the linear feature of VANET topologies. The vehicles' movements in a highway environment are linear due to the fact that their movements are constrained by the road topology. Our scheduling mechanism is also based on the assumption that the highway is equipped with some RSUs (i.e. one RSU for each $2 \times R\ meters$, where R is the communication range). Note that each area is covered by one RSU installed on the side of the highway and in the middle of the corresponding area. The time slots in each TDMA frame are partitioned into two sets S_1, S_2 associated with vehicles in two adjacent RSU areas (see Figure 3). Each frame consists of a constant number of time slots, denoted by τ and each time slot is of a fixed time duration, denoted by s. Each vehicle can detect the start time of each frame as well as the start time of a time slot.

The CTMAC scheduling mechanism uses a slot reuse concept to ensure that vehicles in adjacent areas covered by two RSUs have a collision-free schedule. The channel time is partitioned into frames and each frame is further partitioned into two sets of time slots S_1 and S_2. These sets are associated with vehicles moving in the adjacent RSU areas. These sets of time slots are reused along the highway in such a way that no vehicles belonging to the same set of two-hop neighbors using the same time slot. As shown in Figure 3 , the vehicles in the coverage area of RSU_1 and those in the coverage area of RSU_2 are accessing disjoint sets of time slots. As a result, the scheduling mechanism of CTMAC can decrease the collision rate by avoiding the inter-RSUs interference without using any complex band. Each active vehicle keeps accessing the same time slot on all subsequent frames unless it enters another area covered by another RSU or a merging collision problem occurs. Each vehicle uses only its allocated time slot to transmit its packet on the control channel.

The simulation results reveal that CTMAC significantly outperforms the VeMAC and ADHOC MAC protocols. in terms of transmission collisions and the overhead required to create and maintain the TDMA schedules, see [28].

7.3.1.4. A Flooding-Based Location Service in VANETs

Participants: Selma Boumerdassi, Paul Muhlethaler.

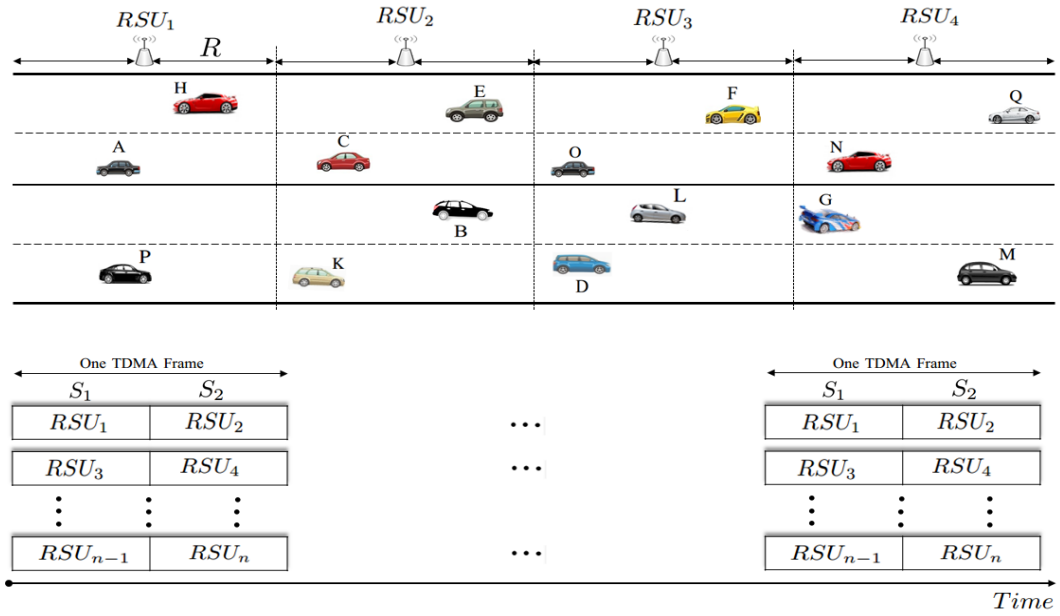

Figure 3. TDMA slots scheduling mechanism of CTMAC

This work was 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.

7.3.2. Models for Wireless Networks and VANETs

7.3.2.1. Performance analysis of IEEE 802.11 broadcast schemes with different inter-frame spacings

Participants: Younes Bouchaala, Paul Muhlethaler, Nadjib Achir.

This work has been 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.

7.3.2.2. Model of IEEE 802.11 Broadcast Scheme with Infinite Queue

Participant: Paul Muhlethaler.

This work has been in collaboration with Guy Fayolle (Inria RITS).

We have analyzed the so-called back-off technique of the IEEE 802.11 protocol in broadcast mode with waiting queues. In contrast to existing models, packets arriving when a station (or node) is in back-off state are not discarded, but are stored in a buffer of infinite capacity. As in previous studies, the key point of our analysis hinges on the assumption that the time on the channel is viewed as a random succession of transmission slots (whose duration corresponds to the length of a packet) and mini-slots during which the back-off of the station is decremented. These events occur independently, with given probabilities. The state of a node is represented by a two-dimensional Markov chain in discrete-time, formed by the back-off counter and the number of packets at the station. Two models are proposed both of which are shown to cope reasonably well with the physical principles of the protocol. The stability (ergodicity) conditions are obtained and interpreted in terms of maximum throughput. Several approximations related to these models are also discussed. The results of this study are in [2].

7.3.2.3. Model and optimization of CSMA

Participants: Younes Bouchaala, Paul Muhlethaler, Nadjib Achir.

This work has been 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. 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.

7.3.2.4. Adaptive CSMA

Participants: Nadjib Achir, Younes Bouchaala, Paul Muhlethaler.

This work has been 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.

FUN Project-Team

7. New Results

7.1. Routing

Participants: Nathalie Mitton, Mouna Masmoudi.

Geographic routing is an attractive routing strategy in wireless sensor networks. It works well in dense networks, but it may suffer from the void problem. For this purpose, a recovery step is required to guarantee packet delivery. Face routing has widely been used as a recovery strategy since proved to guarantee delivery. However, it relies on a planar graph not always achievable in realistic wireless networks and may generate long paths. In [23], [12], we propose GRACO, a new geographic routing algorithm that combines a greedy forwarding and a recovery strategy based on swarm intelligence. During recovery, ant packets search for alternative paths and drop pheromone trails to guide next packets within the network. GRACO avoids holes and produces near optimal paths. Simulation results demonstrate that GRACO leads to a significant improvement of routing performance and scalability when compared to the literature algorithms.

GRACO has first been designed in the general case. We then studied its applicability to the Virtual Power Plants and their specific data packets with different priorities [23], [12]. Indeed, the Smart Grid (SG) incorporates communication networks to the conventional electricity system in order to intelligently integrate distributed energy resources (DERs) and allow for demand side management. The move to Smart grid in developing countries has to cope with great disparities of ICT infrastructures even within the same city. Besides, individual DERs are often too small to be allowed access to energy market, likewise power utilities are unable to effectively control and manage small DERs. We propose the use of affordable and scalable wireless communication technology to aggregate geographically sparse DERs into a single virtual power plant. The enrollment of prosumers in the VPP is conditional to financial performance of the plant. Thus, the VPPs are dynamic and are expected to scale up as more and more prosumers are attracted by their financial benefits. the communication network has to follow this progression and therefore to be scalable and rapidly deploy-able. We present a routing algorithm for data communication within the VPP to support centralized, decentralized or fully distributed control of the VPP's DERs.

Based on this study, we adapted GRACO so it can fit the specific cases of Smart Grid [23], [12] and more specifically to the Neighbor Area Networks (NAN) of Smart Grids, or distribution segment of the power system in the smart grid (SG). The deployment of ICT to support conventional grid will solve legacy problems that used to prevent implementation of smart services such as smart metering, demand side management or the integration of Distributed Energy Resources (DERs) within the smart grid. We demonstrate the effectiveness of GRACO in terms of scalability, peer-to-peer routing, end-to-end delay and delivery rate.

In another context, we made the observation that typical betweenness centrality metrics neglect the potential contribution of nodes that are near but not exactly on shortest paths. The idea of [35] is to give more value to these nodes. We propose a weighted betweenness centrality, a novel metric that assigns weights to nodes based on the stretch of the paths they intermediate against the shortest paths. We compare the proposed metric with the traditional and the distance-scaled betweenness metrics using four different network datasets. Results show that the weighted betweenness centrality pinpoints and promotes nodes that are underestimated by typical metrics, which can help to avoid network disconnections and better exploit multipath protocols.

7.2. Cloud and IoT

Participants: Valeria Loscri, Nathalie Mitton, Riccardo Petrolo.

Innovative and effective solutions to the fragmentation issues in the Internet of Things (IoT) landscape have been designed and proof of concept have been implemented to show the feasibility and effectiveness of the Cloud of Things (CoT) paradigm. In other words, we have focused on the convergence of Web semantic technologies and the Cloud computing concept as key enabler of an horizontal integration of various IoT applications and platforms [21]. The heterogeneity has to be considered not only in terms of applications and platforms, but another "type of heterogeneity" that deserves to be considered and analyzed is based on different devices and their interoperability.

A feasible solution to make different and heterogeneous devices to "interoperate" is based on the exploitation of a gateway. In particular, we have considered a Gateway-as-a-Service (Gaas) in [36], where we have shown that it is an efficient and lightweight device, which can be shared between several final users. Through the container virtualization technologies, we have been able to show how several platform requirements can be met, in a context where constrained devices have been considered. This study has demonstrated the Gateway-as-a-Service (GaaS) effectiveness and its exploitability in several IoT contexts, such as smart home, buildings, farms, agriculture environments, etc.

A different and complementary, to the previous solutions, perspective of IoT paradigm is represented by the management of the huge amount of data that have to be treated in the different IoT based applications. In [45], an infer algorithm has been proposed and more specifically an Bayesian Inference Approach (BIA) with the amin objective to avoid the transmission of high spatio-temporal correlated data.

7.3. Resource management in FUN

Participants: Cristina Cano Bastidas, Valeria Loscri, Simon Duquennoy.

A standard solution for reliable low-power mesh networks was defined in IEEE802.15.4e-2012, through the new MAC layer TSCH. TSCH (Time-Slotted Channel Hopping) provides a globally synchronized network that enables scheduling and channel hopping. Our review paper [28] details the TSCH technology as well as the 6LoWPAN and 6TiSCH protocols. It gathers authors from all major open-source IoT OSes: Contiki, OpenWSN, RIOT and TinyOS. The paper presents architectural considerations when it comes to implementing portable TSCH stacks, and presents preliminary evaluation results.

TSCH networks require global synchronization. The more precise the synchronization, the more energy-efficient the network. We address the challenge of reaching micro-second time synchronization over multiple hops in TSCH networks [31], at low power. The key idea is to use two crystal oscillators, one at low-frequency for low-power timekeeping, one at high-frequency for intra-slot precision. Along with adaptive drift compensation, this method is proven effective through an experimental assessment.

Beaconing is usually employed to allow network discovery and to maintain synchronisation in mesh networking protocols, such as those defined in the IEEE 802.15.4e and IEEE 802.11s standards. Thus, avoiding persistent or consecutive collisions of beacons is crucial in order to ensure correct network operation. Beacons are also used in receiver-initiated medium access protocols to advertise that nodes are awake. Consequently, effective beacon scheduling can enable duty-cycle operation and reduce energy consumption. We propose [56] a completely decentralised and low-complexity solution based on learning techniques to schedule beacon transmissions in mesh networks. We show the algorithm converges to beacon collision-free operation almost surely in finite time and evaluate converge times in different mesh network scenarios.

In [54] we focus on new methods, architectures, and applications for the management of Cyber Physical Objects (CPOs) in the context of the Internet of Things (IoT). The book covers a wide range of topics related to CPOs, such as resource management, hardware platforms, communication and control, and control and estimation over networks. It also discusses decentralized, distributed, and cooperative optimization as well as effective discovery, management, and querying of CPOs. Other chapters outline the applications of control, real-time aspects, and software for CPOs and introduce readers to agent-oriented CPOs, communication support for CPOs, real-world deployment of CPOs, and CPOs in Complex Systems. There is a focus on the importance of application of IoT technologies for Smart Cities.

Finally, we address software security and in particular the challenge of formally verifying the source code of IoT OSes. This is the topic of the yet-to-be-started H2020 VESSEDIA project. Our preliminary study [32] demonstrated the feasibility of applying Frama-C to a memory allocation module of the Contiki OS.

7.4. Smart Cities

Participants: Nathalie Mitton, Valeria Loscri, Riccardo Petrolo.

Smart City represents one of the most promising, prominent and challenging Internet of Things (IoT) applications, but recent ICT trends suggest more and more that cities could also benefit from Cloud computing. The convergence of IoT paradigm and Cloud computing technology, can play a fundamental role for developing of highly level and organized cities form an ICT point of view, but it is of paramount importance to deal a critical analysis to identify the issues and challenges deriving from this synergy.

A novel perspective that we have considered as key factor for the realization of Future Internet is the role of the interconnected objects as active entities in the context of the networked systems [52]. With this perspective in mind, we have proposed CACHACA [43], a ranking mechanism for Sensor Networks that facilitate the discovery of services provided by each network element. Discovery functionality has been also considered in the context of VITAL project, since effective and accurate mechanisms to discover Inter-Connected Objects (ICOs) and new services represents a sine qua non condition to have effective exploration of data-sources that are appropriate for a specific business context as defined by an end-user [42] [11].

On the other hand, a Smart City is a kind of ecosystem characterized with different IoT solutions that have to cooperate and coexist and is in continuous expansion. In order to face with the integration and interoperability challenges of this ecosystem, we have considered VITAL-OS architecture that can monitor, visualize, and control all the operations of a city [44].

7.5. RFID

Participants: Nathalie Mitton, Abdoul Aziz Mbacke.

One of the devices under consideration by the FUN team is RFID. One of the main issues to widely deploy RFID reader is reader-to-reader collision. Indeed, when the electromagnetic fields of the readers overlap, a collision occurs on the tag laying in the overlapping section and cannot be read. Numerous protocols have been proposed to attempt to reduce them, but, remaining reading errors still heavily impact the performances and fairness of dense RFID deployments. In [33], [18] we introduce a new Distributed Efficient & Fair Anticollision for RFID (DEFAR) protocol. It reduces both monochannel and multichannel collisions as well as interference by a factor of almost 90% in comparison with the best state of the art protocols. The fairness of the medium access among the readers is improved to a 99% level. Such improvements are achieved applying a TDMA-based "server-less" approach and assigning different priorities to readers depending on their behavior over precedent rounds. A distributed reservation phase is organized between readers with at least one winning reader afterwards. Then, multiple reading phases occur within a single frame in order to obtain fast coverage and high throughput. The use of different reader priorities based on reading behaviors of previous frames also contributes to improve both fairness and efficiency. Simulation results show the robustness of the proposed solution in terms of different metrics such collision avoidance, fairness and coverage and in comparison with a centralized literature solution.

In order to ensure collision-free reading, a scheduling scheme is needed to read tags in the shortest possible time. We study in [37] this scheduling problem in a stationary setting and the reader minimization problem in a mobile setting. We show that the optimal schedule construction problem is NP-complete and provide an approximation algorithm that we evaluate our techniques through simulation.

7.6. Interferences and failures management

Participants: Nathalie Mitton, Viktor Toldov, Valeria Loscri, Simon Duquennoy.

In the recent years, the Machine-to-Machine (M2M) paradigm together with the integration of wireless sensors networks with the generic infrastructure via $6LoWPAN$ require the implementation of ad hoc communication protocols at the Medium Access Control layer, that do not depend on pre-existing infrastructure. Channel hopping concept has more and more gained consensus as a viable and effective solution for wireless MAC layer coordination with time-synchronized channel hopping (TSCH). In [24] we propose a decentralized multichannel MAC coordination framework (DT-SCS) leveraging the concept of *pulse-coupled oscillators* at the MAC layer. In DT-SCS, nodes randomly join a channel and are automatically spread across the available channels. The nodes then achieve PCO-based coordination via the periodic transmission of beacon packets at the MAC layer. As such, for channels with an equal number of nodes, DT-SCS converges to synchronized beacon packet transmission at the MAC layer in a completely uncoordinated manner. In order to combat the well-know phenomenon of Cross-Technology Interference (CTI) a cross-layer mechanism, CrossZig, has been implemented in [39], based on the exploitation of information at the physical layer in order to detect the presence of CTI in a corrupted packet.

A different perspective of the interference management has been considered in [47] and [41], where a novel solution to allow to secondary users the access of allocated spectrum has been proposed. The study has been based on the major consideration that a big bottleneck in cognitive radio systems is based on finding the best available channel as fast as possible.

A totally different approach to face the enormous quantity of data generated by IoT devices, is to try to reduce the sending of useless data, based on the adoption of effective predictive approaches.

In [50] we have considered the concept of high spatio-temporal correlated data and we have proposed a Belief Propagation (BP) algorithm to derive methods to drastically reduce the number of transmitted messages, by keeping an high accuracy in terms of global information.

Together with interference management approaches it is also important to figure out tools to support network operator for mitigation of the impact of failures on their infrastructures. The need of advanced Network Planning and Management Tool (NPMT) has been considered in [30].

7.7. Vehicular Networks

Participants: Nathalie Mitton, Valeria Loscri.

[27] studies the information delivery delay analysis for roadside unit deployment in a vehicular ad hoc network (VANET) with intermittent connectivity. A mathematical model is developed to describe the relationship between the average delay for delivering road condition information and the distance between two neighbor RSUs deployed along a road. The derived mathematical model considers a straight highway scenario where two RSUs are deployed at a distance without any direct connection and vehicles are sparsely distributed on the road with road condition information randomly generated between the two neighbor RSUs. Moreover, the model takes into account the vehicle speed, the vehicle density, the likelihood of an incident, and the distance between two RSUs. The effectiveness of the derived mathematical model is verified through simulation results. Given the information delivery delay constraint of a time-critical application, this model can be used to estimate the maximum distance allowed between two neighbor RSUs, which can provide a reference for the deployment of RSUs in such scenarios.

But Vehicular Networks can also convey social networks. In [53], we survey recent literature on Vehicular Social Networks that are a particular class of vehicular ad hoc networks, characterized by social aspects and features. Starting from this pillar, we investigate perspectives of next generation vehicles under the assumption of social networking for vehicular applications (i.e., safety and entertainment applications). This paper plays a role as a starting point about socially-inspired vehicles, and main related applications, as well as communication techniques. Vehicular communications can be considered as the "first social network for automobiles", since each driver can share data with other neighbors. As an instance, heavy traffic is a common occurrence in some areas on the roads (e.g., at intersections, taxi loading/unloading areas, and so on); as a consequence, roads become a popular social place for vehicles to connect to each other. Human factors are then involved in vehicular ad hoc networks, not only due to the safety related applications, but also for entertainment

purpose. Social characteristics and human behavior largely impact on vehicular ad hoc networks, and this arises to the vehicular social networks, which are formed when vehicles (individuals) "socialize" and share common interests. This survey describes the main features of vehicular social networks, from novel emerging technologies to social aspects used for mobile applications, as well as main issues and challenges. Vehicular social networks are described as decentralized opportunistic communication networks formed among vehicles. They exploit mobility aspects, and basics of traditional social networks, in order to create novel approaches of message exchange through the detection of dynamic social structures. An overview of the main state-of-the-art on safety and entertainment applications relying on social networking solutions is also provided.

Cognitive Radio (CR) together with vehicular networks have been considered with an integrated and synergic perspective in [55], since CR technology is foreseen as a very effective tool to improve the communication efficiency in the context of vehicular networked systems.

7.8. Self-deployment and coverage

Participants: Nathalie Mitton, Tahiry Razafindralambo.

Controlled mobility in wireless sensor networks can provide many services. One of the most challenging one is coverage. Coverage can be needed either for monitoring control of specific area or point of interest or for deploying a communication network. This latter case is required for instance in post-disaster situations. In post-disaster scenarios, for example, after earthquakes or floods, the traditional communication infrastructure may be unavailable or seriously disrupted and overloaded. Therefore, rapidly deployable network solutions are needed to restore connectivity and provide assistance to users and first responders in the incident area. This work surveys the solutions proposed to address the deployment of a network without any a priori knowledge about the communication environment for critical communications. The design of such a network should also allow for quick, flexible, scalable, and resilient deployment with minimal human intervention. We survey this kind of approaches in [20].

In [13], we present a decentralized deployment algorithm for wireless mobile sensor networks focused on deployment Efficiency, connectivity Maintenance and network Reparation (EMR). We assume that a group of mobile sensors is placed in the area of interest to be covered, without any prior knowledge of the environment. The goal of the algorithm is to maximize the covered area and cope with sudden sensor failures. By relying on the locally available information regarding the environment and neighborhood, and without the need for any kind of synchronization in the network, each sensor iteratively chooses the next-step movement location so as to form a hexagonal lattice grid. Relying on the graph of wireless mobile sensors, we are able to provide the properties regarding the quality of coverage, the connectivity of the graph and the termination of the algorithm. We run extensive simulations to provide compactness properties of the deployment and evaluate the robustness against sensor failures. We show through the analysis and the simulations that EMR algorithm is robust to node failures and can restore the lattice grid. We also show that even after a failure, EMR algorithm call still provide a compact deployment in a reasonable time.

Routing a fleet of robots in a known surface is a complex problem. It consists in the determination of the exact trajectory each robot has to follow to collect information. The objective pursued in [38] is to maximize the exploration of the given surface. To ensure the robots can execute the mission in a collaborative manner, connectivity constraints are considered. These constraints guarantee that robots can communicate among each other and share the collected information. Moreover, the trajectories of the robots need to respect autonomy constraints.

7.9. Controlled Mobility for additional services

Participants: Nathalie Mitton, Valeria Loscri, Jean Cristanel Razafimandimby Anjalalaina.

Wireless sensor networks (WSNs) have been of very high interest for the research community since years, but most of the time, the mobility of nodes have been considered as an obstacle to overcome. In the contrary, in have tried to adopt another perspective and see it as an asset to exploit to provide additional services.

In [19], we leverage on the ability of mobile nodes to replace or recharge static sensors. Two main approaches can be identified that target this objective: either “recharging” or “replacing” the sensor nodes that are running out of energy. Of particular interest are solutions where mobile robots are used to execute the above mentioned tasks to automatically and autonomously maintain the WSN, thus reducing human intervention. Recently, the progress in wireless power transfer techniques has boosted research activities in the direction of battery recharging, with high expectations for its application to WSNs. Similarly, also sensor replacement techniques have been widely studied as a means to provide service continuity in the network. Objective of [19] is to investigate the limitations and the advantages of these two research directions. Key decision points must be identified for effectively supporting WSN self-maintenance: (i) which sensor nodes have to be recharged/replaced; (ii) in which order the mobile robot is serving (i.e., recharging/replacing) the nodes and by following which path; (iii) how much energy is delivered to a sensor when recharged. The influence that a set of parameters, relative to both the sensors and the mobile robot, on the decisions will be considered. Centralized and distributed solutions are compared in terms of effectiveness in prolonging the network lifetime and in allowing network self-sustainability. The performance evaluation in a variety of scenarios and network settings offers the opportunity to draw conclusions and to discuss the boundaries for one technique being preferable to the other.

Mobility can also help for collecting data in wireless sensor networks [29]. The sensor data collection problem using data mules have been studied fairly extensively in the literature. However, in most of these studies, while the mule is mobile, all sensors are stationary. The objective of most of these studies is to minimize the time needed by the mule to collect data from all the sensors and return to the data collection point, from where it embarked on its data collection journey. The problem studied in this paper has two major differences with the earlier studies. First, in this study we assume that both the mule as well as the sensors are mobile. Second, we do not attempt to minimize the data collection time. Instead we minimize the number of mules that will be needed to collect data from all the sensors, subject to the constraint that the data collection process has to be completed within some pre-specified time. We show that the mule minimization problem is NP-Complete and provide a solution by first transforming it to a generalized version of the minimum flow problem in a network and then solving it optimally using Integer Linear Programming. Finally, we evaluate our algorithms through extensive simulation and present the results.

Internet of Robotic Things (IoRT) is a new concept introduced for the first time by ABI Research. Unlike the Internet of Things (IoT), IoRT provides an active sensorization and is considered as the new evolution of IoT.

This new concept will bring new opportunities and challenges, while providing new business ideas for IoT and robotics’ entrepreneurs.

In [46], we focus particularly on two issues: (i) connectivity maintenance among multiple IoRT robots, and (ii) their collective coverage.

We propose (i) IoRT-based, and (ii) a neural network control scheme to efficiently maintain the global connectivity among multiple mobile robots to a desired quality-of-service (QoS) level. The proposed approaches will try to find a trade-off between collective coverage and communication quality.

The IoT-based approach is based on the computation of the algebraic connectivity and the use of virtual force algorithm.

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, connectivity, and energy consumption.

7.10. New and other communication paradigms

Participants: Nathalie Mitton, Valeria Loscri.

Interconnection and self-organized systems are normally populated with heterogeneous and different devices. The differences range from computational capabilities, storage size, etc. Instead of considering the heterogeneity as a limitation, it is possible to "turn it" as a primitive control of the system, in order to realize more robust and more resilient communication systems.

Based on those considerations, we have studied and analyzed the specific features of devices belonging to the category of micro-nano nodes that are however, required to interact with up-sized devices.

In order to improve the understanding of the behavior of micro/nano-sized devices, we have considered fundamental the analysis in specific applications and environment, where this kind of devices can be largely exploited, such as on/in-body networks applications.

Indeed, we retain that bio-medical applications can be advantaged by an effective and efficient communication and cooperation of devices deployed both on top of the body and inside it. Even if the research community recognizes a great importance to the study of interaction between the Human Immune System (HIS) and nano devices, this branch of research is in its infancy due to the major issue to model the HIS. A theoretical derivation of HIS and its interaction with a nanoparticulate system have been proposed in [15]. Some experimental results have been derived in [16], where specific parameters, e.g. temperature variations, Ph, etc. have been considered to establish the biocpmpatibility of TiO2 particles with human tissues.

A step ahead in this direction has consisted in the consideration of alternative particles as potential information carriers always in the context of biological environments. In [40] we have studied *phonons* as information carriers, we have derived a channel modeling and evaluated the theoretical capacity. The main reasons for taking into consideration this type of nanoparticles are twofold. Firstly, phonons represent something that is naturally generated in a biological context with the application of a tolerable electromagnetic field and secondly they represent a straightforward way to implement nanomachines, since their native size.

7.11. Modelling and experimentations of interferences and other PHY effects

Participants: Nathalie Mitton, Valeria Loscri.

In the era of Internet of Things (IoT), the development of Wireless Sensor Networks (WSN) arises different challenges. Two of the main issues are electromagnetic interference and the lifetime of WSN nodes. In [48], we show and evaluate experimentally the relation between interference and energy consumption, which impacts the network lifetime. We present a platform based on commercially available low-cost hardware in order to evaluate the impact of electromagnetic interference in 2.4 GHz ISM band on energy consumption of WSN. The energy measurements are obtained separately from each electronic component in the node. Interference and energy measurements are conducted in an anechoic chamber and in an office-type lab environment. X-MAC protocol is chosen to manage the Radio Duty Cycle of the nodes and its energy performance is evaluated. The energy consumption transmitter nodes is analyzed particularly in this work. Moreover, this energy consumption has been quantified and differentiated according to the number of (re-)transmissions carried out by the transmitter as well as the number of ACK packets sent by the receiver for a single packet. Finally, we use a model of real battery to calculate the lifetime of the node for operation within different interference level zones. This study lays the basis for further design rules of communication protocols and development of WSNs.

In [49], we propose a WSN architecture for wild animal monitoring. The key requirements of the system are long range transmissions and low power consumption. Indeed, the animals could be spread over vast areas. Kruger National Park in South Africa (19485 km2) is the potential zone of implementation of the network. On the other hand, size and weight limitations of wearable devices must be respected, which limits the size and capacity of battery. Moreover, battery replacement is a difficult and expensive process. So, low energy consumption is essential to extend the network lifetime. Some animal tracking projects [3] use GSM to transmit collected data to insure the coverage over a large area. However, high energy consumption of GSM and lack of coverage of the deployment area do not meet the essential requirements of the application. LoRa technology provides both long range transmissions and low power operation. This technology could be an appropriate solution for PREDNET project. The contribution of this work is multiple: 1) we defined communication parameters of LoRa radio for PREDNET WSN; 2) we performed radio propagation simulation for chosen parameters to estimate the coverage area for both urban and wilderness (rural) scenarios; 3) we confirmed the propagation simulations with range tests; 4) we measured experimentally the Packet Error Rate (PER) of transmissions.

Terahertz frequency band is an emerging research area related to nano-scale communications. In this frequency range, specific features can provide the possibility to overcome the issues related to the spectrum scarcity and capacity limitation.

Apart high molecular absorption, and very high reflection loss that represent main phenomena in THz band, we can derive the characteristics of the channel affected by chirality effects occurring in the propagation medium, specifically , in the case where a Giant Optical Activity is present. This effect is typical of the so-called chiral-metamaterials in (4-10) THz band, and is of stimulating interest particularly for millimeter wireless communications.

In [51], [25], we analyze the behavior of specific parameters of a chiral-metamaterial, like the relative electrical permitivity, magnetic permeability and chirality coefficients, and from that we derive the channel behavior both for Line-of-Sight and No Line-of-Sight propagations. We notice the presence of spectral windows, due to peaks of resonance of chiral parameter.

Finally, performances of the chirality-affected channel have been assessed in terms of (i) channel capacity, (ii) propagation delay, and (iii) coherence band-width, for different distances.

Thanks to the exploitation of frequencies in the interval ranging from 0.06 to 10 THz, it is envisioned the possibility to overcome the issues related to the spectrum scarcity and capacity limitation. On the other hand, the design of new channel models, able to capture the inherent features of the phenomenons related with this specific field is of paramount importance. Very high molecular absorption, and very high reflection loss are peculiarities phenomenons that need to be included in these models. In [26], we present a full-wave propagation model of the electromagnetic field that propagates in the THz band both for Line-of-Sight and Non-Line-of-Sight propagation models. In the full-wave model, we also introduce the chirality effects occurring in the propagation medium, i.e., a chiral metamaterial.

GANG Project-Team

6. New Results

6.1. Graph and Combinatorial Algorithms

6.1.1. New Results in Multi-sweep Graph Search

A theoretical model to describe a series of successive graph searches is proposed in [7]. We apply this model to deal with cocomparability graphs (i.e., complement of comparability graphs) in [6] and in [48] or [44]. In this series of papers we provide a general algorithmic framework for many optimization problems on cocomparability graphs, such as Minimum Path Cover, Maximum Independent Set, Maximal interval subgraph, etc.

We also provide a new very simple algorithm for the recognition of cocomparability graphs. This algorithm is also based on a series of successive graph searches in [13].

We mainly use the two well-known Lexicographic graph searches: LBFS and LDFS, but not only. In [48], we also introduced a new graph search LocalMNS which seems to behave nicely on cocomparability graphs.

6.1.2. Studies of Read Networks and Laminar Graphs

In the context of biological networks, in [50] we introduce k-laminar graphs — a new class of graphs which extends the idea of Asteroidal-triple-free graphs. A graph is k-laminar if it admits a diametral path that is k-dominating. This bio-inspired class of graphs was motivated by a biological application dealing with sequence similarity networks of reads. We briefly develop the context of the biological application in which this graph class appeared and then we consider the relationships of this new graph class among known graph classes and then we study its recognition problem. For the recognition of k-laminar graphs, we develop polynomial algorithms when k is fixed. For $k = 1$, our algorithm improves a Deogun and Krastch's algorithm (1999). We finish by an NP-completeness result when k is unbounded.

6.1.3. Further Studies into Shortest Paths, Eccentricity, and Laminarity

From our recent research on diameter computations on graphs we also investigated some reductions between polynomial problems on graphs [3].

We also extend the well-known multisweep BFS to give a better polynomial-time approximation for the Maximum Eccentricity Shortest Path Problem, in relation with the k-Laminarity Problem [20].

6.1.4. Clique Colourings of Perfect Graphs

A *clique-coloring* of a graph G is an assignment of colors to the vertices of G in such a way that no inclusion-wise maximal clique of size at least two of G is monochromatic (as usual, a set of vertices is *monochromatic* if all vertices in the set received the same color). The *clique-chromatic number* of G, denoted by $\chi_C(G)$, is the smallest integer k such that G is admits a clique-coloring using at most k colors. Note that every proper coloring of G is also a clique-coloring of G, and so $\chi_C(G) \leq \chi(G)$. Furthermore, if G is triangle-free, then $\chi_C(G) = \chi(G)$ (since there are triangle-free graphs of arbitrarily large chromatic number, this implies that there are triangle-free graphs of arbitrarily large clique-chromatic number). However, if G contains triangles, $\chi_C(G)$ may be much smaller than $\chi(G)$. For instance, if G contains a dominating vertex, then $\chi_C(G) \leq 2$ (we assign the color 1 to the dominating vertex and the color 2 to all other vertices of G), while $\chi(G)$ may be arbitrarily large. Note that this implies that the clique-chromatic number is not monotone with respect to induced subgraphs, that is, there exist graphs H and G such that H is an induced subgraph of G, but $\chi_C(H) > \chi_C(G)$. (In particular, the restriction of a clique-coloring of G to an induced subgraph H of G need not be a clique-coloring of H.)

A graph G is *perfect* if all its induced subgraphs H satisfy $\chi(H) = \omega(H)$, where $\omega(H)$ denotes the size of a maximum clique. It was asked by Duffus, Sands, Sauer, and Woodrow in a paper from 1991 whether perfect graphs have a bounded clique-chromatic number and indeed it has been proven since that for many sublasses of the class of perfect graphs, this holds. Even more, until now it was not known whether there were any perfect graphs of clique-chromatic number greater than three. The main result of [4] is to prove that there exist perfect graphs of arbitrarily large clique-chromatic number, which gives a negative answer for the question of Duffus et al. mentioned above.

6.1.5. Algorithmic Aspects of Switch Cographs

Cographs are the graphs totally decomposable using series and parallel operations, in [5] we introduced an interesting generalization, namely the class of switch cographs. These are the class of graphs that are totally decomposable w.r.t involution modular decomposition — a generalization of the modular decomposition of 2-structure, which has a unique linear-sized decomposition tree. We use our new decomposition tool to design three practical algorithms for the maximum cut, vertex cover and vertex separator problems. The complexity of these problems was previously unknown for this class of graphs.

6.1.6. Shrinking Maxima, Decreasing Costs: New Online Packing and Covering Problems

In [16], we consider two new variants of online integer programs that are duals. In the packing problem we are given a set of items and a collection of knapsack constraints over these items that are revealed over time in an online fashion. Upon arrival of a constraint we may need to remove several items (irrevocably) so as to maintain feasibility of the solution. Hence, the set of packed items becomes smaller over time. The goal is to maximize the number, or value, of packed items. The problem originates from a buffer-overflow model in communication networks, where items represent information units broken into multiple packets. The other problem considered is online covering: there is a universe to be covered. Sets arrive online, and we must decide for each set whether we add it to the cover or give it up. The cost of a solution is the total cost of sets taken, plus a penalty for each uncovered element. The number of sets in the solution grows over time, but its cost goes down. This problem is motivated by team formation, where the universe consists of skills, and sets represent candidates we may hire. The packing problem was introduced in Emek et al. (SIAM J Comput 41(4):728-746, 2012) for the special case where the matrix is binary; in this paper we extend the solution to general matrices with non-negative integer entries. The covering problem is introduced in this paper; we present matching upper and lower bounds on its competitive ratio.

6.1.7. The Complexity of the Shortest-path Broadcast Problem

In [8], we study the shortest-path broadcast problem in graphs and digraphs, where a message has to be transmitted from a source node s to all the nodes along shortest paths, in the classical telephone model. For both graphs and digraphs, we show that the problem is equivalent to the broadcast problem in layered directed graphs. We then prove that this latter problem is NP-hard, and therefore that the shortest-path broadcast problem is NP-hard in graphs as well as in digraphs. Nevertheless, we prove that a simple polynomial-time algorithm, called MDST-broadcast, based on min-degree spanning trees, approximates the optimal broadcast time within a multiplicative factor $3/2$ in 3-layer digraphs, and $O(\log n/\log\log n)$ in arbitrary multi-layer digraphs. As a consequence, one can approximate the optimal shortest-path broadcast time in polynomial time within a multiplicative factor $3/2$ whenever the source has eccentricity at most 2, and within a multiplicative factor $O(\log n/\log\log n)$ in the general case, for both graphs and digraphs. The analysis of MDST-broadcast is tight, as we prove that this algorithm cannot approximate the optimal broadcast time within a factor smaller than $\Omega(\log n/\log\log n)$.

6.1.8. Setting Ports in an Anonymous Network: How to Reduce the Level of Symmetry

A fundamental question in the setting of anonymous graphs concerns the ability of nodes to spontaneously break symmetries, based on their local perception of the network. In contrast to previous work, which focuses on symmetry breaking under arbitrary port labelings, in [37] we study the following design question: Given an anonymous graph G without port labels, how to assign labels to the ports of G, in interval form at each vertex, so that symmetry breaking can be achieved using a message-passing protocol requiring as few rounds of synchronous communication as possible?

More formally, for an integer $l > 0$, the *truncated view* $\mathcal{V}_l(v)$ of a node v of a port-labeled graph is defined as a tree encoding labels encountered along all walks in the network which originate from node v and have length at most l, and we ask about an assignment of labels to the ports of G so that the views $\mathcal{V}_l(v)$ are distinct for all nodes $v \in V$, with the goal being to minimize l.

We present such efficient port labelings for any graph G, and we exhibit examples of graphs showing that the derived bounds are asymptotically optimal in general. More precisely, our results imply the following statements.

1. For any graph G with n nodes and diameter D, a uniformly random port labeling achieves $l = O(\min(D, \log n))$, w.h.p.
2. For any graph G with n nodes and diameter D, it is possible to construct in polynomial time a labeling that satisfies $l = O(\min(D, \log n))$.
3. For any integers $n \geq 2$ and $D \leq \log_2 n - \log_2 \log_2 n$, there exists a graph G with n nodes and diameter D which satisfies $l \geq \frac{1}{2}D - \frac{5}{2}$.

6.1.9. Robustness of the Rotor-Router Mechanism

The *rotor-router model*, also called the *Propp machine*, was first considered as a deterministic alternative to the random walk. The edges adjacent to each node v (or equivalently, the exit ports at v) are arranged in a fixed cyclic order, which does not change during the exploration. Each node v maintains a *port pointer* π_v which indicates the exit port to be adopted by an agent on the conclusion of the next visit to this node (the "next exit port"). The rotor-router mechanism guarantees that after each consecutive visit at the same node, the pointer at this node is moved to the next port in the cyclic order. It is known that, in an undirected graph G with m edges, the route adopted by an agent controlled by the rotor-router mechanism forms eventually an Euler tour based on arcs obtained via replacing each edge in G by two arcs with opposite direction. The process of ushering the agent to an Euler tour is referred to as the *lock-in problem*. In [Yanovski et al., Algorithmica 37(3), 165–186 (2003)], it was proved that, independently of the initial configuration of the rotor-router mechanism in G, the agent locks-in in time bounded by $2mD$, where D is the diameter of G.

In [2], we examine the dependence of the lock-in time on the initial configuration of the rotor-router mechanism. Our analysis is performed in the form of a game between a player p intending to lock-in the agent in an Euler tour as quickly as possible and its adversary a with the counter objective. We consider all cases of who decides the initial cyclic orders and the initial values π_v. We show, for example, that if a provides its own port numbering after the initial setup of pointers by p, the complexity of the lock-in problem is $O(m \cdot \min\{\log m, D\})$.

We also investigate the robustness of the rotor-router graph exploration in presence of faults in the pointers π_v or dynamic changes in the graph. We show, for example, that after the exploration establishes an Eulerian cycle, if k edges are added to the graph, then a new Eulerian cycle is established within $O(km)$ steps.

6.1.10. The Multi-Agent Rotor-Router on the Ring: A Deterministic Alternative to Parallel Random Walks

Continuing the line of research on the rotor-router model, in [18] we consider the setting in which multiple, indistinguishable agents are deployed in parallel in the nodes of the graph, and move around the graph in synchronous rounds, interacting with a single rotor-router system. We propose new techniques which allow us to perform a theoretical analysis of the multi-agent rotor-router model, and to compare it to the scenario of parallel independent random walks in a graph. Our main results concern the n-node ring, and suggest a strong similarity between the performance characteristics of this deterministic model and random walks.

We show that on the ring the rotor-router with k agents admits a cover time of between $\Theta(n^2/k^2)$ in the best case and $\Theta(n^2/\log k)$ in the worst case, depending on the initial locations of the agents, and that both these bounds are tight. The corresponding expected value of the cover time for k random walks, depending on the initial locations of the walkers, is proven to belong to a similar range, namely between $\Theta(n^2/(k^2/\log^2 k))$ and $\Theta(n^2/\log k)$.

Finally, we study the limit behavior of the rotor-router system. We show that, once the rotor-router system has stabilized, all the nodes of the ring are always visited by some agent every $\Theta(n/k)$ steps, regardless of how the system was initialized. This asymptotic bound corresponds to the expected time between successive visits to a node in the case of k random walks. All our results hold up to a polynomially large number of agents ($1 \leq k < n^{1/11}$).

6.1.11. Bounds on the Cover Time of Parallel Rotor Walks

In [12], we study the parallel rotor-router model in the case of general graphs. We consider the cover time of such a system, i.e., the number of steps after which each node has been visited by at least one walk, regardless of the initialization of the walks. We show that for any graph with m edges and diameter D, this cover time is at most $\Theta(mD/\log k)$ and at least $\Theta(mD/k)$, which corresponds to a speedup of between $\Theta(\log k)$ and $\Theta(k)$ with respect to the cover time of a single walk.

6.2. Distributed Computing

6.2.1. Local Conflict Coloring

Locally finding a solution to symmetry-breaking tasks such as vertex-coloring, edge-coloring, maximal matching, maximal independent set, etc., is a long-standing challenge in distributed network computing. More recently, it has also become a challenge in the framework of centralized local computation. In [30], we introduce conflict coloring as a general symmetry-breaking task that includes all the aforementioned tasks as specific instantiations — conflict coloring includes all locally checkable labeling tasks from [Naor&Stockmeyer, STOC 1993]. Conflict coloring is characterized by two parameters l and d, where the former measures the amount of freedom given to the nodes for selecting their colors, and the latter measures the number of constraints which colors of adjacent nodes are subject to. We show that, in the standard LOCAL model for distributed network computing, if $l/d > \Delta$, then conflict coloring can be solved in $\widetilde{O}(\sqrt{\Delta}) + \log^* n$ rounds in n-node graphs with maximum degree Δ, where $\widetilde{O}$ ignores the polylog factors in Δ. The dependency in n is optimal, as a consequence of the $\Omega(\log^* n)$ lower bound by [Linial, SIAM J. Comp. 1992] for $(\Delta+1)$-coloring. An important special case of our result is a significant improvement over the best known algorithm for distributed $(\Delta+1)$-coloring due to [Barenboim, PODC 2015], which required $\widetilde{O}(\Delta^{3/4}) + \log^* n$ rounds. Improvements for other variants of coloring, including $(\Delta+1)$-list-coloring, $(2\Delta-1)$-edge-coloring, T-coloring, etc., also follow from our general result on conflict coloring. Likewise, in the framework of centralized local computation algorithms (LCAs), our general result yields an LCA which requires a smaller number of probes than the previously best known algorithm for vertex-coloring, and works for a wide range of coloring problems.

6.2.2. A Hierarchy of Local Decision

In [29], we extend the notion of *distributed decision* in the framework of distributed network computing, inspired by recent results on so-called *distributed graph automata*. We show that, by using distributed decision mechanisms based on the interaction between a *prover* and a *disprover*, the size of the certificates distributed to the nodes for certifying a given network property can be drastically reduced. For instance, we prove that minimum spanning tree can be certified with $O(\log n)$-bit certificates in n-node graphs, with just one interaction between the prover and the disprover, while it is known that certifying MST requires $\Omega(\log^2 n)$-bit certificates if only the prover can act. The improvement can even be exponential for some simple graph properties. For instance, it is known that certifying the existence of a nontrivial automorphism requires $\Omega(n^2)$ bits if only the prover can act. We show that there is a protocol with two interactions between the prover and the disprover enabling to certify nontrivial automorphism with $O(\log n)$-bit certificates. These results are achieved by defining and analysing a *local hierarchy* of decision which generalizes the classical notions of *proof-labelling schemes* and *locally checkable proofs*.

6.2.3. Distributed Testing of Excluded Subgraphs

In [35], we study property testing in the context of distributed computing, under the classical CONGEST model. It is known that testing whether a graph is triangle-free can be done in a constant number of rounds,

where the constant depends on how far the input graph is from being triangle-free. We show that, for every connected 4-node graph H, testing whether a graph is H-free can be done in a constant number of rounds too. The constant also depends on how far the input graph is from being H-free, and the dependence is identical to the one in the case of testing triangle-freeness. Hence, in particular, testing whether a graph is K_4-free, and testing whether a graph is C_4-free can be done in a constant number of rounds (where K_k denotes the k-node clique, and C_k denotes the k-node cycle). On the other hand, we show that testing K_k-freeness and C_k-freeness for $k \geq 5$ appear to be much harder. Specifically, we investigate two natural types of generic algorithms for testing H-freeness, called DFS tester and BFS tester. The latter captures the previously known algorithm to test the presence of triangles, while the former captures our generic algorithm to test the presence of a 4-node graph pattern H. We prove that both DFS and BFS testers fail to test K_k-freeness and C_k-freeness in a constant number of rounds for $k \geq 5$.

6.2.4. Asynchronous Coordination Under Preferences and Constraints

Adaptive renaming can be viewed as a coordination task involving a set of asynchronous agents, each aiming at grabbing a single resource out of a set of resources totally ordered by their desirability. Similarly, musical chairs is also defined as a coordination task involving a set of asynchronous agents, each aiming at picking one of a set of available resources, where every agent comes with an a priori preference for some resource. In [22], we foresee instances in which some combinations of resources are allowed, while others are disallowed. We model these constraints, i.e., the restrictions on the ability to use some combinations of resources, as an undirected graph whose nodes represent the resources, and an edge between two resources indicates that these two resources cannot be used simultaneously. In other words, the sets of resources that are allowed are those which form independent sets in the graph. E.g., renaming and musical chairs are specific cases where the graph is stable (i.e., it the empty graph containing no edges). As for musical chairs, we assume that each agent comes with an a priori preference for some resource. If an agent's preference is not in conflict with the preferences of the other agents, then this preference can be grabbed by the agent. Otherwise, the agents must coordinate to resolve their conflicts, and potentially choose non preferred resources. We investigate the following problem: given a graph, what is the maximum number of agents that can be accommodated subject to non-altruistic behaviors of early arriving agents? We entirely solve this problem under the restriction that agents which cannot grab their preferred resources must then choose a resource among the nodes of a predefined independent set. However, the general case, where agents which cannot grab their preferred resource are then free to choose any resource, is shown to be far more complex. In particular, just for cyclic constraints, the problem is surprisingly difficult. Indeed, we show that, intriguingly, the natural algorithm inspired from optimal solutions to adaptive renaming or musical chairs is sub-optimal for cycles, but proven to be at most 1 to the optimal. The main message of this paper is that finding optimal solutions to the coordination with constraints and preferences task requires to design " dynamic " algorithms, that is, algorithms of a completely different nature than the "static" algorithms used for, e.g., renaming.

6.2.5. Making Local Algorithms Wait-Free: The Case of Ring Coloring

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. This work proposes 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 [21] 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 vertex coloring algorithm in which the processes are not required to start simultaneously.

6.2.6. t-Resilient Immediate Snapshot Is Impossible

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 (t-crash system model), the paper [25] is on the construction of t-resilient immediate snapshot objects. In the t-crash system model, a process can obtain values from at least $(n-t)$ processes, and, consequently, t-immediate snapshot is assumed to have the properties of the basic $(n-1)$-resilient immediate snapshot plus the additional property stating that each process obtains values from at least $(n-t)$ processes. The main result of the work is the following. While there is a (deterministic) $(n-1)$-resilient algorithm implementing the basic $(n-1)$-immediate snapshot in an $(n-1)$-crash read/write system, there is no t-resilient algorithm in a t-crash read/write model when $t \in [1...(n-2)]$. This means that, when $t < n-1$, the notion of t-resilience is inoperative when one has to implement t-immediate snapshot for these values of t: the model assumption "at most $t < n-1$ processes may crash" does not provide us with additional computational power allowing for the design of a genuine t-resilient algorithm (genuine meaning that such an algorithm would work in the t-crash model, but not in the $(t+1)$-crash model). To show these results, we rely on well-known distributed computing agreement problems such as consensus and k-set agreement.

6.2.7. Perfect Failure Detection with Very Few Bits

A *failure detector* is a distributed oracle that provides the processes with information about failures. The *perfect* failure detector provides accurate and eventually complete information about process failures. In [34], we show that, in asynchronous failure-prone message-passing systems, perfect failure detection can be achieved by an oracle that outputs at most $\lceil \log \alpha(n) \rceil + 1$ bits per process in n-process systems, where α denotes the inverse-Ackermann function. This result is essentially optimal, as we also show that, in the same environment, no failure detectors outputting a constant number of bit per process can achieve perfect failure detection.

6.2.8. Decentralized Asynchronous Crash-Resilient Runtime Verification

Runtime Verification (RV) is a lightweight method for monitoring the formal specification of a system during its execution. It has recently been shown that a given state predicate can be monitored consistently by a set of crash-prone asynchronous *distributed* monitors, only if sufficiently many different verdicts can be emitted by each monitor. In [27], we revisit this impossibility result in the context of LTL semantics for RV. We show that employing the four-valued logic RVLTL will result in inconsistent distributed monitoring for some formulas. Our first main contribution is a family of logics, called LTL(k), that refines RVLTL incorporating $2k+4$ truth values, for each $k \geq 0$. The truth values of LTL(k) can be effectively used by each monitor to reach a consistent global set of verdicts for each given formula, provided k is sufficiently large. Our second main contribution is an algorithm for monitor construction enabling fault-tolerant distributed monitoring based on the aggregation of the individual verdicts by each monitor.

6.2.9. Asynchronous Consensus with Bounded Memory

The paper [11] presents a bounded memory size Obstruction-Free consensus algorithm for the asynchronous shared memory model. More precisely for a set of n processes, this algorithm uses $n+2$ multi-writer multi-reader registers, each of these registers being of size $O(\log n)$ bits. From this, we get a bounded memory size space complexity consensus algorithm with single-writer multi-reader registers and a bounded memory size space complexity consensus algorithm in the asynchronous message passing model with a majority of correct processes. As it is easy to ensure the Obstruction-Free assumption with randomization (or with leader election failure detector Ω) we obtain a bounded memory size randomized consensus algorithm and a bounded memory size consensus algorithm with failure detector.

6.2.10. Implementing Snapshot Objects on Top of Crash-Prone Asynchronous Message-Passing Systems

Distributed snapshots, as introduced by Chandy and Lamport in the context of asynchronous failure-free message-passing distributed systems, are consistent global states in which the observed distributed application

might have passed through. It appears that two such distributed snapshots cannot necessarily be compared (in the sense of determining which one of them is the "first"). Differently, snapshots introduced in asynchronous crash-prone read/write distributed systems are totally ordered, which greatly simplify their use by upper layer applications. In order to benefit from shared memory snapshot objects, it is possible to simulate a read/write shared memory on top of an asynchronous crash-prone message-passing system, and build then snapshot objects on top of it. This algorithm stacking is costly in both time and messages. To circumvent this drawback, the paper [24] presents algorithms building snapshot objects directly on top of asynchronous crash-prone message-passing system. "Directly" means here "without building an intermediate layer such as a read/write shared memory". To the authors knowledge, the proposed algorithms are the first providing such constructions. Interestingly enough, these algorithms are efficient and relatively simple.

6.2.11. Set-Consensus Collections are Decidable

A natural way to measure the power of a distributed-computing model is to characterize the set of tasks that can be solved in it. In general, however, the question of whether a given task can be solved in a given model is undecidable, even if we only consider the wait-free shared-memory. In [23], we address this question for restricted classes of models and tasks. We show that the question of whether a collection C of (ℓ, j)-set consensus objects, for various ℓ (the number of processes that can invoke the object) and j (the number of distinct outputs the object returns), can be used by n processes to solve wait-free k-set consensus is decidable. Moreover, we provide a simple $O(n^2)$ decision algorithm, based on a dynamic programming solution to the Knapsack optimization problem. We then present an adaptive wait-free set-consensus algorithm that, for each set of participating processes, achieves the best level of agreement that is possible to achieve using C. Overall, this gives us a complete characterization of a read-write model defined by a collection of set-consensus objects through its set-consensus power.

6.2.12. Minimizing the Number of Opinions for Fault-Tolerant Distributed Decision Using Well-Quasi Orderings

The notion of deciding a *distributed language*L is of growing interest in various distributed computing settings. Each process p_i is given an input value x_i, and the processes should collectively decide whether their set of input values $x = (x_i)_i$ is a valid state of the system w.r.t. to some specification, i.e., if $x \in L$. In *non-deterministic* distributed decision each process p_i gets a local certificate c_i in addition to its input x_i. If the input $x \in L$then there exists a certificate $c = (c_i)_i$ such that the processes collectively accept x, and if $x \notin L$, then for every c, the processes should collectively reject x. The collective decision is expressed by the set of *opinions* emitted by the processes, and one aims at minimizing the number of possible opinions emitted by each process. In [33], we study non-deterministic distributed decision in asynchronous systems where processes may crash. In this setting, it is known that the number of opinions needed to deterministically decide a language can grow with n, the number of processes in the system.We prove that every distributed language L can be non-deterministically decided using only three opinions, with certificates of size $\lceil \log \alpha(n) \rceil + 1$ bits, where α grows at least as slowly as the inverse of the Ackerman function. The result is optimal, as we show that there are distributed languages that cannot be decided using just two opinions, even with arbitrarily large certificates. To prove our upper bound, we introduce the notion of *distributed encoding of the integers*, that provides an explicit construction of a long *bad sequence* in the *well-quasi-ordering* $(\{0,1\}^*, \leq_*)$ controlled by the successor function. Thus, we provide a new class of applications for well-quasi-orderings that lies outside logic and complexity theory. For the lower bound we use combinatorial topology techniques.

6.2.13. Collision-Free Network Exploration

In [9], we consider a network exploration setting in which mobile agents start at different nodes of an n-node network. The agents synchronously move along the network edges in a *collision-free* way, i.e., in no round two agents may occupy the same node. An agent has no knowledge of the number and initial positions of other agents. We are looking for the shortest time required to reach a configuration in which each agent has visited all nodes and returned to its starting location. In the scenario when each mobile agent knows the map of the network, we provide tight (up to a constant factor) lower and upper bounds on the collision-free exploration time in arbitrary graphs, and the exact bound for the trees. In the second scenario, where the

network is unknown to the agents, we propose collision-free exploration strategies running in $O(n^2)$ rounds in tree networks and in $O(n^5 \log n)$ rounds in networks with an arbitrary topology.

6.2.14. When Patrolmen Become Corrupted: Monitoring a Graph Using Faulty Mobile Robots

In [10], we consider a setting in which a team of k mobile robots is deployed on a weighted graph whose edge weights represent distances. The robots perpetually move along the domain, represented by all points belonging to the graph edges, not exceeding their maximal speed. The robots need to patrol the graph by regularly visiting all points of the domain. We consider a team of robots (patrolmen), at most f of which may be unreliable, failing to comply with their patrolling duties. What algorithm should be followed so as to minimize the maximum time between successive visits of every edge point by a reliable patrolmen? The corresponding measure of efficiency of patrolling called *idleness* has been widely accepted in the robotics literature. We extend it to the case of untrusted patrolmen; we denote by $I_k^f(G)$ the maximum time that a point of the domain may remain unvisited by reliable patrolmen. The objective is to find patrolling strategies minimizing $I_k^f(G)$.

We investigate this problem for various classes of graphs. We design optimal algorithms for line segments, which turn out to be surprisingly different from strategies for related patrolling problems proposed in the literature. We then use these results to provide algorithms for general graphs. For Eulerian graphs G, we give an optimal patrolling strategy with idleness $I_k^f(G) = (f+1)E(G)/k$, where $E(G)$ is the sum of the lengths of the edges of G. For arbitrary graphs and given ratio r of faulty robots, $r := f/k < 1/2$, we design a strategy which is a $(1+\epsilon)$ approximation of the optimal one, for sufficiently large k. Further, we show the hardness of the problem of computing the idle time for three robots, at most one of which is faulty, by reduction from 3-edge-coloring of cubic graphs — a known NP-hard problem. A byproduct of our proof is the investigation of classes of graphs minimizing idle time (with respect to the total length of edges); an example of such a class is known in the literature under the name of Kotzig graphs.

6.2.15. Noisy Rumor Spreading and Plurality Consensus

Error-correcting codes are efficient methods for handling noisy communication channels in the context of technological networks. However, such elaborate methods differ a lot from the unsophisticated way biological entities are supposed to communicate. Yet, it has been recently shown by Feinerman, Haeupler, and Korman [PODC 2014] that complex coordination tasks such as rumor spreading and majority consensus can plausibly be achieved in biological systems subject to noisy communication channels, where every message transferred through a channel remains intact with small probability $1+\varepsilon$, without using coding techniques. This result is a considerable step towards a better understanding of the way biological entities may cooperate. It has nevertheless been established only in the case of 2-valued opinions: rumor spreading aims at broadcasting a single-bit opinion to all nodes, and majority consensus aims at leading all nodes to adopt the single-bit opinion that was initially present in the system with (relative) majority. In [32], we extend this previous work to k-valued opinions, for any constant $k \geq 2$. Our extension requires to address a series of important issues, some conceptual, others technical. We had to entirely revisit the notion of noise, for handling channels carrying k-valued messages. In fact, we precisely characterize the type of noise patterns for which plurality consensus is solvable. Also, a key result employed in the bivalued case by Feinerman et al. is an estimate of the probability of observing the most frequent opinion from observing the mode of a small sample. We generalize this result to the multivalued case by providing a new analytical proof for the bivalued case that is amenable to be extended, by induction, and that is of independent interest.

6.3. Models and Algorithms for Networks

6.3.1. Beyond Highway Dimension: Small Distance Labels Using Tree Skeletons

The goal of a hub-based distance labeling scheme for a network $G = (V, E)$ is to assign a small subset $S(u) \subseteq V$ to each node $u \in V$, in such a way that for any pair of nodes u, v, the intersection of hub sets $S(u) \cap S(v)$ contains a node on the shortest uv-path. The existence of small hub sets, and consequently efficient shortest path processing algorithms, for road networks is an empirical observation. A theoretical

explanation for this phenomenon was proposed by Abraham et al. (SODA 2010) through a network parameter they called highway dimension, which captures the size of a hitting set for a collection of shortest paths of length at least r intersecting a given ball of radius $2r$. In [38], we revisit this explanation, introducing a more tractable (and directly comparable) parameter based solely on the structure of shortest-path spanning trees, which we call skeleton dimension. We show that skeleton dimension admits an intuitive definition for both directed and undirected graphs, provides a way of computing labels more efficiently than by using highway dimension, and leads to comparable or stronger theoretical bounds on hub set size.

6.3.2. Sublinear-Space Distance Labeling using Hubs

Continuing work in the previously discussed framework of hub-based distance labeling schemes, in [36], [39], we present a hub labeling which allows us to decode exact distances in sparse graphs using labels of size sublinear in the number of nodes. For graphs with at most n nodes and average degree Δ, the tradeoff between label bit size L and query decoding time T for our approach is given by $L = O(n \log\log_\Delta T / \log_\Delta T)$, for any $T \leq n$. Our simple approach is thus the first sublinear-space distance labeling for sparse graphs that simultaneously admits small decoding time (for constant Δ, we can achieve any $T = \omega(1)$ while maintaining $L = o(n)$), and it also provides an improvement in terms of label size with respect to previous slower approaches.

By using similar techniques, we then present a 2-additive labeling scheme for general graphs, i.e., one in which the decoder provides a 2-additive-approximation of the distance between any pair of nodes. We achieve almost the same label size-time tradeoff $L = O(n \log^2 \log T / \log T)$, for any $T \leq n$. To our knowledge, this is the first additive scheme with constant absolute error to use labels of sublinear size. The corresponding decoding time is then small (any $T = \omega(1)$ is sufficient).

We believe all of our techniques are of independent value and provide a desirable simplification of previous approaches.

6.3.3. Labeling Schemes for Ancestry Relation

In [17], we solve the ancestry-labeling scheme problem which aims at assigning the shortest possible labels (bit strings) to nodes of rooted trees, so that ancestry queries between any two nodes can be answered by inspecting their assigned labels only. This problem was introduced more than twenty years ago by Kannan et al. [STOC '88], and is among the most well-studied problems in the field of informative labeling schemes. We construct an ancestry-labeling scheme for n-node trees with label size $\log_2 n + O(\log\log n)$ bits, thus matching the $log_2 n + \Omega(\log\log n)$ bits lower bound given by Alstrup et al. [SODA '03]. Our scheme is based on a simplified ancestry scheme that operates extremely well on a restricted set of trees. In particular, for the set of n-node trees with depth at most d, the simplified ancestry scheme enjoys label size of $\log_2 n + 2\log_2 d + O(1)$ bits. Since the depth of most XML trees is at most some small constant, such an ancestry scheme may be of practical use. In addition, we also obtain an adjacency-labeling scheme that labels n-node trees of depth d with labels of size $\log_2 n + 3\log_2 d + O(1)$ bits. All our schemes assign the labels in linear time, and guarantee that any query can be answered in constant time. Finally, our ancestry scheme finds applications to the construction of small universal partially ordered sets (posets). Specifically, for any fixed integer k, it enables the construction of a universal poset of size $O(n^k)$ for the family of n-element posets with tree-dimension at most k. Up to lower order terms, this bound is tight thanks to a lower bound of $n^{k-o(1)}$ due to Alon and Scheinerman [Order '88].

6.3.4. Independent Lazy Better-Response Dynamics on Network Games

In [43], we study an independent best-response dynamics on network games in which the nodes (players) decide to revise their strategies independently with some probability. We are interested in the convergence time to the equilibrium as a function of this probability, the degree of the network, and the potential of the underlying games.

6.3.5. Forwarding Tables Verification through Representative Header Sets

Forwarding table verification consists in checking the distributed data-structure resulting from the forwarding tables of a network. A classical concern is the detection of loops. We study in [42] this problem in the context

of software-defined networking (SDN) where forwarding rules can be arbitrary bitmasks (generalizing prefix matching) and where tables are updated by a centralized controller. Basic verification problems such as loop detection are NP-hard and most previous work solves them with heuristics or SAT solvers. We follow a different approach based on computing a representation of the header classes, i.e. the sets of headers that match the same rules. This representation consists in a collection of representative header sets, at least one for each class, and can be computed centrally in time which is polynomial in the number of classes. Classical verification tasks can then be trivially solved by checking each representative header set. In general, the number of header classes can increase exponentially with header length, but it remains polynomial in the number of rules in the practical case where rules are constituted with predefined fields where exact, prefix matching or range matching is applied in each field (e.g., IP/MAC addresses, TCP/UDP ports). We propose general techniques that work in polynomial time as long as the number of classes of headers is polynomial and that do not make specific assumptions about the structure of the sets associated to rules. The efficiency of our method rely on the fact that the data-structure representing rules allows efficient computation of intersection, cardinal and inclusion. Finally, we propose an algorithm to maintain such representation in presence of updates (i.e., rule insert/update/removal). We also provide a local distributed algorithm for checking the absence of black-holes and a proof labeling scheme for locally checking the absence of loops.

6.3.6. A Locally-Blazed Ant Trail Achieves Efficient Collective Navigation Despite Limited Information

This work fits into the framework of computationally-inspired analysis of biological systems. Any organism faces sensory and cognitive limitations which may result in maladaptive decisions. Such limitations are prominent in the context of groups where the relevant information at the individual level may not coincide with collective requirements. In [14], we study the navigational decisions exhibited by *Paratrechina longicornis* ants as they cooperatively transport a large food item. These decisions hinge on the perception of individuals which often restricts them from providing the group with reliable directional information. We find that, to achieve efficient navigation despite partial and even misleading information, these ants employ a locally-blazed trail. This trail significantly deviates from the classical notion of an ant trail: First, instead of systematically marking the full path, ants mark short segments originating at the load. Second, the carrying team constantly loses the guiding trail. We experimentally and theoretically show that the locally-blazed trail optimally and robustly exploits useful knowledge while avoiding the pitfalls of misleading information.

6.3.7. Parallel Exhaustive Search without Coordination

In [31], we analyze parallel algorithms in the context of *exhaustive search* over totally ordered sets. Imagine an infinite list of "boxes", with a "treasure" hidden in one of them, where the boxes' order reflects the importance of finding the treasure in a given box. At each time step, a search protocol executed by a searcher has the ability to peek into one box, and see whether the treasure is present or not. Clearly, the best strategy of a single searcher would be to open the boxes one by one, in increasing order. Moreover, by equally dividing the workload between them, k searchers can trivially find the treasure k times faster than one searcher. However, this straightforward strategy is very sensitive to failures (*e.g.,* crashes of processors), and overcoming this issue seems to require a large amount of communication. We therefore address the question of designing parallel search algorithms maximizing their *speed-up* and maintaining high levels of *robustness*, while minimizing the amount of resources for coordination. Based on the observation that algorithms that avoid communication are inherently robust, we focus our attention on identifying the best running time performance of *non-coordinating* algorithms. Specifically, we devise non-coordinating algorithms that achieve a speed-up of $9/8$ for two searchers, a speed-up of $4/3$ for three searchers, and in general, a speed-up of $\frac{k}{4}(1+1/k)^2$ for any $k \geq 1$ searchers. Thus, asymptotically, the speed-up is only four times worse compared to the case of full-coordination. Moreover, these bounds are tight in a strong sense as no non-coordinating search algorithm can achieve better speed-ups. Furthermore, our algorithms are surprisingly simple and hence applicable. Overall, we highlight that, in faulty contexts in which coordination between the searchers is technically difficult to implement, intrusive with respect to privacy, and/or costly in term of resources, it might well be worth giving up on coordination, and simply run our non-coordinating exhaustive search algorithms.

6.3.8. Rumor Spreading in Random Evolving Graphs

Randomized gossip is one of the most popular way of disseminating information in large scale networks. This method is appreciated for its simplicity, robustness, and efficiency. In the Push protocol, every informed node selects, at every time step (a.k.a. round), one of its neighboring node uniformly at random and forwards the information to this node. This protocol is known to complete information spreading in $O(\log n)$ time steps with high probability (w.h.p.) in several families of n-node *static* networks. The Push protocol has also been empirically shown to perform well in practice, and, specifically, toe robust against dynamic topological changes. In [15], we aim at analyzing the Push protocol in *dynamic* networks. We consider the *edge-Markovian* evolving graph model which captures natural temporal dependencies between the structure of the network at time t, and the one at time $t+1$. Precisely, a non-edge appears with probability p, while an existing edge dies with probability q. In order to fit with real-world traces, we mostly concentrate our study on the case where $p = \Omega(\frac{1}{n})$ and q is constant. We prove that, in this realistic scenario, the Push protocol does perform well, completing information spreading in $O(\log n)$ time steps w.h.p. Note that this performance holds even when the network is, w.h.p., disconnected at every time step (e.g., when $p \ll \frac{\log n}{n}$). Our result provides the first formal argument demonstrating the robustness of the Push protocol against network changes. We also address another range of parameters p and q, namely $p + q = 1$ with arbitrary p and q. Although this latter range does not precisely fit with the measures performed on real-world traces, they can be of independent interest for other settings. The result in this case confirms the positive impact of dynamism.

6.3.9. Sparsifying Congested Cliques and Core-Periphery Networks

The *core-periphery* network architecture proposed by Avin et al. [ICALP 2014] was shown to support fast computation for many distributed algorithms, while being much sparser than the *congested clique*. For being efficient, the core-periphery architecture is however bounded to satisfy three axioms, among which is the capability of the core to emulate the clique, i.e., to implement the all-to-all communication pattern, in $O(1)$ rounds in the CONGEST model. In [26], we show that implementing all-to-all communication in k rounds can be done in n-node networks with roughly n^2/k edges, and this bound is tight. Hence, sparsifying the core beyond just saving a fraction of the edges requires to relax the constraint on the time to simulate the congested clique. We show that, for $p \gg \sqrt{\log n/n}$, a random graph in $\mathcal{G}_{n,p}$ can, w.h.p., perform the all-to-all communication pattern in $O(\min\{\frac{1}{p^2}, np\})$ rounds. Finally, we show that if the core can emulate the congested clique in t rounds, then there exists a distributed MST construction algorithm performing in $O(t \log n)$ rounds. Hence, for $t = O(1)$, our (deterministic) algorithm improves the best known (randomized) algorithm for constructing MST in core-periphery networks by a factor $\Theta(\log n)$.

6.3.10. Core-periphery Clustering and Collaboration Networks

In [28], we analyse the core-periphery clustering properties of collaboration networks, where the core of a network is formed by the nodes with highest degree. In particular, we first observe that, even for random graph models aiming at matching the degree-distribution and/or the clustering coefficient of real networks, these models produce synthetic graphs which have a spatial distribution of the triangles with respect to the core and to the periphery which does not match the spatial distribution of the triangles in the real networks. We therefore propose a new model, called CPCL, whose aim is to distribute the triangles in a way fitting with their real core-periphery distribution, and thus producing graphs matching the core-periphery clustering of real networks.

INFINE Project-Team

6. New Results

6.1. Online Social Networks (OSN)

Community detection; bandit algorithms; privacy preservation; reward mechanisms

6.1.1. Capacity of Information Processing Systems

Participants: Laurent Massoulié, 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 accepted and presented under the title "On the capacity of information processing systems" at the COLT 2016 conference.

6.2. Spontaneous Wireless Networks and Internet of Things

internet of things; wireless sensor networks; dissemination; resource management

6.2.1. Platform Design for the Internet of Things

Participants: Emmanuel Baccelli, Cedric Adjih, Oliver Hahm, Francisco Acosta, Hauke Petersen.

Within this activity, we have further developed the platforms we champion for the Internet of Things: the open source operating system RIOT on one hand, and open-access IoT-lab testbeds on the other hand. RIOT now aggregates open source contributions from 130+ people (and counting) from all over the world, coming both from academia and from industry, and received financial backing from top companies including Cisco and Google. We further developed RIOT for low-cost mobile robots and received the Best Demo Award at the ACM EWSN'16 conference for our work on this topic. As steering RIOT community members, we also participated in the prestigious Internet Architecture Board (IAB) workshop on IoT Software Updates, a hot and essential topic for the future of Internet of Things. The year culminated in this domain with the successful organization of the first RIOT Summit in Berlin, where 100+ participants from all over the world, from industry, academia as well as hackers/makers involved in RIOT gathered to discuss various aspects of the future of RIOT and open source IoT software. In addition, 2016, at the site of Saclay, one of the testbeds from FIT IoT-LAB was opened: the platform of Saclay includes more than 300 IoT nodes (175 A8-M3, 12 M3, 120 WSN430, some Arduinos and some SAMR21-xpro). In parallel, the platform from Freie Universitat Berlin also joined the OneLab/FIT IoT-LAB testbed federation.

6.2.2. Energy-Efficient Communication Protocols for the Internet of Things

Participants: Oliver Hahm, Emmanuel Baccelli, Cedric Adjih, Matthias Waehlisch, Thomas Schmidt.

Within this activity, we have designed distributed algorithms providing improved trade-off between content availability and energy efficiency (which plays a crucial role). The approach we developed leverages distributed caching for IoT content, based on an information-centric networking paradigm. We extended the NDN protocol with a variety of caching and replacement strategies, and we analyzed alternative approaches for extending NDN to accommodate such IoT use cases. Based on extensive experiments on real IoT hardware in a network gathering hundreds of nodes, we demonstrate these caching strategies can bring 90% reduction in energy consumption while maintaining IoT content availability above 90%. This work was published in IEEE Globecom'16 workshop on Named Data Networks for Challenged Communication Environments.

We also have designed new mechanisms to jointly exploit ICN communication patterns and dynamically optimize the use of TSCH (Time Slotted Channel Hopping), a wireless link layer technology increasingly popular in the IoT. Through a series of experiments on FIT IoT-LAB interconnecting typical IoT hardware, we find that our proposal is fully robust against wireless interference, and almost halves the energy consumed for transmission when compared to CSMA. Most importantly, our adaptive scheduling prevents the time-slotted MAC layer from sacrificing throughput and delay. Our work on ICN and on TSCH was published at NTMS'16, at ACM ICN'16, and in Proceedings of the IEEE.

6.2.3. Standards for Spontaneous Wireless Networks

Participant: Emmanuel Baccelli.

Within this activity, we have contributed to new network protocol standards for spontaneous wireless networking, applied to ad hoc networks and the Internet of Things. In particular, collaborating with Fraunhofer, we have published RFC 7779, standardizing Directional Airtime Metric (DAT), a new wireless metric standard targeting wireless mesh networks. Furthermore, collaborating with ARM and Sigma Designs, we published RFC 7733, which provides guidance in the configuration and use of protocols from the RPL protocol suite to implement the features required for control in building and home environments. In collaboration with various industrial partners, with have also published a number of other Internet drafts, including an analysis of the characteristics of multi-hop ad hoc wireless communication between interfaces in the context of IP networks, and an analysis of the challenges of information-centric networking in the Internet of Things.

6.2.4. Spatio-Temporal Predictability of Cellular Data Traffic

Participants: Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, Sahar Hoteit, Carlos Sarraute.

The ability to foresee the data traffic activity of subscribers opens new opportunities to reshape mobile network management and services. In this work, we leverage two large-scale real-world datasets collected by a major mobile carrier in Mexico to study how predictable are the cellular data traffic demands generated by individual users. We focus on the predictability of mobile traffic consumption patterns in isolation. Our results show that it is possible to anticipate the individual demand with a typical accuracy of 85%, and reveal that this percentage is consistent across all user types. Despite the heterogeneity in usage patterns of users, we also find a lack of significant variability in predictability when considering demographic factors or different mobility or mobile service usage. We also analyze the joint predictability of the traffic demands and mobility patterns. We find that the two dimensions are correlated, which improves the predictability upper bound to 90%on average. This first work is in submission in an international conference.

6.2.5. Completion of Sparse Call Detail Records for Mobility Analysis

Participants: Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, Sahar Hoteit.

Call Detail Records (CDRs) have been widely used in the last decades for studying different aspects of human mobility. The accuracy of CDRs strongly depends on the user-network interaction frequency: hence, the temporal and spatial sparsity that typically characterize CDR can introduce a bias in the mobility analysis. In this work, we evaluate the bias induced by the use of CDRs for inferring important locations of mobile subscribers, as well as their complete trajectories. Besides, we propose a novel technique for estimating real human trajectories from sparse CDRs. Compared to previous solutions in the literature, our proposed technique reduces the error between real and estimated human trajectories and at the same time shortens the temporal

period where users' locations remain undefined. This work has been published as an invited paper at the ACM CHANTS 2016 workshop in conjunction with ACM MobiCom 2016. Related to CDRs, we have also investigated whether the information of user's instantaneous whereabouts provided by CDRs enables us to estimate positions over longer time spans. Our results confirm that CDRs ensure a good estimation of radii of gyration and important locations, yet they lose some location information. Most importantly, we show that temporal completion of CDRs is straightforward and efficient: thanks to the fact that they remain fairly static before and after mobile communication activities, the majority of users' locations over time can be accurately inferred from CDRs. Finally, we observe the importance of user's context, i.e.,of the size of the current network cell, on the quality of the CDR temporal completion. This work is in submission in an international conference. Finally, driven by real-world data across a large population, we propose two approaches as the refinement of the legacy solution, which complete CDR data adaptively according to the information of users and activities. Our proposed methods outperform the legacy solution in terms of the combination of accuracy and temporal coverage. Besides, our work reveals the important factors to the data completion. This paper has been accepted for publication at the IEEE DAWM workshop in conjunction with IEEE Percom 2017.

6.2.6. Completion of Sparse Call Detail Records for Mobility Analysis

Participants: Panagiota Katsikouli, Aline Carneiro Viana, Marco Fiore, Alessandro Nordio, Alberto Tarable.

The increasing usage of smart devices and location-tracking systems has made it possible to study and understand the behaviour of users as well as human mobility at an unprecedented scale. The insights of such studies can help improve many aspects of our everyday lives, from road network infrastructure to mobile network quality of service. Human mobility is repetitive and regular.In addition to our tendency to revisit the same locations, those visits happen with relevant temporal regularity, where each visited location has been assigned with an ID. The daily interaction with our smart devices, such as smartphones, results in collecting fine grained information on our activities and whereabouts. This information can be used to detect and analyze the routinary behaviour of humans but also to discover interests, preferences and hidden patterns of mobility. However, frequent recording of data tends to quickly drain the battery of the smartphone. A natural alternative is to sample the collected data. Maintaining a summary or sample as close to the original collected data as possible is the key challenge. Deciding what constitutes a representative sample depends on the type of information we wish to maintain from the data collected. In this work, we wish to sparsely sample mobility traces of GPS data with the goal to reconstruct the movement of the users both in space and time at the desired granularity. An ideal sample would allow us to reconstruct the traces in such a way that we preserve the frequency of visits and the time spent to the various locations. Therefore, the problem we tackle here is to *sparsely sample the mobility trace of a user with the goal to reconstruct her complete trace in space and time*. This is an on-going work and will be submitted to an international conference in the next months.

6.3. Resource and Traffic Management

Traffic offloading; infrastructure deployment; opportunistic routing; traffic modeling; intermittently connected networks.

6.3.1. Utility Optimization Approach to Network Cache Design

Participants: Mostafa Dehghan, Laurent Massoulié, Don Towsley, Daniel Menasche, Y.c. Tay.

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 Project-Team

6. New Results

6.1. Monitoring

6.1.1. Quality of Experience Monitoring

Participants: Isabelle Chrisment [contact], Thibault Cholez, Vassili Rivron.

We have pursued our work on smartphone usage monitoring. In [26], we presented an exploratory smartphone usage study with logs collected from users in the wild, combined with the sociodemographic, technological and cultural information provided by them. We have shown that application usage among users is highly diverse. However when the applications are grouped as services, interesting relations appear between user profiles and types of services used. We found significant correlations between service usage and socio-demographic profile. We have proposed several possible use cases of how sociological information can be used to renew the official statistics, to recommend suitable applications to potential users.

6.1.2. Active Monitoring

Participants: Abdelkader Lahmadi [contact], Jérôme François, Frédéric Beck [LHS], Loic Rouch [LHS].

Following preliminary work in 2015, we pursued our assessment of industrial system exposition in the Internet. Industrial systems are composed of multiple components whose security has not been addressed for a while. Even if recent propositions target to improve it, they are still often exposed to vulnerabilities, since their components are hard to update or replace. In parallel, they tend to be more and more exposed in the public Internet for convenience. Although awareness of such a problem has been raised, there is no precise evaluation of such a risk. We thus defined a methodology to measure the exposure of industrial systems through Internet. In particular, a carefully designed scanning approach and software with a low footprint, named WiScan, consists in optimizing the distance between consecutively scanned IP addresses but being fast to compute. It has been applied on the entire IPv4 address space, by targeting specific SCADA ports. This work is reported in [20].

During the year 2016, we are also working with the regional PME TracIP http://www.tracip.fr on the development of attack assessment and forensics platform dedicated to industrial control system. The platform involves multiple PLC from different manufactures and real devices of factory automation systems.

6.1.3. SDN enhanced monitoring

Participants: Jérôme François [contact], Lautaro Dolberg [University of Luxembourg].

Software-Defined Networking (SDN) provides a highly flexible flow management platform through a logically centralized controller that exposes network capabilities to the applications. However, most applications do not natively use SDN. An external entity is thus responsible for defining the corresponding flow management policies. This is mainly the role of the network administrator, which also prefers to keep the control of its network rather than fully let the control to users or applications.

Usually network operators prefer to control the flow management policies, rather than granting full control to the applications. Although IP addresses and port numbers can suffice to identify users and applications in ISP networks and determine the policies applicable to their flows, such an assumption does not hold strongly in cloud environments. IP addresses are allocated dynamically to the users, while port numbers can be freely chosen by users or cloud-based applications. These applications, like computing or storage frameworks, use diverse port numbers which amplifies this phenomenon. We have proposed higher-level abstractions for defining user- and application-specific policies. In this scope, our framework transparently maps application-level policies (involving application and user names) to OpenFlow rules (IP addresses, protocols and port numbers), which alleviates the necessity for the control applications (those interacting with the Northbound

interface of the controller) to keep track of the network characteristics (like location) of users and applications themselves. To achieve this end, application-level information is retrieved in real-time through local remote system agents, which can be easily deployed in a cloud platform where both network and computational infrastructure are hosted by the same entity.

Thus our work enables the association of flows with applications and users. It led to a publication [19].

6.1.4. Service-level Monitoring of HTTPS traffic

Participants: Thibault Cholez [contact], Shbair Wazen, Jérôme François, Isabelle Chrisment.

We previously investigated the latest technique for HTTPS traffic filtering that is based on the Server Name Indication (SNI) field of TLS and which has been recently implemented in many firewall solutions. We showed that SNI has two weaknesses, regarding (1) backward compatibility and (2) multiple services using a single certificate. On the other side, HTTPS proxy suffers from privacy issues by decrypting users' sensitive traffic. This led us to the development of new reliable methods to investigate the increasing number of HTTPS traffic with a proper level of identification (service-level) that allows the management of the traffic while other methods are either too broad (protocol-lvl identification) or too precise (page-level identification).

We proposed to improve HTTPS traffic monitoring based on SNI. Our investigation shows that 92% of the HTTPS websites surveyed can be accessed with fake-SNI. Our approach verifies the coherence between the real destination server and the claimed value of SNI by relying on a trusted DNS service. Experimental results show the ability to overcome the shortage of SNI-based monitoring by detecting forged SNI values while having a very small false positive rate (1.7%). The overhead of our solution only adds negligible delays to access HTTPS websites. The proposed method opens the door to improve global HTTPS monitoring and firewall systems and was published in the IEEE STAM workshop [31].

We 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. This work was published in IFIP/IEEE NOMS [30] and is now extended in the frame of the CNRS PEPS NEFAE project to address the challenge of real-time monitoring of HTTPS. We extract statistical features on TLS handshake packets and progressively on application data packets, so that we can identify HTTPS services very early in the session. Extensive experiments performed over a significant and open dataset show that our method offers a good accuracy and a prototype implementation confirms that the real-time requirement of monitoring HTTPS services is satisfied.

6.1.5. Sensor networks monitoring

Participants: Rémi Badonnel, Isabelle Chrisment, Olivier Festor, Abdelkader Lahmadi [contact], Anthea Mayzaud.

This year, we have pursued our work on IoT security monitoring, based on our distributed architecture specified in [24]. This one exploits the multi-instance mechanisms of the RPL protocol, to build a monitoring plane using high-order nodes, in the context of Advanced Metering Infrastructures (AMI). It permits to preserve more constrained node resources, by passively monitoring the network. We have shown in [23] its benefits for detecting version number attacks. A DODAG versioning system is incorporated into the RPL protocol, in order to ensure an optimized topology. However, an attacker can exploit this mechanism to damage the network and reduce its lifetime. We have therefore proposed a detection strategy with a set of algorithms capable of identifying malicious nodes performing such attacks. We have evaluated our solution through experiments and have analyzed the performance according to defined metrics. We have shown that false positive rates can be reduced by a strategic monitoring node placement. In particular, we have addressed scalability considerations, as an optimization problem which can be easily adapted to different topologies. By resolving this problem, we were able to quantify the number of monitoring nodes required to ensure an acceptable false positive rate for different topologies.

Our taxonomy on security attacks in these networks has also been published in [2]. The RPL protocol is exposed to a large variety of attacks, whose consequences can be quite significant in terms of network performance and resources. The attacks against resources reduce network lifetime through the generation of fake control messages or the building of loops. The attacks against the topology make the network converge to a sub-optimal configuration or isolated nodes. Attacks against network traffic let a malicious node capture and analyse large part of the traffic. This classification serves as a support to prioritize attacks depending on the damages they may cause to the network, and can be exploited for risk management purposes in order to select counter-measures.

6.2. Security

6.2.1. Security analytics

Participants: Jérôme François [contact], Abdelkader Lahmadi, Giulia de Santis, Marc Coudriau, Olivier Festor.

During 2016, active collaboration with the High Security Lab in Nancy continues especially in the context of the FUI HuMa project. First we developed a method to automatically analyze darknet data. A darknet or telescope is a whole subnetwork, which is announced over Internet such that packets sent to the IP addresses are properly routed over but not replied to. In our case, the darknet is a /20 network meaning that we monitor 2^{12} addresses. The main challenge we faced was to cope with the volume of data in order to extract inter-twined phenomena characterized by groups of packets. We proposed a clustering and visualisation method derived from the Mapper algorithm that has been developed in the field of Topological Data Analysis (TDA). The developed method and its associated tool are able to analyze a large number of IP packets in order to make malicious activity patterns easily observable by security analysts. The results show that our method is able to exhibit observable patterns which have been missed by Suricata, a widely used State-of-the-Art IDS https://hal.inria.fr/hal-01403950/document.

Second scannings have been particularly studied as they represent the first phase of recognition in advanced persistent threats. While it is known that every exposed systems is always being actively scanned from multiple sources, it is still challenging to fingerprint them, in particular to identify what are the distributed sources of a single synchronized scan and what is the tool used to generate it. As a first step, we proposed a methodology based on Hidden Markov Models (HMMs) to model scanning activities monitored by a darknet [18]. The HMMs of scanning activities are built on the basis of the number of scanned IP addresses within a time window and fitted using mixtures of Poisson distributions.

We are also still maintaining an IRTF draft [50] to promote a standardization effort towards the extension of IP Flow-based monitoring with geographic information. Associating Flow information with their measurement points geographic locations will enable security applications to detect anomalous activities. In the case of mobile devices, the characterization of communication patterns using only time and volume is not enough to detect unusual location-related communication patterns. The draft went through an IRSG review and a feedback is still required from the OPSWAG IETF working group.

6.2.2. DDoS Signaling

Participants: Jérôme François [contact], Abdelkader Lahmadi, Giovane Moura [SIDN Labs, Netherland], Marco Davids [SIDN Labs, Netherlands].

A distributed denial-of-service (DDoS) attack aims at rendering machines or network resources unavailable. These attacks have grown in frequency, intensity and target diversity. In the context of Flamingo, Madynes contributed to the definition of an opportunistic signaling protocol in cooperation with SIDN Labs in Netherlands. The goal is to provide an efficient mechanism where nodes in an IPv6 network facing a DDoS attack can deliver a DOTS (DDoS Open Threat Signaling) signal message sent by a DOTS client to the DOTS server. The specified mechanism does not generate transport packets to carry the DOTS signal message but it only relies on existing IPv6 packets in the network to include within them a hop-by-hop extension header which contains an encoded DOTS signal message.

This work is done under the umbrella of our standardization activities especially within the IETF DOTS working group [45] and was presented during IETF 96 in Berlin.

6.2.3. NDN Security

Participants: Thibault Cholez [contact], Xavier Marchal, Olivier Festor.

Named-Data Networking (NDN) is one of the most advanced ICN architecture but the security of NDN or NFD (NDN Forwarding Deamin) is still in the early stages and not ready for real deployment. In the context of the ANR Doctor project, we investigate NDN security in order to unveil issues and propose remediations.

First, we discovered a new vulnerability of NDN which is easy to exploit and can lead to very serious attacks, badly affecting the network. This vulnerability is due to an absence of control at the precise moment when NFD receives an incoming Data. In fact, NFD only checks two points: if the Data belongs to the localhost scope, or if it matches an existing PIT entry, but not if the Data comes from a valid Face. This is a critical shortage because it allows malicious users to directly send Data to perform attacks like DoS and cache poisoning without having to register a prefix in the router's FIB beforehand to receive legitimate Interests. After these checks, NFD continues to process the Data packet. The NDN protocol makes the hypothesis that a node cannot send a Data packet without having previously received the corresponding Interest (receiver driven communication). However, NFD should consider malicious nodes that decide to not follow the standard way to proceed with NDN communications and send Data unexpectedly. We further described two serious attack scenarios exploiting this vulnerability based on the fact that malicious nodes can send unexpected Data that can consume legitimate PIT entries. We also propose two ways to prevent it with minor modifications in NFD. This work was published and demonstrated at the ACM-ICN conference [46].

Second, we addressed the Content Poisoning Attack (CPA), known to be one of the major threats in NDN. 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 the context of the ANR Doctor Project, and in collaboration with UTT, we demonstrated through an experimental measurement campaign that CPA can easily and widely affect NDN. We proposed three realistic attack scenarios relying on both protocol design and implementation weaknesses and presented their implementation and evaluation in a testbed based on the latest NFD version. We analyzed their impact on the different ICN nodes composing a realistic topology (clients, access and core routers, content provider) in order to fully characterize the CPA. This work has been accepted and will be published in IFIP/IEEE IM 2017 conference.

6.2.4. Configuration security automation

Participants: Rémi Badonnel [contact], Abdelkader Lahmadi, Olivier Festor, Nicolas Schnepf, Maxime Compastie.

We have pursued during year 2016 our efforts on the orchestration of security functions in the context of mobile smart environments, with a joint work with Stephan Merz of the VeriDis project-team at Inria Nancy. In particular, Nicolas Schnepf studied during his Master thesis formal techniques for the automatic verification of chains of security functions in a setting of software-defined networks (SDN). Concretely, he defined an extension of the Pyretic language [51] which takes into account the data plane of SDN controllers and implemented a translation of that extension to the input languages of the nuXmv model checker and of SMT solvers. The approach and its scalability were validated over crafted security chains, and a conference paper describing the results is going to be submitted shortly. Nicolas Schnepf started a PhD thesis on the same topic in October 2016 with joint supervision by members of the Madynes and VeriDis Inria project-teams.

In addition, we have analyzed and evaluated the usage of OpenFlow messages for security applications [29], jointly with Bundeswehr University of Munich. The purpose was to quantify the performances of security solutions that are built on top of software-defined networking infrastructures. We have considered overloading attacks and information gathering attacks, that are quite common in these environments, and have detailed regular and sdn-based mitigation mechanisms that have been designed for tackling them. We have then analyzed for each category the dependencies of these mechanisms to the OpenFlow protocol

commonly supporting the communications between sdn controllers and switches. These dependencies have been identified through the mapping of OpenFlow messages to security functionalities in that context. Based on this analysis, we performed series of experiments on two different testbeds for comparing and evaluating the accuracy and reliability that can be expected with respect to these messages.

We have also investigated in [16] a software- defined security framework, for supporting the enforcement of security policies in distributed cloud environments. These latter require security mechanisms able to address their multi-tenancy and multi-cloud properties. 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. It exploits a security orchestrator, policy decision points (PDP) and policy enforcement point (PEP) interacting according to a dedicated set of protocols, and will take advantage of facilities offered by unikernel and micro-services techniques to reduce the security exposure of cloud resources. The proposed framework has been evaluated so far through a set of validation scenarios corresponding to a realistic use cases including cloud resource allocation/deallocation, cloud resource state change, and dynamic access control.

6.3. Experimentation, Emulation, Reproducible Research

This section covers our work on experimentation on testbeds (mainly Grid'5000), on emulation (mainly on Distem), and on Reproducible Research.

6.3.1. Grid'5000 design and evolutions

Participants: Jérémie Gaidamour, Arthur Garnier, Lucas Nussbaum [contact], Clément Parisot, Florent Didier.

The team was again heavily involved in the evolutions and the governance of the Grid'5000 testbed.

First, we finished the installation and setup of several new clusters in the Nancy site, which brought several new local users, from various teams, to the testbed.

In the context of ADT LAPLACE, Jérémie Gaidamour added support for the control of CPU parameters such as Hyperthreading, Turboboost, P-states and C-states. It is expected that this work will enable interesting usages in the areas of HPC runtimes and energy-aware computing.

Finally, in the context of his roles in the *bureau*, *comité d'architectes* and *comité des responsables de sites* of Grid'5000, Lucas Nussbaum managed the purchase of the new clusters at Nancy mentioned above, and gave several presentations about the testbed, at the *Grid'5000 School* [5] [38], at a GENI-FIRE collaboration workshop [9], [8], [6], [7].

6.3.2. Emulation with Distem

Participants: Emmanuel Jeanvoine, Lucas Nussbaum [contact], Cristian Ruiz.

Several improvements have been made around Distem, mostly in the context of ADT COSETTE.

A paper demonstrating the use of Distem to evaluate fault tolerance and load balancing strategies implemented in Charm++ was accepted at CCGrid'2016 [28].

We continued our work on using Distem to experiment on NDN infrastructures. We obtained promising results, especially in terms of scale. This work is still pending publication.

Finally, we also evaluated the porting of Distem to other testbeds (ChameleonCloud and CloudLab), which was interesting for Distem specifically, but also to understand differences between those testbeds [43].

6.3.3. Management of experiments

Participants: Tomasz Buchert, Emmanuel Jeanvoine, Lucas Nussbaum [contact], Cristian Ruiz.

We continued work on Ruby-Cute, a library that aggregates various useful functionality in the context of such tools. Several releases were made in 2016. We hope that it will be useful as a basis for future tools, and ease testing of new ideas in that field. The library is available on http://ruby-cute.github.io/.

Tomasz Buchert defended his PhD thesis, entitled *Managing large-scale, distributed systems research experiments with control-flows*, in January 2016 [1].

6.3.4. Experimentation methodologies on Big Data

Participants: Abdulqawi Saif, Lucas Nussbaum [contact], Ye-Qiong Song [contact].

Abdulqawi Saif started his PhD on experimentation methodologies for Big Data at the end of 2015. His first work [35] is a multi-criteria analysis of NFS performance using statistical Design of Experiments techniques.

6.4. Routing

6.4.1. Probabilistic Energy-Aware Routing for Wireless Sensor Networks

Participants: Evangelia Tsiontsiou, Bernardetta Addis, Alberto Ceselli [Universita degli Studi di Milano], Ye-Qiong Song [contact].

Healthcare applications are considered as promising fields for Wireless Sensor Networks (WSNs) and globally IoT. Thanks to WSNs, patients can be monitored in hospitals or smart home environments, providing health improvement, or emergency care. Network lifetime is still the key issue when we deploy wireless sensor networks and IoT solutions in real-world applications. Current WSN research trends include duty-cycling at MAC layer and energy efficient routing at network layer, among others. We proposed an Optimal Probabilistic Energy-Aware Routing Protocol (OPEAR) for duty-cycled WSNs which aims at maximizing the network lifetime by keeping low energy consumption and balancing network traffic between nodes. Our experimental campaign reveals that our OPEAR protocol outperforms the popular Energy Aware Routing Protocol (EAR) from the literature, proving to be more effective in extending the network lifetime [33]. It is part of Lorraine AME Satelor project granted by Lorraine Region.

6.4.2. NDN router with P4

Participants: Salvatore Signorello [University of Luxembourg], Olivier Festor [contact], Radu State [University of Luxembourg], Jérôme François.

Although content-awareness at the network level is becoming more and more needed, Information-Centric Networking (ICN)-based solutions struggle to emerge. Research on ICN has already produced insightful outputs, nevertheless architecture-tied designs of ICN devices cannot be easily deployed and tested in operational networks; further those designs are hard to share. In the meantime, the vision of Software-Defined Networking has grown and taken new shapes. Network players desire to change devices' behavior often and drastically, even though performances are still crucial to operate at line-speed. This has been leading to a rethink of network devices designs with the aim to offer full-programmability through high-level programming languages for packet processors, like P4. It is a programming language to describe the forwarding plane of network devices. The language abstracts how packets are processed by different devices in target- independent programs. Then, compilers map those programs to different forwarding devices with as final goal a single specification which can be automatically mapped to hardware or software implementations. Although high-level protocols like ICN with advanced parsing mechanisms are usually handled by software switch with standard programming capacity, P4 would allow more efficient implementation on specific platform. Our preliminary implementation strives to implement many features of the NDN routing by using native P4 language constructs only [32].

6.4.3. NDN/HTTP cohabitation

Participants: Thibault Cholez [contact], Xavier Marchal, Olivier Festor.

Network operators are reluctant to deploy globally Named Data Networking (NDN) because of the huge investment costs required and the uncertainty about the security and the manageability of such disruptive network protocols when deployed in production, while the return of investment is also uncertain. Meanwhile, Network Functions Virtualization (NFV) greatly facilitates the deployment of novel networking architectures by reducing the costs thanks to the usage of commodity hardware in place of dedicated equipments. Consequently, leveraging NFV to ease the deployment of NDN infrastructures appears as a strong mean to incite network operators to adopt this technology. In this context, the challenge we address in the ANR DOCTOR project is to fulfil the requirements needed to move NDN from a solution restricted to labs or tesbeds to a fully operational one by developing NDN-specific Virtual Network Functions (VNF).

In this effort, one of the main first questions which arise is about the integration of NDN into the existing Internet, and particularly the collocation of NDN with IP and HTTP. We think that a good way to deploy NDN consists in creating virtualized NDN island that can be crossed by specific content-related traffic, such as HTTP, and thus benefit from NDN properties (caching, aggregation, etc.). We proposed and developed an early version of a fully-capable NDN/HTTP gateway that can seamlessly connect a NDN network to the rest of the World Wide Web. This work was published and demonstrated at the ACM-ICN conference [47].

6.5. Multi-modeling and co-simulation

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Yannick Presse, Julien Vaubourg, Alexandre Tan, Benjamin Segault, Thomas Paris.

Vincent Chevrier (former Maia team, Dep 5, LORIA) is a collaborator and the correspondant for the MS4SG/MECSYCO project, Benjamin Camus, and Christine Bourjot (former MAIA team, Dep 5, LORIA) are collaborators for AA4MM/MECSYCO. Julien Vaubourg and Thomas Paris's PhDs are under the co-direction of V. Chevrier and L. Ciarletta.

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 on 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 [52] 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) projet 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.

In 2016 we worked on the following research topics:

- Assessment and evaluation of complex systems.
- Cyber Physical Systems.

  We have pursued the design and implementation of the Aetournos platform at Loria. The collective movements of a flock of flying communicating robots / UAVs, evolving in potentially perturbed environment constitute a good example of a Cyber Physical System.

We have maintained thanks to the ADT UASS a set of tools: multi-simulation behavior / network / physics and generic software development using ROS (Robot Operating System). The UAVs carry a set of sensors for location awareness, their own computing capabilities and several wireless networks.

- MS4SG / MECSYCO-NG opportunity to link simulations tools with a strong focus on FMI (Functional Mockup Interface) and network simulators (NS3/Omnet++) thanks to our MECSYCO (formerly AA4MM) framework. 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 major improvements in 2016 thanks to the ressources provided by the MECSYCO-NG project in collaboration with EDF R&D (http://www.mecsyco.com).

  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.

  We have developed software components and plugins that interconnects within MECSYCO heterogeneous simulators from different domains: FMU (working with the 1.0 and 2.0 FMI standard for CoSimulation) ou non-FMU such as NS3 or Omnet++.

  Several EDF oriented advanced use cases have demonstrated multi-simulations.

In addition to technical reports [41], several publications have been accepted in 2016 on these subjects [25], [13] and [34].

6.6. Pervasive or Ubiquious Computing

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Emmanuel Nataf, Thomas Paris, Benjamin Segault, Antoine Richard, Petro Aksonenko.

P. Aksonenko PhD is under the co-direction of L. Ciarletta and Patrick Henaff from Loria Dep 5. Thomas Gurriet, now PhD student at Georgia Tech under the supervision of Prs Eric Feron and Aaron Ames is contributing to the topic of CPS safety.

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, increasingly numerous and heterogeneous, are embedded in the fabric of our daily lives. Our initial subject of interest is to study them with regards to their complexity: Those 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 their surroundings and the environment. Secondly, the potential multiplicity and heterogeneity of devices, services, communication protocols, and the constant mobility and reorganization also need to be addressed. Thirdly we are aking into account their dynamcity, with regards to their mobility and evolving context.

Our research on 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.

In 2016 we mainly worked on the Cyber Physical Systems.

We maintained the Aetournos platform at Loria in collaboration with 6PO and the support of ADT UASS. We are studying the collective movements of a flock of flying communicating robots / UAVs, evolving in potentially perturbed environment constitute a good example of a Cyber Physical System.

The effort put in the UAVs gathers academic and research ressources from the Aetournos platform, the Inria ADT R2D2 and the 6PO project, while applied, industrial and more R&D projects have been pursued this year (Medical Express / Outback Joe Search and Rescue Challenge, Alerion, Hydradrone, and a CIFRE PhD with Thales for example) .

This also led to two new accepted projects:

- one Interreg “Grone”, a generic project about drones in industrial and agricultural environments, started in October 2016
- and one FUI22 “CEOS”, about insuring safety in UAVs at the system level that will start in 2017

  One of the emerging topic in this area is the safety of Mobile IoT / CPS with regards to their environment and users. This gave first results on how to design the internal communication system [21], the overal system [15], specific safety solutions [14] and a US Patent has been filled on a termination system led by Georgia Tech [Optimal Emergency Termination System for Unmanned Aerial Vehicles by Destructive Rotor Surface Reduction, Application No.: 62/378,923].
- Smart * (MECSYCO)

  We have studied scientific problems around models and simulators composition. We have also looked into practical and implementation issues in the frame of our MECSYCO /AA4MM solutions. We have added to our Smart Grid scenarios a smart appartment complex use case.
- (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.

6.7. Quality-of-Service

6.7.1. Self-adaptive MAC protocol for both QoS and energy efficiency

Participants: Kévin Roussel, Shuguo Zhuo, Olivier Zendra, Ye-Qiong Song [contact].

WSN research focus has progressively been moved from the energy issue to the QoS issue. Typical example is the MAC protocol design, which cares about not only low duty-cycle at light traffic, but also high throughput with self-adaptation to dynamic traffic bursts.

We have mainly contributed to enhancing the implementation of the high efficient traffic self-adaptive MAC protocols. As part of RIOT ADT project, we have improved and implemented a fully functional iQueue-MAC which provides not only the unique feature of high traffic self-adaptivity, but also the robustness by using two control channels (https://github.com/RIOT-OS/RIOT/pull/5618).

As part of LAR project, we were interested by using the Cooja/MSPSim network simulation framework for RIOT OS based platforms. We have showed that Cooja is not limited only to the simulation of the Contiki OS based systems and networks, but can also be extended to perform simulation experiments of other OS based platforms, especially that with RIOT OS. Moreover, when performing our own simulations with Cooja and MSPSim, we observed timing inconsistencies with identical experimentations made on actual hardware. Such inaccuracies clearly impair the use of the Cooja/MSPSim framework as a performance evaluation tool, at least for time-related performance parameters. The detailed results of our investigations on the inaccuracy problems, as well as the consequences of this issue, and possible ways to fix or avoid it are available in [27].

6.7.2. QoS and fault-tolerance in distributed real-time systems

Participants: Florian Greff, Laurent Ciarletta, Arnauld Samama [Thales TRT], Eric Dujardin [Thales TRT], 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 started a study on the real-time dependability of distributed multi-criticity systems interconnected by an embedded mesh network (RapidIO). For easing the QoS specification at the higher level, DDS middleware is used. We postulate that enhancing QoS for real-time applications entails the development of a cross-layer support of high-level requirements, thus requiring a deep knowledge of the underlying networks. This year, we proposed and implemented a new simulation/emulation/experimentation framework called ERICA, for designing such a feature. ERICA integrates both a network simulator (Ptolemy) and an actual hardware network to allow implementation and evaluation of different QoS-guaranteeing mechanisms. It also supports real-software-in-the-loop, i.e. running of real applications and middleware over these networks [21].

We have also dealt with mesh networking of embedded components. Our approach is to allow applications to make online real-time flow resource requests and consequently dynamically allot network resources according to these requirements. To this end, additional mechanisms must be provided in order to meet the real-time constraints while the platform remains as dynamic as possible. We gather these mechanisms into a Software-Defined Real-time Network (SDRN) paradigm. The online admission control and pathfinding algorithms have been developed allowing the controller to dynamically configure the real-time network nodes. We have evaluated several pathfinding algorithms.

6.7.3. Wireless sensor and actuator networks

Participants: Lei Mo, Adrian Guenard, Yifei Qi [Zhejiang University], Jiming Chen [Zhejiang University], Ye-Qiong Song [contact].

Wireless sensor and actuator networks provide a key technology for fully interacting within a CPS (Cyber-Physical System). However, the introduction of the mobile actuator nodes in a network rises some new challenging issues. In this context, we addressed two important issues: the multiple target tracking using both fixed and mobile sensors and the optimal scheduling of mobile wireless energy chargers (actuators) for fixed sensor nodes.

In the low-cost and large-scale deployment of mobile sensor nodes for target tracking, due to the constraints of limited sensing range, it is of great importance to design node coordination mechanism for reliable tracking so that at least the target can always be detected with a high probability, while the total network energy cost can be reduced for longer network lifetime. In [3], we dealt with this problem considering both the unreliable wireless channel and the network energy constraint. We transfer the original problem into a dynamic coverage problem and decompose it into two subproblems. By exploiting the online estimate of target location, we first decide the locations where the mobile nodes should move into so that the reliable tracking can be guaranteed. Then, we assign different mobile nodes to each location in order that the total energy cost in terms of moving distance can be minimized. Extensive simulations under various system settings have shown the effectiveness of our solution.

We also investigated the multiple mobile chargers coordination problem that is minimizing the energy expenditure of the mobile chargers while guaranteeing the perpetual operation of the wireless sensor network. We extended our previous result (published in IPCCC2015) by taking into account mobile charger's charging ability. We formulated this problem as a mixed-integer linear program (MILP), and proposed a novel decentralized method which is based on Benders decomposition. The convergence of proposed method is analyzed theoretically. Simulation results demonstrate the effectiveness and scalability of the proposed method.

6.7.4. NDN performance evaluation

Participants: Thibault Cholez [contact], Xavier Marchal, Olivier Festor.

NDN (Named Data Networking) is a promising protocol that can help to reduce congestion at Internet scale by putting content at the center of communications instead of hosts. NDN can also natively authenticate transmitted content with a mechanism similar to website certificates that allows clients to assess the original provider. But this security feature comes at a high cost, as it relies heavily on asymmetric cryptography which affects server performance when NDN Data are generated. This is particularly critical for many services dealing with real-time data (VOIP, live streaming, etc.), but current tools are not adapted for a realistic server-side performance evaluation of NDN traffic generation when digital signature is used. We propose a new tool, NDNperf, to perform this evaluation and show that creating NDN packets is a major bottleneck of application performances. On our testbed, 14 server cores only generate $\sim$400 Mbps of new NDN Data with default packet settings. We gave recommendation about the configuration of NDN (packet size, cryptographic function) and proposed practical improvements to the NDN library that all combined can vastly increase the performance of server-side NDN Data generation (x8,5). This work was published in the ACM-ICN conference [22].

MAESTRO Project-Team

7. New Results

7.1. Network Science

Participants: Eitan Altman, Konstantin Avrachenkov, Arun Kadavankandy, Jithin Kazhuthuveettil Sreedharan, Hlib Mykhailenko, Giovanni Neglia, Alina Tuholukova.

7.1.1. Computation on Large Graphs

The MAESTRO team has been working on how to partition large graphs in distributed computation frameworks in order to speed up the execution time.

In [43], H. Mykhailenko and G. Neglia in collaboration with F. Huet (Univ. Côte d'Azur, CNRS, I3S), provide an overview of existing edge partitioning algorithms. However, based only on published work, it is not possible to draw a clear conclusion about the relative performances of these partitioners. For this reason, the authors compare all the edge partitioners currently available for the widely-used framework for graph processing Apache GraphX. Preliminary results suggest that the Hybrid-Cut partitioner provides the best performance.

In [44], H. Mykhailenko and G. Neglia in collaboration with F. Huet (Univ. Côte d'Azur, CNRS, I3S), focus on vertex-cut graph partitioning and they investigate how it is possible to evaluate the quality of a partition before running the computation. To this purpose the authors scrutinize a set of metrics proposed in literature. They carry experiments with Apache GraphX and they perform an accurate statistical analysis. Preliminary experimental results show that communication metrics like vertex-cut and communication cost are effective predictors on most of the cases.

7.1.2. Network centrality measures

In [19], K. Avrachenkov in collaboration with V. Mazalov (Karelian Institute of Applied Mathematical Research, Russia), L. Trukhina (Baikal State Univ. of Economics and Law, Russia) and B. Tsynguev (Transbaikal State Univ., Russia) worked on network centrality measures based on game-theoretic concepts. The betweenness centrality is one of the basic concepts in the analysis of the social networks. Initial definition for the betweenness of a node in the graph is based on the fraction of the number of geodesics (shortest paths) between any two nodes that given node lies on, to the total number of the shortest paths connecting these nodes. This method has polynomial complexity. We propose a new concept of the betweenness centrality for weighted graphs using the methods of cooperative game theory. The characteristic function is determined by special way for different coalitions (subsets of the graph). Two approaches are used to determine the characteristic function. In the first approach the characteristic function is determined via the number of direct and indirect weighted connecting paths in the coalition. In the second approach the coalition is considered as an electric network and the characteristic function is determined as a total current in this network. We use Kirchhoff's law. After that the betweenness centrality is determined as the Myerson value. The results of computer simulations for some examples of networks, in particular, for the popular social network "VKontakte", as well as the comparing with the PageRank method are presented.

7.1.3. Sampling and Inference of Complex Networks

In [32] K. Avrachenkov, G. Neglia and A. Tuholukova study chain-referral methods for sampling in social networks. These methods rely on subjects of the study recruiting other participants among their set of connections. This approach gives us the possibility to perform sampling when the other methods, that imply the knowledge of the whole network or its global characteristics, fail. Chain- referral methods can be implemented with random walks or crawling in the case of online social networks. However, the estimations made on the collected samples can have high variance, especially with small sample size. The other drawback is the potential bias due to the way the samples are collected. We suggest and analyze a subsampling technique, where some users are requested only to recruit other users but do not participate to the study. Assuming that

the referral has lower cost than actual participation, this technique takes advantage of exploring a larger variety of population, thus decreasing significantly the variance of the estimator. We test the method on real social networks and on synthetic ones. As by-product, we propose a Gibbs-like method for generating synthetic networks with desired properties.

Function estimation on Online Social Networks (OSN) is an important field of study in complex network analysis. An efficient way to do function estimation on large networks is to use random walks. We can then defer to the extensive theory of Markov chains to do error analysis of these estimators. In [29], K. Avrachenkov, A. Kadavankandy and J.K. Sreedharan in collaboration with V. Borkar (IIT Bombay, India) compare two existing techniques, Metropolis-Hastings MCMC and Respondent-Driven Sampling, that use random walks to do function estimation and compare them with a new reinforcement learning based technique. We provide both theoretical and empirical analyses for the estimators we consider.

In [33] K. Avrachenkov and J.K. Sreedharan in collaboration with B. Ribeiro (Purdue Univ., USA) develop random walk based methods for inference in Online Social Networks (OSNs) to answer questions like are OSN users more likely to form friendships with those with similar attributes? Do users at an OSN A score content more favorably than OSN B users? Such questions frequently arise in the context of Social Network Analysis (SNA) but often crawling an OSN network via its Application Programming Interface (API) is the only way to gather data from a third party. To date, these partial API crawls are the majority of public datasets and the synonym of lack of statistical guarantees in incomplete-data comparisons, severely limiting SNA research progress. Using regenerative properties of the random walks, we propose estimation techniques based on short crawls that have proven statistical guarantees. Moreover, our short crawls can be implemented in massively distributed algorithms. We also provide an adaptive crawler that makes our method parameter-free, significantly improving our statistical guarantees. We then derive the Bayesian approximation of the posterior of the estimates, and in addition, obtain an estimator for the expected value of node and edge statistics in an equivalent configuration model or Chung-Lu random graph model of the given network (where nodes are connected randomly) and use it as a basis for testing null hypotheses. The theoretical results are supported with simulations on a variety of real-world networks.

In [30] K. Avrachenkov in collaboration with L. Iskhakov and M. Mironov (Moscow Institute of Physics and Technology, Russia) consider pairwise Markov random fields which have a number of important applications in statistical physics, image processing and machine learning such as Ising model and labeling problem to name a couple. Our own motivation comes from the need to produce synthetic models for social networks with attributes. First, we give conditions for rapid mixing of the associated Glauber dynamics and consider interesting particular cases. Then, for pairwise Markov random fields with submodular energy functions we construct monotone perfect simulation.

7.1.4. Distributed algorithms for complex network analysis

In [31] K. Avrachenkov and J.K. Sreedharan in collaboration with P. Jacquet (Nokia Bell Labs, France) address the problem of finding top-k eigenvalues and corresponding eigenvectors of symmetric graph matrices in networks in a distributed way. We propose a novel idea called complex power iterations in order to decompose the eigenvalues and eigenvectors at node level, analogous to time-frequency analysis in signal processing. At each node, eigenvalues correspond to the frequencies of spectral peaks and respective eigenvector components are the amplitudes at those points. Based on complex power iterations and motivated from fluid diffusion processes in networks, we devise distributed algorithms with different orders of approximation. We also introduce a Monte Carlo technique with gossiping which substantially reduces the computational overhead. An equivalent parallel random walk algorithm is also presented. We validate the algorithms with simulations on real-world networks. Our formulation of the spectral decomposition can be easily adapted to a simple algorithm based on quantum random walks. With the advent of quantum computing, the proposed quantum algorithm will be extremely useful.

In [56] K. Avrachenkov in collaboration with V. Borkar and K. Saboo (IIT Bombay, India) propose two asynchronously distributed approaches for graph-based semi-supervised learning. The first approach is based on stochastic approximation, whereas the second approach is based on randomized Kaczmarz algorithm. In

addition to the possibility of distributed implementation, both approaches can be naturally applied online to streaming data. We analyse both approaches theoretically and by experiments. It appears that there is no clear winner and we provide indications about cases of superiority for each approach.

7.1.5. Random Matrix Theory for Complex Networks

In [41] A. Kadavankandy and K. Avrachenkov in collaboration with L. Cottatellucci (Eurecom, France) describe a test statistic based on the L1-norm of the eigenvectors of a modularity matrix to detect the presence of an embedded Erdos-Renyi (ER) subgraph inside a larger ER random graph. An embedded subgraph may model a hidden community in a large network such as a social network or a computer network. We make use of the properties of the asymptotic distribution of eigenvectors of random graphs to derive the distribution of the test statistic under certain conditions on the subgraph size and edge probabilities. We show that the distributions differ sufficiently for well defined ranges of subgraph sizes and edge probabilities of the background graph and the subgraph. This method can have applications where it is sufficient to know whether there is an anomaly in a given graph without the need to infer its location. The results we derive on the distribution of the components of the eigenvector may also be useful to detect the subgraph nodes.

7.1.6. Network Growth Models

Network growth and evolution is a fundamental theme that has puzzled scientists for the past decades. A number of models have been proposed to capture important properties of real networks. In an attempt to better describe reality, more recent growth models embody local rules of attachment, however they still require a primitive to randomly select an existing network node and then some kind of global knowledge about the network (at least the set of nodes and how to reach them). In [28] G. Neglia, in collaboration with B. Amorim, D. Figueiredo and G. Iacobelli (Federal Univ. of Rio de Janeiro, Brazil), proposes a purely local network growth model that makes no use of global sampling across the nodes. The model is based on a continuously moving random walk that after s steps connects a new node to its current location, but never restarts. Through extensive simulations and theoretical arguments, they analyze the behavior of the model finding a fundamental dependency on the parity of s, where networks with either exponential or a conditional power law degree distribution can emerge. As s increases parity dependency diminishes and the model recovers the degree distribution of Barabási-Albert preferential attachment model. The proposed purely local model indicates that networks can grow to exhibit interesting properties even in the absence of any global rule, such as global node sampling.

7.1.7. Competition over popularity in online social networks

In [24] E. Altman in collaboration with A. Jain and Y. Hayel (UAPV) consider a stochastic game that describes competition through advertisement over the popularity of their content. They show that the equilibrium may or may not be unique, depending on the system's parameters. They identify structural properties of the equilibria. In particular, they show that a finite improvement property holds on the best response pure policies which implies the existence of pure equilibria. They further show that all pure equilibria are fully ordered in the performance they provide to the players and propose a procedure to obtain the best equilibrium.

7.1.8. Trend detection in social networks using Hawkes processes

In [18], J. C. Louzada Pinto and T. Chahed from Telecom SudParis in collaboration with E. Altman propose a general Hawkes-based framework to model information diffusion in social networks. The proposed framework takes into consideration the hidden interactions between users as well as the interactions between contents and social networks, and can also accommodate dynamic social networks and various temporal effects of the diffusion, which provides a complete analysis of the hidden influences in social networks. This framework can be combined with topic modeling, for which modified collapsed Gibbs sampling and variational Bayes techniques are derived. We provide an estimation algorithm based on nonnegative tensor factorization techniques, which together with a dimensionality reduction argument are able to discover the latent community structure of the social network. We provide numerical examples from real-life networks: a Game of Thrones and a MemeTracker datasets.

7.1.9. Potential Game approach to defense against virus attacks in networks

The Susceptible-Infected-Susceptible (SIS) model is a classical epidemic model where agents alternate between a sane (susceptible) and an infected state. SIS epidemic non-zero sum games have been recently used to analyse virus protection in networks. A potential game approach was proposed for solving the game for the case of a fully connected network. In [42], F.-X. Legenvre and Y. Hayel (UAPV) in collaboration with E. Altman extend this result to an arbitrary topology by showing that the general topology game is a generalized ordinal potential game. We apply this result to study numerically some examples.

7.2. Wireless Networks

Participants: Sara Alouf, Eitan Altman, Giovanni Neglia, Alina Tuholukova.

7.2.1. Control of Delay–Tolerant Networks

In [5] E. Altman and G. Neglia, in collaboration with F. De Pellegrini (Create-Net, Italy) and D. Miorandi (U-Hopper, Italy), study optimal stochastic control of delay tolerant networks. First, the structure of optimal two-hop forwarding policies is derived. In order to be implemented, such policies require knowledge of certain global system parameters such as the number of mobiles or the rate of contacts between mobiles. But, such parameters could be unknown at system design time or may even change over time. In order to address this problem, adaptive policies are designed that combine estimation and control: based on stochastic approximation techniques, such policies are proved to achieve optimal performance in spite of lack of global information. Furthermore, the paper studies interactions that may occur in the presence of several DTNs which compete for the access to a gateway node. The latter problem is formulated as a cost-coupled stochastic game and a unique Nash equilibrium is found. Such equilibrium corresponds to the system configuration in which each DTN adopts the optimal forwarding policy determined for the single network problem.

7.2.2. Performance Evaluation of Train Moving-Block Control

In moving block systems for railway transportation a central controller periodically communicates to the train how far it can safely advance. On-board automatic protection mechanisms stop the train if no message is received during a given time window. In [45], [63] G. Neglia, S. Alouf, and A. Tuholukova in collaboration with A. Dandoush (SME Sudria, France, formerly engineer with MAESTRO) and S. Simoens, P. Dersin, J. Billion and P. Derouet (all from ALSTOM Transport) consider as reference a typical implementation of moving-block control for metro and quantify the rate of spurious Emergency Brakes (EBs), i.e. of train stops due to communication losses and not to an actual risk of collision. Such unexpected EBs can happen at any point on the track and are a major service disturbance.

The general formula for the EB rate found in [45] requires a probabilistic characterization of losses and delays. Calculations are surprisingly simple in the case of homogeneous and independent packet losses. More complex loss scenarios are studied in [59]. The approach is computationally efficient even when emergency brakes are very rare (as they should be) and can no longer be estimated via discrete-event simulations.

The analytical models have also been validated using ns-3 simulations [35].

7.2.3. Speed estimation

After several years of cooperation with Nokia (formerly Alcatel-Lucent) Bell Labs in developing tools for speed estimation from measurement of the radio channel, we have now started to publish our joint patented work. This includes the work on mobility state estimation in LTE by D.-G. Herculea, V. Capdevielle, C. S. Chen, N. Ben Rached and F. Ratovelomanana from Nokia-Bell Labs in collaboration with E. Altman and M. Haddad (UAPV), see [38].

7.2.4. Sonorous cartography for sighted and blind people

E. Altman has been invited by D. Josselin from UMR Espace in UAPV to co-advise a Master project and later a thesis financed by the CNRS on Sonorous cartography. Other persons with whom we collaborate are D. Roussel, S. Boularouk, A. Saidi, M. Driss (from UAPV) and O. Bonin (Laboratoire Ville, Mobilité, Transport) all coauthors of [40] which won the best short paper award in the AGILE conference. In this article, we test the usability of a cartographic tool mixing maps and sounds. This tool is developed within QuantumGIS as a plugin prototype. We first present some theoretical reflections about synesthesia. Secondly, we explain the way we "sonificate" the images, by associating colors and recorded chords and sounds. Then we present the results of several usability tests in France with different users, including blind people.

To help blind people compensate visual perception and to better understand their outdoor environment, S. Boularouk and D. Josselin from UAPV in collaboration with E. Altman, proposed in [49] a method using human-computer interaction via Text-to-Speach. It helps visually impaired people to know surrounding places from OpenStreetMap data by hearing. The principal idea is to convey spatial information by voice synthesis and receive requests from blind people by voice recognition.

7.2.5. Scheduling for mobile users with non-stationary mobility

H. Zaaraoui and Z. Altman from Orange Labs in collaboration with T. Jiménez (UAPV) and E. Altman have studied scheduling in an environment with non-stationary mobility (cars are moving on a road and may have to stop at red lights). They propose scheduling schemes for such mobility patterns and study their performance in in [55] and in [48].

7.2.6. User Association in Multi-user MIMO Small Cell Networks

Dense Networks and large MIMO are two key enablers to achieve high data rates towards next generation 5G networks. In this context, S. Ramanath (Lekha Wireless Solutions and IIT Mumbai) and M. Debbah (Huawei) in collaboration with E. Altman study in [47] user association in an interference limited Multiuser MIMO Small Cell Network. Extending on previous findings, they derive explicit expressions for the optimal ratio of the number of antennas at the base station to the number of users that can associate to a base station in such a Network. The expressions are used to compute the actual number of users that can associate for a given interference level and other system parameters. Simulation results and numerical examples are provided to support our theoretical findings.

7.3. Network Engineering Games

Participants: Eitan Altman, Konstantin Avrachenkov, Giovanni Neglia, Nessrine Trabelsi.

7.3.1. Network formation games

Network formation games have been proposed as a tool to explain the topological characteristics of existing networks. They assume that each node is an autonomous decision-maker, ignoring that in many cases different nodes are under the control of the same authority (e.g. an Autonomous System) and then they operate as a team. In [11] K. Avrachenkov and G. Neglia in collaboration with V.V. Singh (LRI, Univ. Paris-Sud, France) introduce the concept of network formation games for teams of nodes and show how very different network structures can arise also for some simple games studied in the literature. Beside extending the usual definition of pairwise stable networks to this new setting, we define a more general concept of stability toward deviations from a specific set C of teams' coalitions (C-stability). We study then a trembling-hand dynamics, where at each time a coalition of teams can create or sever links in order to reduce its cost, but it can also take wrong decisions with some small probability. We show that this stochastic dynamics selects C-stable networks or networks from closed cycles in the long run as the error probability vanishes.

7.3.2. Routing Games

A central question in routing games has been to establish conditions for uniqueness of the equilibrium, in terms of network topology or of costs. This question is well understood in two classes of routing games. In [27], E. Altman and C. Touati (Inria Grenoble - Rhône-Alpes) study two other frameworks of routing games in which each of several players has an integer number of connections (which are population of packets) to route and where there is a constraint that a connection cannot be split. Through a particular game with a simple three link topology, we identify various novel and surprising properties of games within these frameworks. We show in particular that equilibria are non unique even in the potential game setting of Rosenthal with strictly convex link costs.

7.3.3. Game theory applied to the Internet and social networks

In [25] E. Altman, A. Jain (UAPV) and C. Touati (Inria Grenoble - Rhône-Alpes) in collaboration with N. Shimkin (Technion), present an overview of the use of dynamic games for analyzing competition in the Internet and in on-line social networks. A special emphasis is put on identifying phenomena and tools that are novel with respet to game theory applied to other types of networks.

7.3.4. Resilience of Routing in Parallel Link Networks

E. Altman, C. Touati and A. Singhal (Inria Grenoble - Rhône-Alpes), in collaboration with J. Li (Tsukuba Univ. Japan), use a game approach in [26] to study the resilience problem of routing traffic in a parallel link network with a malicious player. They consider two players: the first wishes to split its traffic so as to minimize its average delay, which the second player, i.e., the malicious player, tries to maximize. The first player has a demand constraint on the total traffic it routes. The second player controls the link capacities: it can decrease by some amount the capacity of each link under a constraint on the sum of capacity degradation. We first show that the average delay function is convex both in traffic and in capacity degradation over the parallel links and thus does not have a saddle point. We identify best responses strategies of each player and compute both the max-min and the min-max values of the game. We provide stable algorithms for computing both max-min and min-max strategies as well as for best responses.

7.3.5. A game theoretic solution for Resource Allocation in LTE Cellular Networks

Due to Orthogonal Frequency Division Multiple Access (OFDMA) mechanism adopted in LTE cellular networks, intra-cell interference is nearly absent. Yet, as these networks are designed for a frequency reuse factor of 1 to maximize the utilization of the licensed bandwidth, inter-cell interference coordination remains an important challenge. In both homogeneous and heterogeneous cellular networks, there is a need for scheduling coordination techniques to efficiently distribute the resources and mitigate inter-cell interference. In [54], N. Trabelsi and E. Altman in collaboration with C. S. Chen, L. Roullet from Nokia Bell Labs and with R. El-Azouzi from UAPV propose a dynamic solution of inter-cell interference coordination performing an optimization of frequency sub-band reuse and transmission power in order to maximize the overall network utility. The proposed framework, based on game theory, permits to dynamically define frequency and transmission power patterns for each cell in the coordinated cluster.

7.4. Green Networking and Smart Grids

Participants: Sara Alouf, Eitan Altman, Alain Jean-Marie, Giovanni Neglia, Dimitra Politaki.

7.4.1. Power Demand Control

Demand-Response (DR) programs, whereby users of an electricity network are encouraged by economic incentives to rearrange their consumption in order to reduce production costs, are envisioned to be a key feature of the smart grid paradigm. Several recent works proposed DR mechanisms and used analytical models to derive optimal incentives. Most of these works, however, rely on a macroscopic description of the population that does not model individual choices of users. In [34], [57] G. Neglia and A. Benegiamo (PhD student in MAESTRO at the submission time), in collaboration with P. Loiseau, conduct a detailed analysis of those models and argue that the macroscopic descriptions hide important assumptions that can jeopardize

the mechanisms' implementation (such as the ability to make personalized offers and to perfectly estimate the demand that is moved from a timeslot to another). Then, they start from a microscopic description that explicitly models each user's decision. They introduce four DR mechanisms with various assumptions on the provider's capabilities. Contrarily to previous studies, they find that the optimization problems that result from these mechanisms are not convex. Local optimizers can be found numerically through a heuristic. The authors present numerical simulations that compare the different mechanisms and their sensitivity to forecast errors. At a high level, their results show that the performance of DR mechanisms under reasonable assumptions on the provider's capabilities are significantly lower than those suggested by previous studies, but that the gap reduces when the population's flexibility increases.

In [22] A. Jean-Marie and G. Neglia in collaboration with I. Tinnirello, L. Giarré, M. Ippolito (Univ. of Palermo, Italy) and G. Di Bella (Telecom Italia, Italy) investigate a realistic and low-cost deployment of large scale direct control of inelastic home appliances whose energy demand cannot be shaped, but simply deferred. The idea is to exploit 1) some simple actuators to be placed on the electric plugs for connecting or disconnecting appliances with heterogeneous control interfaces, including non-smart appliances, and 2) the Internet connections of customers for transporting the activation requests from the actuators to a centralized controller. The solution requires no interaction with home users: in particular, it does not require them to express their energy demand in advance. A queuing theory model is derived to quantify how many users should adopt this solution in order to control a significant aggregated power load without significantly impairing their quality of service.

7.4.2. Geographical Load Balancing across Green Datacenters

"Geographic Load Balancing" is a strategy for reducing the energy cost of data centers spreading across different terrestrial locations. In [20] G. Neglia, in collaboration with M. Sereno (Univ. of Torino, Italy) and G. Bianchi (Univ. of Roma "Tor Vergata", Italy), focuses on load balancing among micro-datacenters powered by renewable energy sources. They model via a Markov Chain the problem of scheduling jobs by prioritizing datacenters where renewable energy is currently available. Not finding a convenient closed form solution for the resulting chain, they use mean field techniques to derive an asymptotic approximate model which instead is shown to have an extremely simple and intuitive steady state solution. After proving, using both theoretical and discrete event simulation results, that the system performance converges to the asymptotic model for an increasing number of datacenters, they exploit the simple closed form model's solution to investigate relationships and trade-offs among the various system parameters.

7.4.3. Stochastic models for solar energy

The recent popularization of renewable energy sources makes it urgent to have realistic and practical models for the renewable energy harvested by photovoltaic panels for instance. Solar radiation is intrinsically stochastic and exhibits fluctuations at several time scales. Due to the sun's position during the day with respect to a given point on Earth, there is a periodic day-night pattern that is observed on top of which short-time burstiness occurs due to fluctuating weather conditions. In [64], D. Politaki and S. Alouf propose a stochastic model for the global solar radiation. They introduce a multiplicative factor that is the ratio between the actual global solar radiation and the idealized clear sky global radiation. The latter is obtained using known astronomical models and captures the day-night pattern of the solar radiation at any given point on Earth. On the other hand, the multiplicative factor captures the short-time burstiness caused by cloudiness. A semi-Markov model is proposed for the latter such that most of the time correlation found in measured data can be reproduced in synthetic traces.

7.5. Content-Oriented Systems

Participants: Sara Alouf, Eitan Altman, Konstantin Avrachenkov, Philippe Nain, Giovanni Neglia, Dimitra Tsigkari.

7.5.1. Modeling modern DNS caches

In-network caching is a widely adopted technique to provide an efficient access to data or resources on a world-wide deployed system while ensuring scalability and availability. In previous years, S. Alouf and N. Choungmo Fofack (former PhD student at MAESTRO, currently at Ingima) have focused on hierarchical systems that rely on expiration-based policies to manage their caches. Each cache in the system maintains for each item a timer that indicates its duration of validity. The Domain Name System (DNS) is a valid application case. The objective was to assess the performance of a polytree of caches. This work has now been published in [4].

7.5.2. Caching policies

In [46], [60], G. Neglia and D. Tsigkari, in collaboration with D. Carra (Univ. of Verona), M. Feng (Akamai Technologies), V. Janardhan (Akamai Technologies) and P. Michiardi (Eurecom), present a new cache replacement policy that takes advantage of a hierarchical caching architecture, and, in particular, of access-time difference between memory and hard disk. They prove that the proposed policy is optimal when requests follow the independent reference model, and significantly reduces the hard-disk load, as they show through their realistic trace-driven evaluation.

7.5.3. Analyzing Caching and Shaping Timeline Networks

Cache networks are one of the building blocks of information centric networks (ICNs). Most of the recent work on cache networks has focused on networks of request driven caches, which are populated based on users requests for content generated by publishers. However, user generated content still poses the most pressing challenges. For such content timelines are the de facto sharing solution. In [53], A. Reiffers-Masson (PhD student in MAESTRO at the time of submission) and E. Altman in collaboration with E. Hargreaves, W. Caarls and D. Sadoc Menasché from UFRJ (Brazil) establish a connection between timelines and publisher-driven caches. We propose simple models and metrics to analyze publisher-driven caches, allowing for variable-sized objects. Then, we design two efficient algorithms for timeline workload shaping, leveraging admission and price control in order, for instance, to aid service providers to attain prescribed service level agreements.

7.5.4. Cooperative view on Caching

The non-cooperative nature of relations between economic actors in todays networks may lead to inefficiencies and may not provide incentives for investing in deploying new technologies. In [36] E. Altman in cooperation with V. Douros and S. Elayoubi (Orange Labs) in collaboration with Y. Hayel (UAPV) have studied the question of how to split costs for deploying caches between Content Providers and Internet Service Providers. They have designed the cost sharing by casting the problem into a coalition game which they solved using the Shapely value concept.

7.5.5. Streaming optimization

The Quality of Experience (QoE) of streaming service is often degraded by frequent play-back interruptions. To mitigate the interruptions, the media player prefetches streaming contents before starting playback, at a cost of initial delay. In [23], Y. Yu and Y. Yu from Fudan Univ. in collaboration with S. Elayoubi (Orange Labs) R. El-Azouzi (UAPV) and E. Altman, study the QoE of streaming from the perspective of flow dynamics. Firstly, a framework is developed for QoE when streaming users join the network randomly and leave after downloading completion. We model the distribution of prefetching delay using partial differential equations (PDEs), and the probability generating function of playout buffer starvations using ordinary differential equations (ODEs) for constant bit-rate (CBR) streaming. Explicit form starvation probabilities and mean start-up delay are obtained. Secondly, we extend our framework to characterize the throughput variation caused by opportunistic scheduling at the base station, and the playback variation of variable bit-rate (VBR) streaming. Our study reveals that the flow dynamics is the fundamental reason of playback starvation. The QoE of streaming service is dominated by the first moments such as the average throughput of opportunistic scheduling and the mean playback rate. While the variances of throughput and playback rate have very limited impact on starvation behavior in practice.

7.6. Advances in Methodological Tools

Participants: Eitan Altman, Konstantin Avrachenkov, Alain Jean-Marie.

7.6.1. Control theory

Linear programming formulations for the discounted and long-run average Markov Decision Processes have evolved along separate trajectories. In 2006, E. Altman conjectured that the linear programming formulations of these two models are, most likely, a manifestation of general properties of singularly perturbed linear programs. In [8] K. Avrachenkov in collaboration with J. Filar and A. Stillman (Flinders Univ., Australia) and V. Gaitsgory (Macquarie Univ., Australia) demonstrate that this is, indeed, the case.

A. Jean-Marie, together with E. Hyon (Univ. Paris-Ouest Nanterre La Défense), completed the analysis of optimal admission control in a single-server queue with impatience. In the presence of a server startup cost, linear holding costs for the queue and individual costs for departures due to impatience, the optimal policy is to either serve customers whenever some are present, or never serve any customer. The situation is decided by a simple criterion comparing the cost of starting the server to a combination of the other parameters. Proving the optimality of such a simple policy is more difficult than expected, and involves the propagation of properties through the dynamic programming operator of a suitably approximated sequence of problems, following methods and results of Blok, Bhulai and Spieksma.

7.6.2. Game theory

7.6.2.1. Uniqueness of equilibrium

E. Altman in cooperation with M. Kumar (IIT Mumbai) and R. Sundaresan (IICs) have derived in [6] a new sufficient condition for uniqueness of equilibrium which extends the Diagonal Strict Concavity condition of Rosen. They further apply the condition to various networking examples.

7.6.2.2. Hybrid games

In collaboration with V. Gaitsgory, I. Brunetti (former member of MAESTRO) and E. Altman have studied in [15] a non-zero sum game in which there are two components of the state space: one is a finite (controlled) Markov chain and the other is a vector of real numbers. Only the Markov chain is controlled; the other part of the state space evolves according to some differential equations whose parameters are the state and actions of the Markov chain. The authors have shown the existence of an asymptotic stationary equilibrium. They show how to derive epsilon equilibria policies for the original problem based on policies that are asymptotically equilibria.

7.6.2.3. Finite games

In [13] K. Avrachenkov in collaboration with V.V. Singh (LRI, Univ. Paris-Sud 11, France) consider coalition formation among players in an n-player finite strategic game over infinite horizon. At each time a randomly formed coalition makes a joint deviation from a current action profile such that at new action profile all the players from the coalition are strictly benefited. Such deviations define a coalitional better-response (CBR) dynamics that is in general stochastic. The CBR dynamics either converges to a K-stable equilibrium or becomes stuck in a closed cycle. We also assume that at each time a selected coalition makes mistake in deviation with small probability that add mutations (perturbations) into CBR dynamics. We prove that all K-stable equilibria and all action profiles from closed cycles, that have minimum stochastic potential, are stochastically stable. Similar statement holds for strict K-stable equilibria. We apply the CBR dynamics to study the dynamic formation of the networks in the presence of mutations. Under the CBR dynamics all strongly stable networks and closed cycles of networks are stochastically stable.

7.6.2.4. Dynamic Games

In a collaboration with M. Tidball (INRA, France), A. Jean-Marie considered the extension of an infinite-horizon dynamic game of groundwater extension [51], due to Provencher and Burt. As usual in this kind of models, the marginal extraction cost depends on the level of the groundwater. The goal of this paper is to point out the importance of the moment where this cost is announced to the players. We consider the case where the cost is announced before the extraction is made and the case where is announced after extractions. For both

cases, we also analyse the possibility of taking into account the rainfall or not. The current literature considers only the case where the cost is announced before rain and harvesting. We characterize the equilibrium in the linear-quadratic case. We compare solutions as functions of the discount factor, with the particular cases of zero discount (myopic model) and no discount (maximization of the steady state) from the economic and the environmental points of view. We show that when the level of the groundwater is small, announcing costs after harvesting and rainfall is better from the economic and environmental point of view than the case of announcing it before harvesting and rainfall.

7.6.3. Queueing Theory

7.6.3.1. Retrial queues

In [10] K. Avrachenkov in collaboration with E. Morozov (Karelian Institute of Applied Mathematical Research, Russia) and B. Steyaert (Gent Univ., Belgium) study multi-class retrial queueing systems with Poisson inputs, general service times, and an arbitrary numbers of servers and waiting places. A class-i blocked customer joins orbit i and waits in the orbit for retrial. Orbit i works like a single-server $\cdot/M/1$ queueing system with exponential retrial time regardless of the orbit size. Such retrial systems are referred to as retrial systems with constant retrial rate. Our model is motivated by several telecommunication applications, such as wireless multi-access systems, optical networks and transmission control protocols, but represents independent theoretical interest as well. Using a regenerative approach, we provide sufficient stability conditions which have a clear probabilistic interpretation. We show that the provided sufficient conditions are in fact also necessary, in the case of a single-server system without waiting space and in the case of symmetric classes. We also discuss a very interesting case, when one orbit is unstable, whereas the rest of the system is stable.

In [9] K. Avrachenkov in collaboration with E. Morozov, R. Nekrasova (Karelian Institute of Applied Mathematical Research, Russia), and B. Steyaert (Gent Univ., Belgium) study the stability of a single-server retrial queueing system with constant retrial rate, general input and service processes. First, we present a review of some relevant recent results related to the stability criteria of similar systems. Sufficient stability conditions were obtained by (Avrachenkov and Morozov, 2014), which hold for a rather general retrial system. However, only in case of Poisson input an explicit expression is provided; otherwise one has to rely on simulation. On the other hand, the stability criteria derived by (Lillo, 1996) can be easily computed, but only hold for the case of exponential service times. We present new sufficient stability conditions, which are less tight than the ones obtained by (Avrachenkov and Morozov, 2010), but have an analytical expression under rather general assumptions. A key assumption is that interarrival times belongs to the class of *new better than used* (NBU) distributions. We illustrate the accuracy of the condition based on this assumption (in comparison with known conditions when possible) for a number of non-exponential distributions.

7.6.3.2. Polling Systems

In [12] K. Avrachenkov in collaboration with E. Perel and U. Yechiali (Tel Aviv Univ., Israel) consider a system of two separate finite-buffer $M/M/1$ queues served by a single server, where the switching mechanism between the queues is threshold-based, determined by the queue which is not being served. Applications may be found in data centers, smart traffic-light control and human behavior. We analyse both work-conserving and non-work-conserving policies. We present occasions where the non-work-conserving policy is more economical than the work-conserving policy when high switching costs are involved. An intrinsic feature of the process is an oscillation phenomenon: when the occupancy of one queue decreases, the occupancy of the other queue increases. This fact is illustrated and discussed. By formulating the system as a three-dimensional continuous-time Markov chain we provide a probabilistic analysis of the system and investigate the effects of buffer sizes and arrival rates, as well as service rates, on the system's performance. Numerical examples are presented and extreme cases are investigated.

MUSE Team

6. New Results

6.1. Home Network or Access Link? Locating Last-mile Downstream Throughput Bottlenecks

Participants: Srikanth Sundaresan (ICSI), Nick Feamster (Princeton), Renata Teixeira

As home networks see increasingly faster downstream throughput speeds, a natural question is whether users are benefiting from these faster speeds or simply facing performance bottlenecks in their own home networks. We studied the problem whether downstream throughput bottlenecks occur more frequently in their home networks or in their access ISPs. We identified lightweight metrics that can accurately identify whether a throughput bottleneck lies inside or outside a user's home network and developed a detection algorithm that locates these bottlenecks. We validated this algorithm in controlled settings and characterized bottlenecks on two deployments, one of which included 2,652 homes across the United States. We found that wireless bottlenecks are more common than access- link bottlenecks—particularly for home networks with downstream throughput greater than 20 Mbps, where access-link bottlenecks are relatively rare.

6.2. Characterizing Home Device Usage From Wireless Traffic Time Series

Participants: Katsiaryna Mirylenka (IBM), Vassilis Christophides, Themis Palpanas (University Rene Descartes), Ioannis Pefkianakis (HP Labs), Martin May (Technicolor)

We conducted a thorough analysis of traffic dynamics of heterogeneous wireless (WiFi) devices connected to 196 real RGWs, which are subscribers of a major European ISP. We focus on a time-oriented analysis of continuous traffic data to extract previously unknown patterns recurring of internet consumption that happen within, or across homes. We also assess the impact of different types of devices, such as laptops, desktops (classiffied as fixed devices), and tablets, smartphones (classiffied as portables), on these patterns. Unsupervised learning techniques are used for patterns discovery as the ground truth data regarding home activities are not available. Rather than partitioning homes or devices into distinct behavioral clusters, we are looking to extract informative motifs of bandwidth consumption within or across homes. The main contributions of this work are:

- We propose a novel analysis framework for wireless home traffic data, namely: (a) a correlation-based similarity measure, which exploits the evolution characteristics, rather than the absolute traffic values, and is invariant to scaling;(b) a notion of strong stationarity that in addition to the similarity of data distributions imposes a correlation similarity across non-overlapping time windows; and (d) a definition of dominant devices based on the correlation similarity, that enables an intuitive and statistically grounded interpretation of the results.
- We evaluate the effectiveness of the proposed framework using real data of wireless traffic observations and report the main findings: (a) there are many repetitive patterns within and across RGWs which describe the intrinsic user behavior of users and valuable to ISPs; (b) as networking time series are not stationary certain aggregation should be performed in order to find statistically significant patterns. The best time windows to aggregate home traffic data is found to be 8 hours for weekly patterns and 3 hours for daily patterns; (c) frequent weekly patterns correspond to heavy bandwidth usage both during weekdays and weekends, and frequent daily patterns correspond to (mostly) evening usage, (d) weekend usage tends to rely on portable devices, weekday usage relies more on fixed devices, while discontinuous usage within a day (mostly active in the evening or the morning) is still due to portable devices; and (e) almost every RGW involves a device that dominates its overall traffic, thus the behavior of this device should be mainly considered by ISPs while planning the updates.

6.3. Towards a Causal Analysis of Video QoE from Network and Application QoS

Participants: Michalis Katsarakis, Renata Teixeira, Maria Papadopouli (Univeristy of Crete), Vassilis Christophides

We have exploited an original framework for mining causal relationships among a 5-start rating of user QoE and various QoS metrics at network and application level. In particular, we have analysed QoE scores provided by a set of users for YouTube video streaming applications under different network conditions. We found that optimal QoE predictors we can be build using a minimal signature of only three features from application or network QoS metrics compared to four when features from both layers are considered. A thorough comparative analysis of the prediction accuracy of three models build using minimal signatures composed of (i) only network QoS, (ii) only application QoS, and (iii) both QoS features demonstrated that we can predict the QoE using only network QoS metrics and more surprisingly, predicting the QoE from network QoS metrics is as accurate as when using application QoS metrics. This work is the first step towards our ambition to assess QoE directly from network QoS metrics obtained via passive measurements of real traffic generated by online users. We will rely on the extracted minimal QoE/QoS signatures to build real-time predictors and compare their accuracy when using only network, only application or both QoS metrics. Last but not least, we plan to extend our experimental setting for other online applications such as teleconferencing services.

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

Participants: Diego Neves da Hora, Renata Teixeira, Karel Van Doorselaer (Technicolor), Koen Van Oost (Technicolor)

We developed a model that predicts the effect of Wi-Fi quality on Web QoE, using solely Wi-Fi metrics commonly available in commercial APs. We trained our predictor during controlled experiments on a Wi-Fi testbed and assess its accuracy through cross-validation, obtaining an RMSE of 0.6432 MO, and by applying it on a separate validation dataset, obtained on an uncontrolled environment, finding an RMSE of 0.9283. Finally, we apply our predictor on Wi-Fi metrics collected in the wild from 4,880 APs over a period of 40 days. We find that Wi-Fi quality is mostly good for Web—in more than 60% of samples Wi-Fi quality does not degrade Web QoE. When we consider average complexity Web pages, however, Wi-Fi quality degrades Web QoE in 11% of samples. Moreover, we saw that 21% of devices present more than 20% of poor Web QoE samples, with 5% of these showing highly intermittent QoE degradations, which are particularly hard to diagnose, indicating the need for a long-term monitoring approach to detect and fix problems.

6.5. Passive Wi-Fi Link Capacity Estimation on Commodity Access Points

Participants: Diego Neves da Hora, Karel Van Doorselaer (Technicolor), Koen Van Oost (Technicolor), Renata Teixeira, Christophe Diot (Safran)

We propose an algorithm to estimate the link capacity based on passive metrics from APs, which is ready to be deployed at scale. We show that it is possible to estimate the link capacity per PHY rate based on a limited set of parameters related to the particular AP instance. Then, we extend the initial model to estimate the link capacity when the PHY rate varies. We measured the link capacity in different link quality conditions and found that more than 90% of the estimations present error below 15% without prior parameter tuning, and more than 95% present estimation error below 5% with appropriate parameter tuning using fixed PHY rate tests.

6.6. Content-Based Publish/Subscribe System for Web Syndication

Participants: Zeinab Hmedeh CNAM, Harry Kourdounakis (FORTH-ICS, Vassilis Christophides, Cedric du Mouza (CNAM), Michel Scholl (CNAM), and Nicolas Travers (CNAM)

Content syndication has become a popular way for timely delivery of frequently updated information on the Web. Today, web syndication technologies such as RSS or Atom are used in a wide variety of applications spreading from large-scale news broadcasting to medium-scale information sharing in scientific and professional communities. However, they exhibit serious limitations for dealing with information overload in Web 2.0. There is a vital need for efficient real-time filtering methods across feeds, to allow users to effectively follow personally interesting information.

To efficiently check whether all keywords of a subscription also appear in an incoming item (i.e., broad match semantics), we need to index the subscriptions. Count-based (CI) and tree-based (TI) are two main indexing schemes proposed in the literature for counting explicitly and implicitly the number of contained key-words. The majority of related data structures cannot be employed for conjunctions of keywords (rather than attribute-value pairs) due to the space high-dimensionality. In this paper, we are interested in efficient implementations of both indexing schemes using inverted lists (IL) for CI and a variant for distinct terms of ordered tries (OT) for TI and study their behavior for critical parameters of realistic web syndication workloads. Although these data structures have been employed to evaluate broad match queries in the context of selective information dissemination and sponsored search or for mining frequent item sets, their memory and matching time requirements appear to be quite different in our setting. This is due to the peculiarities of web syndication systems which are characterized 1) by information items of average length (25?36 distinct terms) which are greater than advertisement bids (4?5 terms) and smaller than documents of Web collections (12K terms) and 2) by very large vocabularies of terms (up to 1.5M terms). Note also that due to broad match semantics, information retrieval techniques for optimizing ILs (e.g., early pruning) are not suited in our setting.

We present analytical models for memory requirements and matching time and we conduct a thorough experimental evaluation to exhibit the impact of critical parameters of realistic web syndication workloads. We found that for small vocabularies, POT matching time is one order of magnitude faster than the best IL (RIL), while for large vocabularies (like the one used on the Web), RIL outperforms the matching POT, which uses almost four times more memory space. The actual distribution of term occurrences has almost no impact on the size of the three indexing structures while it significantly affects the number of nodes that need to be visited upon matching something that justifies OT performance gains. The smaller the subscription length, the larger the OT factorization gain w.r.t. IL and the larger the rank of the term from which the OT substructure degenerates to an IL.

RAP Project-Team

4. New Results

4.1. Random Graphs

Participant: Nicolas Broutin.

Self-similar real trees defined as fixed-points [15]: Random trees that are fixed points of some random decompositions are ubiquitous: the essential building blocks of the scaling limits of graphs, but also various other trees associated to combinatorial models are such trees. We study a general class of fixed-points equations in spaces of measure metric spaces that yield such objects, and study the existence/uniqueness of the fixed-points in the natural spaces of interest. We also obtain geometric information such as fractal dimension or estimates about the degrees directly from the equations. This is joint work with Henning Sulzbach.

4.2. Resource Allocation in Large Data Centres

Participants: Christine Fricker, Philippe Robert, Guilherme Thompson, Veronica Quintuna Rodriguez.

Efficient resource allocation in large data centers has become crucial matter since the expansion in volume and in variety of the internet based services and applications. Everyday examples, such as Video-on-Demand and Cloud Computing are part of this change in the internet environment, bringing new perspectives and challenges with it. Resource pooling (gathering resources to avoid idleness) and resource decentralization (to bring the service “closer" to the user) are too an important topic in service design, specially because of the inherent dichotomy presented in this discussion. Understanding and assessing the performance of such systems ought enable to better resource management and, consequently, better quality of service.

Currently, most systems operate under decentralized policies due to the complexity of managing data exchange on large scale. In such systems, customer demands are served respecting their initial service requirements (a certain video quality, amount of memory or processing power etc.) until the system reaches saturation, which then leads to the blockage of subsequent customer demands. Strategies that rely on the scheduling of tasks are often not suitable to address this load balancing problem as the users expect instantaneous service usage in real time applications, such as video transmission and elastic computation. Our research goal is to understand and redesign its algorithms in order to develop decentralized schemes that can improve global performance using local instantaneous information. This research is made in collaboration with Fabrice Guillemin, from Orange Labs.

In a first approach to this problem, we examined offloading schemes in fog computing context, where one data centers are installed at the edge of the network. We analyze the case with one data center close to user which is backed up by a central (usually bigger) data center. In [10], when a request arrives at an overloaded data center, it is forwarded to the other data center with a given probability, in order to help dealing with saturation and reducing the rejection of requests. In [17], we studied another scheme, where requests are systematically forwarded by the small data to a larger one, but with some trunk reservation to ensure service performance in the second one. We have been able to demonstrate the behavior and performance of these systems, using the invariant distribution of a random walks in the quarter plane, and obtaining explicit expressions for both schemes. Those two papers shed some light in the effectiveness of this fog computing design, by investigating two basic and intuitive policies, whose advantages can now be compared.

In [11] and [16], we investigated allocation schemes which consist in reducing the bandwidth of arriving requests to a minimal value. In the first, this process is initiated when the system is saturated and in the second when the system is close to saturation. We analyzed the effectiveness of such a downgrading policies. In the case of downgrading at saturation, we were able to find an explicit expression of the key performance metrics when two types of customers share a resource and type two asks for the double of resources compared to type one. And, for the second case, we could show that if the system is correctly designed then we can stop losing clients. We developed a mathematical model which allows us to predict system behavior under such a policy and calculate the optimal threshold (in the same scale as the resource) after which downgrading should be initiated. We proved the existence of a unique equilibrium point, around which we have been able to determine the probability a customer receives service at requested quality. We have also shown that system blockage becomes indeed negligible. This policy finds a natural application in the framework of video streaming services and other real time applications. Notably, we are able to derive explicit and simple expressions for many aspects of this system, giving special predictability the outcome of such policy.

Recently, we started to investigate the framework of network function virtualization, another emergent stream stream of research in resource allocation. We start by considering the execution of Virtualized Network Functions (VNFs) in data centers whose capacities are limited and service execution time is constrained by telecommunication protocols. Virtualization practices play a crucial role in the evolution of telecommunications network archiÂtectures, since the service providers can reduce the investment on the edge and share resource more efficiently. Macrofunctions are virtuzatized into micro ones and treat individually. Through simulations and basic mathematical models, we aroused the discussion of three different prioritization policies and their *trade-offs*. The have shown that in for parallelizable macrofunctions (i.e. no order of execution), the greedy algorithm ensures the best performance in terms of execution delay. For chained ones, macrofunctions whose microfunctions need to be run in a certain order, this algorithm is not suitable, the Round Robin and the Dedicated Core policies perform with the same level.

With these results in mind, we have extend our research towards more complex systems, investigating the behaviour of multiple resource systems (such as a Cloud environment, where computational power is provided using unities of CPU and GB of RAM). We analyzed cooperation between data centers offering multiple resources and under imbalanced loads, a problem that naturally arises from the decentralization of resources. Again, we consider instantaneous service. By forwarding some clients across the system, we could design a policy that is allows cooperation between system and preserves service quality at both data centers. We consider two types of demands asking for two types of resources; particularly, type one clients demand more of type one resource (and symmetrically for type two). We have shown that under our forwarding scheme, which offloads clients requiring most of the saturated resource locally at each data center, we can eliminate losses (in a well design system). Some other interesting properties that can help systems designers are as well derived, such as the minimum threshold for the sustainability of such scheme and the offloading rates. A document is being written to further publication.

4.3. Ressource allocation in vehicle sharing systems

Participants: Christine Fricker, Hanene Mohamed, Thanh-Huy Nguyen.

Vehicle sharing systems are becoming an urban mode of transportation, and launched in many cities, as Velib' and Autolib' in Paris. One of the major issues is the availability of the resources: vehicles or free slots to return them. These systems became an important topic in Operation Research and now the importance of stochasticity on the system behavior is commonly admitted. The problem is to understand the system behavior and how to manage these systems in order to provide both resources to users.

Our stochastic model is the first one taking into account the finite number of spots at the stations.

Equivalence of ensembles We used limit local theorems to obtain the asymptotic stationary joint distributions of several station states when the system is large (both numbers of stations and bikes), in the case of finite capacities of the stations. This gives the asymptotic independence property for node states. This widely extends the existing results on heterogeneous bike-sharing systems.

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. 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. It allows to compare local choice, no choice and choice between two chosen at random.

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 limit as the system is large, the so-called mean-field limit, and its equilibrium point. Results on performance mainly involve an original closed form expression of the stationary blocking probability in the classical join-the-shortest-queue model. These results are compared by simulations with the policy where the users choose the least loaded station between two stations to return close to their destination. It turns out that, because of randomness, the choice between two neighbours gives better performance than grouping stations two by two.

Bike-sharing model with waiting In real systems, if the customer does not find the resource (a bike or an place to return), he can either leave, or search in a neighbouring station, or wait. We extend a basic model to take into account waiting.

4.4. Scaling Methods

Participants: Philippe Robert, Wen Sun.

4.4.1. Fluid Limits in Wireless Networks

This is a collaboration with Amandine Veber (CMAP, École Polytechnique). The goal is to investigate the stability properties of wireless networks when the bandwidth allocated to a node is proportional to a function of its backlog: if a node of this network has x requests to transmit, then it receives a fraction of the capacity proportional to $\log(1+x)$, the logarithm of its current load. This year we completed the analysis of a star network topology with multiple nodes. Several scalings were used to describe the fluid limit behaviour.

4.4.2. Large Unreliable Stochastic Networks

The reliability of a large distributed system is studied. The framework is a system where files have several copies on different servers. When one of these servers breaks down, all copies stored on it are lost. These copies can be retrieved afterwards if there is another copy of the same files stored on other servers. In the case where no other copy of a given file is present in the system, it definitely lost. We study two math model on this problem.

In the first model, it is assumed that the duplication process is local, any server has a capacity to make copies to another server, but the capacity can only be used for the copies present on this server. We have studied the asymptotic behavior of this system, i.e. the number of servers is large, via mean field methods. We have shown that asymptotically, the load of each server can be described by a non-linear Markov process. This limiting process can also give an exponential decay of the number of files. This is a joint work with Reza Aghajani, Brown University.

In the second model, two policies for the reassignment of files are studied. It is assumed that each server has a neighborhood, that consists of a set of servers in the system. When a server breaks down, it restarts immediately but empty. Copies on it are reassigned to other servers in the neighborhood, following "Random Choice" (RC) policy or "Power of choices" (PoC) policy.

- (RC) Each copy join a server in the neighborhood at random.
- (PoC) Each copy chooses several servers in the neighborhood 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. This is a joint work with Inria/UPMC Team Regal.

4.5. Stochastic Models of Biological Networks

Participants: Renaud Dessalles, Sarah Eugene, Philippe Robert, Wen Sun.

4.5.1. 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.

4.5.1.1. Feedback model

We have also investigated the production of a single protein, with the transcription and the translation steps, but we also introduced a direct feedback on it: the protein tends to bind on the promoter of its own gene, blocking therefore the transcription. The protein remains on it during an exponential time until its detachment caused by thermal agitation.

The mathematical analysis aims at understanding the nature of the internal noise of the system and to quantify it. We tend to test the hypothesis usually made that such feedback permits a noise reduction of protein distribution compared to the "open loop" model. We have made the mathematical analysis of the model (using a scaling to be able to have explicit results), it appeared that reduction of variance compared to an "open loop" model is limited: the variance cannot be reduced for more than 50%.

We proposed another possible effect of the feedback loop: the return to equilibrium is faster in the case of a feedback model compared to the open loop model. Such behaviour can be beneficial for the bacteria to change of command for a new level of production of a particular protein (due, for example, to a radical change in the environment) by reducing the respond time to reach this new average. This study has been mainly performed by simulation and it has been shown that the feedback model can go 50% faster than the open loop results.

4.5.1.2. 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.

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.

4.5.1.3. Multi-protein Model

In continuation of the previous models, we propose a model that still considers the division and the gene replication but which also integrates the sharing of common resources: the different genes are in competition for the limited quantity of RNA-polymerases and ribosomes in order to produce the mRNAs and proteins. The goal is to examine if fluctuations in the availability of these macromolecules have an important impact on the protein variability, as it has been suggested in literature. As the model considers the interaction between the different protein productions, one needs to represent all the genes of the bacteria altogether: it is therefore a multi-protein model.

As this model is too complex to be studied analytically, we have developed a procedure to estimate the parameters so that they correspond to real experimental measures. We then perform simulations in order to determine the variance of each protein and compare them with the one predicted by the models with cell cycle previously presented. It appears that the common sharing of RNA-polymerases and ribosomes has a limited impact on the protein production: for most of proteins the variance increases of at most 10%.

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.

4.5.2. 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.

The second part concerns the study of telomeres. This work is made in collaboration with Zhou Xu, Teresa Teixeira, from IBCP in Paris.
In eukaryotic cells, at each mitosis, chromosomes are shortened, because the DNA polymerase is not able to duplicate one ending of the chromosome. To prevent loss of genetic information- which could be catastrophic for the cell-chromosomes are equipped with telomeres at their endings. These telomeres do not contain any genetic information; they are a repetition of the sequence T-T-A-G-G-G thousands times. At each mitosis, there is therefore a loss of telomere. As it has a finite length, when the telomeres are too short, the cell cannot divide anymore: they enter in replicative senescence. Our model tries to captures the two phases of the shortening of telomeres: first, the initial state of the cells, when the telomerase is still active to repair the telomeres. Second, when the telomerase is inhibited, we try to estimate the senescence threshold, when the replication of the cells stops. See [8].

SOCRATE Project-Team

6. New Results

6.1. Flexible Radio Front-End

6.1.1. Wake-Up Radio

The last decades have been really hungry in new ways to reduce energy consumption. That is especially true when talking about wireless sensor networks in general and home multimedia networks in particular, since electrical energy consumption is the bottleneck of the network. One of the most energy-consuming functional block of an equipment is the radio front end, and methods to switch it off during the time intervals where it is not active must be implemented. This previous study has proposed a wake-up radio circuit which is capable of both addressing and waking up not only a more efficient but also more energy-consuming radio front end. By using a frequency footprint to differentiate each sensor, awaking all the sensors except for the one of interest is avoided. The particularity of the proposed wake-up receiver [22] is that the decision is taken in the radio-frequency part and no baseband treatment is needed. The global evaluation in theory and in simulation was performed, and a first testbed of this technology was fabricated, demonstrating that this principle actually works in practice [21].

6.1.1.1. Full-Duplex

An important work was done in this axis previously around Full-Duplex systems, in order to enhance throughput, flexibility, and, potentially security of wireless links. A PhD thesis grant from DGA and Inria has allowed us to extend this through a collaboration with axis 2, focusing on Physical layer security mechanisms based on Full-Duplex systems. Starting by a theoretical study of the secrecy capacity in the presence of an eavesdropper, this work studies [13] the duality between wiretap channels and state-dependent channels. This represents a basic framework to extend in a near future this study to Full-Duplex scenarios, where the Full-Duplex capability of a node could increase the secrecy of the wireless communication.

6.1.1.2. SDR for SRDs

The technologies employed in urban sensor networks are permanently evolving, and thus the gateways of these networks have to be regularly upgraded. The existing method to do so is to stack-up receivers dedicated to one communication protocol. However, this implies to have to replace the gateway every time a new protocol is added to the network. A more practical way to do this is to perform a digitization of the full band and to perform digitally the signal processing, as done in Software-Defined Radio (SDR). The main hard point in doing this is the dynamic range of the signals: indeed the signals are emitted with very different features because of the various propagation conditions. It has been proved that the difference of power between two signals can be so important that no existing Analog-to-Digital Converter (ADC) is able to properly digitize the signals. We propose a solution to reduce the dynamic range of signals before digital conversion. In this study [9], the assumption is made that there is one strong signal, and several weak signals. This assumption is made from the existing urban sensor networks topology. A receiver architecture with two branches is proposed with a “Coarse Digitization Path” (CDP) and a “Fine Digitization Path” (FDP). The CDP allows to digitize the strong signal and to get data on it that is used to reconfigure the FDP. The FDP then uses a notch filter to attenuate the strong signal (and then to reduce the dynamic range of the signals) and digitizes the rest of the band.

6.2. Multi-User Communications

6.2.1. Fundamental Limits

6.2.1.1. Approximate Capacity Region of the Gaussian Interference Channel with Feedback

An achievability region and a converse region for the two-user Gaussian interference channel with noisy channel-output feedback (G-IC-NOF) are presented [42], [30], [43], [47]. The achievability region is obtained using a random coding argument and three well-known techniques: rate splitting, superposition coding and backward decoding. The converse region is obtained using some of the existing perfect-output feedback outer-bounds as well as a set of new outer-bounds that are obtained by using genie-aided models of the original G-IC-NOF. Finally, it is shown that the achievability region and the converse region approximate the capacity region of the G-IC-NOF to within a constant gap in bits per channel use.

6.2.1.2. Full Characterization of the Capacity Region of the Linear Deterministic Interference Channel with Feedback

The capacity region of the two-user linear deterministic (LD) interference channel with noisy output feedback (IC-NOF) has been fully characterized [29]. This result allows the identification of several asymmetric scenarios in which implementing channel-output feedback in only one of the transmitter-receiver pairs is as beneficial as implementing it in both links, in terms of achievable individual rate and sum-rate improvements w.r.t. the case without feedback. In other scenarios, the use of channel-output feedback in any of the transmitter-receiver pairs benefits only one of the two pairs in terms of achievable individual rate improvements or simply, it turns out to be useless, i.e., the capacity regions with and without feedback turn out to be identical even in the full absence of noise in the feedback links.

6.2.1.3. Full Characterization of the Information Equilibrium Region of the Multiple Access Channel

The fundamental limits of decentralized information transmission in the K-user Gaussian multiple access channel (G-MAC), with $K \geqslant 2$, are fully characterized [38]. Two scenarios are considered. First, a game in which only the transmitters are players is studied. In this game, the transmitters autonomously and independently tune their own transmit configurations seeking to maximize their own information transmission rates, $R_1, R_2, \cdots, R_K$, respectively. On the other hand, the receiver adopts a fixed receive configuration that is known a priori to the transmitters. The main result consists of the full characterization of the set of rate tuples $(R_1, R_2, ..., R_K)$ that are achievable and stable in the G-MAC when stability is considered in the sense of the η-Nash equilibrium (NE), with $\eta > 0$ arbitrarily small. Second, a sequential game in which the two categories of players (the transmitters and the receiver) play in a given order is presented. For this sequential game, the main result consists of the full characterization of the set of rate tuples $(R_1, R_2, ..., R_K)$ that are stable in the sense of an η-sequential equilibrium, with $\eta > 0$.

6.2.1.4. Full Characterization of the Information-Energy Capacity Region of the Multiple Access Channel with Energy Harvester with and without Feedback

The fundamental limits of simultaneous information and energy transmission in the two-user Gaussian multiple access channel (G-MAC) with and without feedback have been fully characterized [10], [15]. More specifically, all the achievable information and energy transmission rates (in bits per channel use and energy-units per channel use, respectively) are identified. In the case without feedback, an achievability scheme based on power-splitting and successive interference cancelation is shown to be optimal. Alternatively, in the case with feedback (G-MAC-F), a simple yet optimal achievability scheme based on power-splitting and Ozarow's capacity achieving scheme is presented. Two of the most important observations in this work are: (a) The information-energy capacity region of the G-MAC without feedback can be a proper subset of the information-energy capacity region of the G-MAC-F and (b) Feedback can at most double the energy rate when the information transmission rate is kept fixed at the sum-capacity of the G-MAC.

6.2.1.5. Full Characterization of the Information-Energy Equilibrium Region of the Multiple Access Channel with Energy Harvester

The fundamental limits of decentralized simultaneous information and energy transmission in the two-user Gaussian multiple access channel (G-MAC) have been fully characterized for the case in which a minimum energy transmission rate b is required for successful decoding [14], [39]. All the achievable and stable information-energy transmission rate triplets (R_1, R_2, B) are identified. R_1 and R_2 are in bits per channel use measured at the receiver and B is in energy units per channel use measured at an energy-harvester (EH). Stability is considered in the sense of an η-Nash equilibrium (NE), with $\eta > 0$ arbitrarily small. The main

result consists of the full characterization of the η-NE information-energy region, i.e., the set of information-energy rate triplets (R_1, R_2, B) that are achievable and stable in the G-MAC when: (a) both transmitters autonomously and independently tune their own transmit configurations seeking to maximize their own information transmission rates, R_1 and R_2 respectively; (b) both transmitters jointly guarantee an energy transmission rate B at the EH, such that $B > b$. Therefore, any rate triplet outside the η-NE region is not stable as there always exists one transmitter able to increase by at least η bits per channel use its own information transmission rate by updating its own transmit configuration.

6.2.1.6. Duality Between State-Dependent Channels and Wiretap Channels

A duality between wiretap and state-dependent channels with non-causal channel state information at the transmitter has been established [13]. First, a common achievable scheme is described for a certain class of state-dependent and wiretap channels. Further, state-dependent and wiretap channels for which this scheme is capacity (resp. secrecy capacity) achieving are identified. These channels are said to be dual. This duality is used to establish the secrecy capacity of certain state-dependent wiretap channels with non-causal channel state information at the transmitter. Interestingly, combatting the eavesdropper or combatting the lack of state information at the receiver turn out to be two non-concurrent tasks.

6.2.1.7. Energy efficiency - Spectral Efficiency (EE-SE) Tradeoffs in Wireless RANs

Even for a point-to-point communication, the Shannon capacity can be interpreted for a Gaussian channel as a fundamental spectral and energy efficiency (SE-EE) trade-off. Extending this fundamental trade-off in the context of multi-user communications is not straightforward as it may depend on many parameters. We proposed in [8] a simple and effective method to study this trade-off in cellular networks, an issue that has attracted significant recent interest in the wireless community. The proposed theoretical framework is based on an optimal radio resource allocation of transmit power and bandwidth for the downlink direction, applicable for an orthogonal cellular network. The analysis is initially focused on a single cell scenario, for which in addition to the solution of the main SE-EE optimization problem, it is proved that a traffic repartition scheme can also be adopted as a way to simplify this approach. By exploiting this interesting result along with properties of stochastic geometry, this work is extended to a more challenging multi-cell environment, where interference is shown to play an essential role and for this reason several interference reduction techniques are investigated. Special attention is also given to the case of low signal to noise ratio (SNR) and a way to evaluate the upper bound of EE in this regime is provided. This methodology leads to tractable analytical results under certain common channel properties, and thus allows the study of various models without the need for demanding system level simulations.

6.2.1.8. Spatial Continuum Channel Models

In the context of the deployment of Internet of Things (see next section for more details about our protocol developments), it is expected that a unique cell could serve millions of radio nodes transmitting sporadic short packets. In [18] and [41], our objective is to study this problem from an information theory point of view to derive the fundamental limit in terms of maximal information rates that can be transmitted in such a dense cell. This work proposes a new model called spatial continuum asymmetric channels to study the channel capacity region of asymmetric scenarios in which either one source transmits to a spatial density of receivers or a density of transmitters transmit to a unique receiver. This approach is built upon the classical broadcast channel (BC) and multiple access channel (MAC). For the sake of consistency, the study is limited to Gaussian channels with power constraints and is restricted to the asymptotic regime (zero-error capacity). The reference scenario comprises one base station in Tx or Rx mode, a spatial random distribution of nodes (resp. in Rx or Tx mode) characterized by a probability spatial density of users $u(x)$ where each of them requests a quantity of information with no delay constraint, thus leading to a requested rate spatial density $\rho(x)$. This system is modeled as an ‚àû‚àíuser asymmetric channel (BC or MAC). To derive the fundamental limits of this model, a spatial discretization is first proposed to obtain an equivalent BC or MAC. Then, a specific sequence of discretized spaces is defined to refine infinitely the approximation. Achievability and capacity results are obtained in the limit of this sequence while the access capacity region $\mathcal{C}(Pm)$ is defined as the set of requested rates spatial densities $\rho(x)$ that are achievable with a transmission power Pm. The uniform capacity defined as the maximal symmetric achievable rate is also computed.

6.2.1.9. Finite Block-Length Coding in Wireless Networks

In the context of IoT, the information to be transmitted will be divided in very small packets especially when control and commands will be transmitted over the network. The classical asymptotic information theory relies on the statistic properties of channels and information sources, when the coding block-length tends to infinity. Therefore this framework is not appropriate to study the fundamental limits of short packets transmission over wireless networks. Fortunately, information theory is not only about the asymptotic regime. Shannon himself derived the preliminary foundations of a theory for finite block-length. Later, Gallager extended this framework. Recently this question gained interest after the work of Y. Polyanskiy which extended former results on finite block length to Gaussian channels. This fundamental contribution opens a way for studying wireless networks under finite block-length regime. But this relatively new paradigm suffers from strong problems relative to the compexity of the underlying estimation problem. Starting to work on this topic in the framework of the associated team with Princeton, we exploited in [35] the recent results on the non-asymptotic coding rate for fading channels with no channel state information at the transmitter and we analyzed the goodput in additive white Gaussian noise (AWGN) and the energy-efficiency spectral-efficiency (EE-SE) tradeoff where the fundamental relationship between the codeword length and the EE is given. Finally, the true outage probability in Ricean and Nakagami-m block fading channels is investigated and it is proved that the asymptotic outage capacity is the Laplace approximation of the average error probability in finite blocklength regime. This preliminary work constitutes one of the starting point for our future works in the framework of the ANR project ARBURST.

6.2.2. Algorithm and Protocol Design for Multi-User Communication Scenarios

6.2.2.1. Interference Management in OFDM/MIMO Wireless Networks

Modern cellular networks in traditional frequency bands are notoriously interference-limited especially in urban areas, where base stations are deployed in close proximity to one another. The latest releases of Long Term Evolution (LTE) incorporate features for coordinating downlink transmissions as an efficient means of managing interference. In [4], we review recent field trial results and theoretical studies of the performance of joint transmission (JT) coordinated multi-point (CoMP) schemes. These schemes revealed, however, that their gains are not as high as initially expected, despite the large coordination overhead. These schemes are known to be very sensitive to defects in synchronization or information exchange between coordinating bases stations as well as uncoordinated interference. In this article, we review recent advanced coordinated beamforming (CB) schemes as alternatives, requiring less overhead than JT CoMP while achieving good performance in realistic conditions. By stipulating that, in certain LTE scenarios of increasing interest, uncoordinated interference constitutes a major factor in the performance of CoMP techniques at large, we hereby assess the resilience of the state-of-the-art CB to uncoordinated interference. We also describe how these techniques can leverage the latest specifications of current cellular networks, and how they may perform when we consider standardized feedback and coordination. This allows us to identify some key roadblocks and research directions to address as LTE evolves towards the future of mobile communications.

Among the different techniques described above, we studied in [32] an interference Alignment (IA) technique that, in a large sense, makes use of the increasing signal dimensions available in the system through MIMO and OFDM technologies in order to globally reduce the interference suffered by users in a network. In this paper, we addressed the problem of downlink cellular networks, the so-called interfering broadcast channels, where mobile users at cell edges may suffer from high interference and thus, poor performance. Starting from the downlink IA scheme proposed by Suh et al., a new approach is proposed where each user feeds back multiple selected received signal directions with high signal-to-interference gain. A exhaustive search based scheduler selects a subset of users to be served simultaneously, balancing between sum-rate performance and fairness, but becomes untractable in dense network scenarios where many users send simultaneous requests. Therefore, we develop a sub-optimal scheduler that greatly decreases the complexity while preserving a near-optimal data rate gain. More interestingly, our simulations show that the IA scheme becomes valuable only in correlated channels, whereas the matched filtering based scheme performs the best in the uncorrelated scenarios.

6.2.2.2. Performance of Ultra-NarrowBand Techniques for Internet of Things

This section makes echo to the section entitled Spatial Continuum Channel Models where fundamental limits are studied for a similar scenario. In this section, we investigate the scenario for an existing PHY layer technology, Ultra Narrow Band (UNB) technique, proposer by Sigfox. The ALOHA protocol is regaining interest in the context of the Internet of Things (IoT), especially for UNB signals (dedicated to long range and low power transmission in IoT networks). In this case, the classical assumption of channelization is not verified anymore, modifying the ALOHA performances. Indeed, UNB signals suffer from a lack of precision on the actual transmission carrier frequency, leading to a behavior similar to a frequency unslotted random access. More precisely, the channel access is Random-FTMA, where nodes select their time and frequency in a random and continuous way. The frequency randomness prevents from allocating orthogonal resources for transmission, and induces uncontrolled interference.

In [19], the success probability and throughput of ALOHA is generalized to further describe frequency-unslotted systems such as UNB. The main contribution of this work is the derivation of a generalized expression of the throughput for the random time-frequency ALOHA systems, when neglecting channel attenuation. Besides, this study permits to highlight the duality of ALOHA in time and frequency domain.

Besides, in [26] and [27], to introduce diversity, we propose the use of replication mechanism to enhance the reliability of UNB wireless network. Considering the outage probability, we theoretically evaluate the system performance and show that there exists an optimal number of transmissions. Finally, we highlight that this number of repetitions can be easily optimized by considering a unique global parameter.

Finally, in [28], we also take into consideration the channel effect for such specific network. Indeed, the UNB randomness leads to a new behavior of the interference which has not been theoretically analyzed yet, when considering the pathloss of nodes located randomly in an area. In this work, in order to quantify the system performance, we derive and exploit a theoretical expression of the packet error rate in a UNB based IoT network, when taking into account both interference due to the spectral randomness and path loss due to the propagation.

6.2.2.3. Algorithms and Protocols for BANs

Body Area Networks (BANs) represent a challenging area of research for networking design. Indeed, the topology of these networks differe significantly from classical networks. BANs are dynamic, multi-scale, energy limited and require real time protocols for many applications related to localization. Our work is related to the design of dynamic protocols to gather and exploit localization information in dynamic BANs. Our first contribution is related to the context of group navigation and was developed in the framework of the FUI SMACS project dealing with the localisation of runners during bike races. The problem is to develop fast and reliable protocols to dynamically gather mobility information from moving nodes toward moving sinks.

Our second contribution is relative to the mobility of a single BAN and with the objective of improving localization algorithms based on ranging measures between nodes spread on the body. This work was done in the framework of the ANR CORMORAN project with the PhD of Arturo Gimenez-Guizar who defended his PhD in October 2016 [1].

6.2.2.3.1. Information Gathering in a Group of Mobile Users

In [16], we propose an efficient approach to collect data in mobile wireless sensor networks, with the specific application of sensing in bike races. Recent sensor technology permits to track GPS position of each bike. Because of the inherent correlation between bike positions in a bike race, a simple GPS log is inefficient. The idea presented in this work is to aggregate GPS data at sensors using compressive sensing techniques. We enforce, in addition to signal sparsity, a spatial prior on biker motion because of the group behaviour (peloton) in bike races. The spatial prior is modeled by a graphical model and the data aggregation problem is solved, with both the sparsity and the spatial prior, by belief propagation. We validate our approach on a bike race simulator using trajectories of motorbikes in a real bike race.

6.2.2.3.2. MAC Protocols and Algorithms for Localization at the Body Scale

In this work [20], we have considered the positioning success rate for localization applications deployed in Wireless Body Area Networks (WBAN). Localization is performed with Ultra Wide Band (UWB) pulses, which permits to estimate distances as defined by 3 Way Ranging protocol (3WR). Two channels are considered : the empirical channel CM3, and with our model obtained from our measurement campaign. We first evaluate the positioning loss when considering an aggregation and broadcast scheduling strategy (A&B) upon TDMA MAC. We highlight the channel effects depending on the targeted receiver sensitivity. We then improve the performances by proposing a cooperative algorithm based on conditional permutation of anchors.

6.2.3. Cyber-Physical Systems

6.2.3.1. Attacks in the Electricity Grids

Multiple attacker data injection attack construction in electricity grids with minimum-mean-square-error state estimation has been studied for centralized and decentralized scenarios [6], [11]. A performance analysis of the trade-off between the maximum distortion that an attack can introduce and the probability of the attack being detected by the network operator is considered. In this setting, optimal centralized attack construction strategies are studied. The decentralized case is examined in a game-theoretic setting. A novel utility function is proposed to model this trade-off and it is shown that the resulting game is a potential game. The existence and cardinality of the corresponding set of Nash Equilibria (NEs) of the game is analyzed. Interestingly, the attackers can exploit the correlation among the state variables to facilitate the attack construction. It is shown that attackers can agree on a data injection vector construction that achieves the best trade-off between distortion and detection probability by sharing only a limited number of bits offlline. For the particular case of two attackers, numerical results based on IEEE test systems are presented.

6.2.3.2. Recovering Missing Data in Electricity Grids

The performance of matrix completion based recovery of missing data in electricity distribution systems has been analyzed [17]. Under the assumption that the state variables follow a multivariate Gaussian distribution the matrix completion recovery is compared to estimation and information theoretic limits. The assumption about the distribution of the state variables is validated by the data shared by Electricity North West Limited. That being the case, the achievable distortion using minimum mean square error (MMSE) estimation is assessed for both random sampling and optimal linear encoding acquisition schemes. Within this setting, the impact of imperfect second order source statistics is numerically evaluated. The fundamental limit of the recovery process is characterized using Rate-Distortion theory to obtain the optimal performance theoretically attainable. Interestingly, numerical results show that matrix completion based recovery outperforms MMSE estimator when the number of available observations is low and access to perfect source statistics is not available.

6.3. Software Radio Programming Model

6.3.1. Dataflow programming model

The advent of portable software-defined radio (SDR) technology is tightly linked to the resolution of a difficult problem: efficient compilation of signal processing applications on embedded computing devices. Modern wireless communication protocols use packet processing rather than infinite stream processing and also introduce dependencies between data value and computation behavior leading to dynamic dataflow behavior. Recently, parametric dataflow has been proposed to support dynamicity while maintaining the high level of analyzability needed for efficient real-life implementations of signal processing computations. The team developed a new compilation flow [5] that is able to compile parametric dataflow graphs. Built on the LLVM compiler infrastructure, the compiler offers an actor-based C++ programming model to describe parametric graphs, a compilation front end for graph analysis, and a back end that currently matches the Magali platform: a prototype heterogeneous MPSoC dedicated to LTE-Advanced. We also introduce an innovative scheduling technique, called microscheduling, allowing one to adapt the mapping of parametric dataflow programs to the specificities of the different possible MPSoCs targeted. A specific focus on FIFO sizing

on the target architecture is presented. The experimental results show compilation of 3GPP LTE-Advanced demodulation on Magali with tight memory size constraints. The compiled programs achieve performance similar to handwritten code.

The memory subsystem of modern multi-core architectures is becoming more and more complex with the increasing number of cores integrated in a single computer system. This complexity leads to profiling needs to let software developers understand how programs use the memory subsystem. Modern processors come with hardware profiling features to help building tools for these profiling needs. Regarding memory profiling, many processors provide means to monitor memory traffic and to sample read and write memory accesses. Unfortunately, these hardware profiling mechanisms are often very complex to use and are specific to each micro-architecture. The numap library [44], [31] is dedicated to the profiling of the memory subsytem of modern multi-core architectures. numap is portable across many micro-architectures and comes with a clean application programming interface allowing to easily build profiling tools on top of it.

This numap library as been officially integrated into Turnus, a profiler dedicated to dynamic dataflow programs.

6.3.2. Implementation of filters and FFTs on FPGAs

In collaboration with two researchers from Inria AriC, we have worked on a digital filter synthesis flow targeting FPGAs [46]. Based on a novel approach to the filter coefficient quantization problem, this approach produces results which are faithful to a high-level frequency-domain specification. An automated design process is also proposed where user intervention is limited to a very small number of relevant input parameters. Computing the optimal value of the other parameters not only simplifies the user interface: the resulting architectures also outperform those generated by mainstream tools in accuracy, performance, and resource consumption.

In collaboration with researchers from Isfahan, Iran, a multi-precision Fast Fourier Transform (FFT) module with dynamic run-time reconfigurability has been proposed [3] to trade off accuracy with energy efficiency in an SDR-based architecture. To support variable-size FFT, a reconfigurable memory-based architecture is investigated. It is revealed that the radix-4 FFT has the minimum computational complexity in this architecture. Regarding implementation constraints such as fixed-width memory, a noise model is exploited to statistically analyze the proposed architecture. The required FFT word-lengths for different criteria, (bit-error rate (BER), modulation scheme, FFT size, and SNR) are computed analytically and confirmed by simulations in AWGN and Rayleigh fading channels. At run-time, the most energy-efficient word-length is chosen and the FFT is reconfigured until the required application-specific BER is met. Evaluations show that the implementation area and the number of memory accesses are reduced. The results obtained from synthesizing basic operators of the proposed design on an FPGA show energy consumption saving of over 80 %.

6.3.3. Tools for FPGA development

The pipeline infrastructure of the FloPoCo arithmetic core generator has been completely overhauled [34], [23]. 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. These allow pipeline-dependent changes to the circuit graph for finer optimization. This is particularly important for the filter structures already mentioned [46].

In parallel, we also started to study the integration of arithmetic optimizations in high-level synthesis (HLS) tools [48]. HLS is a big step forward in terms of design productivity. However, it restricts data-types and operators to those available in the `C` language supported by the compiler, preventing a designer to fully exploit the FPGA flexibility. To lift this restriction, a source-to-source compiler may rewrite, inside critical loop nests of the input `C` code, selected floating-point additions into sequences of simpler operator using non-standard

arithmetic formats. This enables hoisting floating-point management out the loop. What remains inside the loop is a sequence of fixed-point additions whose size is computed to enforce a user-specified, application-specific accuracy constraint on the result. Evaluation of this method demonstrates significant improvements in the speed/resource usage/accuracy trade-off.

6.3.4. Computer Arithmetic

In collaboration with researchers from Istanbul, Turkey, operators have also been developed for division by a small positive constant [49]. 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.

On the software side, we have also shown, in collaboration with researchers from LIP and the Kalray company, that correctly rounded elementary functions can be implemented more efficiently using only fixed-point arithmetic than when classically using floating-point arithmetic [24]. A purely integer implementation of the correctly rounded double-precision logarithm outperforms the previous state of the art, with the worst-case execution time reduced by a factor 5. This work also introduces variants of the logarithm that input a floating-point number and output the result in fixed-point. These are shown to be both more accurate and more efficient than the traditional floating-point functions for some applications.

URBANET Team

7. New Results

7.1. Network deployment and characterization

Participants: Ahmed Boubrima, Angelo Furno, Walid Bechkit, Khaled Boussetta, Hervé Rivano, Razvan Stanica.

7.1.1. Deployment of Wireless Sensor Networks for Pollution Monitoring

Monitoring air quality has become a major challenge of modern cities, where the majority of population lives, 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. Wireless sensor networks seem to be a good solution to this problem, thanks to sensors' low cost and autonomy, as well as their fine-grained deployment. A careful deployment of sensors is therefore necessary to get better performances, while ensuring a minimal financial cost.

We have tackled the issue of WSN deployment for air pollution monitoring in a series of papers this year. In [10], we tackled the optimization problem of sensor deployment and we proposed an integer programming model, which allows to find the optimal network topology while ensuring air quality monitoring with a high precision and the minimum financial cost. Most of existing deployment models of wireless sensor networks are generic and assume that sensors have a given detection range. This assumption does not fit pollutant concentrations sensing. Our model takes into account interpolation methods to place sensors in such a way that pollution concentration is estimated with a bounded error at locations where no sensor is deployed. This solution was further tested and evaluated on a data set of the Lyon city [9], giving insights on how to establish a good compromise between the deployment budget and the precision of air quality monitoring.

In practice, multiple pollution sources can be present in an area. For this reason, in [11] we propose to apply a spatial clustering algorithm to the air pollution data in order to determine pollution zones that are due to the same pollutant sources and group them together to find candidate sites for the deployment of sensors. This approach was tested on real world data, namely the Paris pollution data, which was recorded in March 2014.

A very important deployment parameter is the height at which the sensor is placed. In [12], we demonstrate the impact of this parameter, usually neglected in the literature. This pushed us to study a 3D deployment model, based on an air pollution dispersion model issued from real experiments, performed in wind tunnels emulating the pollution emitted by a steady state traffic flow in a typical street canyon.

7.1.2. Access Point Deployment

The problem of designing wireless local networks (WLANs) involves deciding where to install the access points (APs), and assigning frequency channels to them with the aim to cover the service area and to guarantee enough capacity to users. In [5], we propose different solutions to the problems related to the WLAN design. In the first part, we focus on the problem of designing a WLAN by treating separately the AP positioning and the channel assignment problems. For the AP positioning issue, we formulate it as a set covering problem. Since the computation complexity limits the exact solution, we propose two heuristics to offer efficient solutions. On the other hand, for the channel assignment, we define this issue as a minimum interference frequency assignment problem and propose three heuristics: two of them aim to minimize the interference at AP locations, and the third one minimizes the interference at the TPs level. In the second part, we treat jointly the two aforementioned issues based on the concept of virtual forces. In this case, we start from an initial solution provided by the separated approach and try to enhance it by adjusting the APs positions and reassigning their operating frequencies.

7.1.3. Mobile Traffic Analysis

The analysis of operator-side mobile traffic data is a recently emerged research field, and, apart a few outliers, relevant works cover the period from 2005 to date, with a sensible densification over the last four years. In [8], we provided a thorough review of the multidisciplinary activities that rely on mobile traffic datasets, identifying major categories and sub-categories in the literature, so as to outline a hierarchical classification of research lines and proposing a complete introductory guide to the research based on mobile traffic analysis.

The usage of these datasets in the design of new networking solutions, in order to achieve the so-called cognitive networking paradigm, is one of the most important applications of these analytics methods. In fact, 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 and realize a more dynamic management of network resources, that adapts to the significant spatiotemporal fluctuations of the mobile demand. In [6], 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.

7.2. Data Collection in Multi-hop Networks

Participants: Jin Cui, Jad Oueis, Hervé Rivano, Razvan Stanica, Fabrice Valois.

7.2.1. Data Aggregation in Wireless Sensor Networks

Wireless Sensor Networks (WSNs) have been regarded as an emerging and promising field in both academia and industry. Currently, such networks are deployed due to their unique properties, such as self-organization and ease of deployment. However, there are still some technical challenges needed to be addressed, such as energy and network capacity constraints. Data aggregation, as a fundamental solution, processes information at sensor level as a useful digest, and only transmits the digest to the sink. The energy and capacity consumptions are reduced due to less data packets transmission.

As a key category of data aggregation, aggregation function, solving how to aggregate information at sensor level, was investigated in the Ph.D. thesis of Jin Cui [1]. In this work, we make four main contributions: firstly, we propose two new networking-oriented metrics to evaluate the performance of aggregation function: aggregation ratio and packet size coefficient. Aggregation ratio is used to measure the energy saving by data aggregation, and packet size coefficient allows to evaluate the network capacity change due to data aggregation. Using these metrics, we confirm that data aggregation saves energy and capacity whatever the routing or MAC protocol is used. Secondly, to reduce the impact of sensitive raw data, we propose a data-independent aggregation method which benefits from similar data evolution and achieves better recovered fidelity. This solution, named Simba, is detailed in [15] as well. Thirdly, a property-independent aggregation function is proposed to adapt the dynamic data variations. Comparing to other functions, our proposal can fit the latest raw data better and achieve real adaptability without assumption about the application and the network topology. Finally, considering a given application, a target accuracy, we classify the forecasting aggregation functions by their performance. The networking-oriented metrics are used to measure the function performance, and a Markov Decision Process is used to compute them. Dataset characterization and classification framework are also presented to guide researcher and engineer to select an appropriate functions under specific requirements.

7.2.2. Energy Harvesting in Wireless Sensor Networks

Energy harvesting capabilities are challenging our understanding of wireless sensor networks by adding recharging capacity to sensor nodes. This has a significant impact on the communication paradigm, as networking mechanisms can benefit from these potentially infinite renewable energy sources. In [23], we study photovoltaic energy harvesting in wireless sensor networks, by building a harvesting analytical model, linking three components: the environment, the battery, and the application. Given information on two of

the components, limits on the third one can be determined. To test this model, we adopt several use cases with various indoor and outdoor locations, battery types, and application requirements. Results show that, for predefined application parameters, we are able to determine the acceptable node duty cycle given a specific battery, and vice versa. Moreover, the suitability of the deployment environment (outdoor, well lighted indoor, poorly lighted indoor) for different application characteristics and battery types is discussed .

In a second contribution [22], we study the consequences of implementing photovoltaic energy harvesting on the duty cycle of a wireless sensor node, in both outdoor and indoor scenarios. We show that, for the static duty cycle approach in outdoor scenarios, very high duty cycles, in the order of tens of percents, are achieved. This further eliminates the need for additional energy conservation schemes. In the indoor case, our analysis shows that the dynamic duty cycle approach based solely on the battery residual energy does not necessarily achieve better results than the static approach. We identify the main reasons behind this behavior, and test new design considerations by adding information on the battery level variation to the duty cycle computation. We demonstrate that this approach always outperforms static solutions when perfect knowledge of the harvestable energy is assumed, as well as in realistic deployments, where this information is not available.

7.2.3. Data Collection with Moving Nodes

Patrolling with mobile nodes (robots, drones, cars) is mainly used in situations where the need of repeatedly visiting certain places is critical. In [24], we consider a deployment of a wireless sensor network (WSN) that cannot be fully meshed because of the distance or obstacles. Several robots are then in charge of getting close enough to the nodes in order to connect to them, and perform a patrol to collect all the data in time. We discuss the problem of multi-robot patrolling within the constrained wireless networking settings. We show that this is fundamentally a problem of vertex coverage with bounded simple cycles (CBSC). We offer a formalization of the CBSC problem and prove it is NP-hard and at least as hard as the Traveling Salesman Problem (TSP). Then, we provide and analyze heuristics relying on clusterings and geometric techniques. The performances of our solutions are assessed in regards to robot limitations (storage and energy), networking parameters, but also to random and particular graph models.

Also related to data collection, in [3], we advocate the use of conventional vehicles equipped with storage devices as data carriers whilst being driven for daily routine journeys. The road network can be turned into a large-capacity transmission system to offload bulk transfers of delay-tolerant data from the Internet. The challenges we address include how to assign data to flows of vehicles and while coping with the complexity of the road network. We propose an embedding algorithm that computes an offloading overlay where each logical link spans over multiple stretches of road from the underlying road infrastructure. We then formulate the data transfer assignment problem as a novel linear programming model we solve to determine the optimal logical paths matching the performance requirements of a data transfer. We evaluate our road traffic allocation scheme using actual road traffic counts in France. The numerical results show that 20% of vehicles in circulation in France equipped with only one Terabyte of storage can offload Petabyte transfers in a week.

7.2.4. Network Resilience

The notion of Shared Risk Link Groups (SRLG) captures survivability issues when a set of links of a network may fail simultaneously. The theory of survivable network design relies on basic combinatorial objects that are rather easy to compute in the classical graph models: shortest paths, minimum cuts, or pairs of disjoint paths. In the SRLG context, the optimization criterion for these objects is no longer the number of edges they use, but the number of SRLGs involved. Unfortunately, computing these combinatorial objects is NP-hard and hard to approximate with this objective in general. Nevertheless some objects can be computed in polynomial time when the SRLGs satisfy certain structural properties of locality which correspond to practical ones, namely the star property (all links affected by a given SRLG are incident to a unique node) and the span 1 property (the links affected by a given SRLG form a connected component of the network). The star property is defined in a multi-colored model where a link can be affected by several SRLGs while the span property is defined only in a mono-colored model where a link can be affected by at most one SRLG. In [4], we extend these notions to characterize new cases in which these optimization problems can be solved in polynomial time. We also

investigate the computational impact of the transformation from the multi-colored model to the mono-colored one. Experimental results are presented to validate the proposed algorithms and principles.

7.3. Networks in the Internet of Things

Participants: Soukaina Cherkaoui, Alexis Duque, Guillaume Gaillard, Hervé Rivano, Razvan Stanica, Fabrice Valois.

7.3.1. Service Level Agreements in the Internet of Things

With the growing use of distributed wireless technologies for modern services, the deployments of dedicated radio infrastructures do not enable to ensure large-scale, low-cost and reliable communications. The Ph.D. thesis of Guillaume Gaillard [2] aims at enabling an operator to deploy a radio network infrastructure for several client applications, hence forming the Internet of Things (IoT). We evaluate the benefits earned by sharing an architecture among different traffic flows, in order to reduce the costs of deployment, obtaining a wide coverage through efficient use of the capacity on the network nodes. We thus need to ensure a differentiated Quality of Service (QoS) for the flows of each application.

We propose to specify QoS contracts, namely Service Level Agreements (SLAs), in the context of the IoT. SLAs include specific Key Performance Indicators (KPIs), such as the transit time and the delivery ratio, concerning connected devices that are geographically distributed in the environment. The operator agrees with each client on the sources and amount of traffic for which the performance is guaranteed. Secondly, we describe the features needed to implement SLAs on the operated network, and we organize them into an SLA management architecture. We consider the admission of new flows, the analysis of current performance and the configuration of the operator's relays. Based on a robust, multi-hop technology, IEEE Std 802.15.4-2015 TSCH mode, we provide two essential elements to implement the SLAs : a mechanism for the monitoring of the KPIs [19], and KAUSA, a resource allocation algorithm with multi-flow QoS constraints [18]. The former uses existing data frames as a transport medium to reduce the overhead in terms of communication resources. We compare different piggybacking strategies to find a tradeoff between the performance and the efficiency of the monitoring. With the latter, KAUSA, we dedicate adjusted time-frequency resources for each message, hop by hop. KAUSA takes into account the interference, the reliability of radio links and the expected load to improve the distribution of allocated resources and prolong the network lifetime [17]. We show the gains and the validity of our contributions with a simulation based on realistic traffic scenarios and requirements.

7.3.2. Channel Access in Machine-to-Machine Communications

The densification of the urban population and the rise of smart cities applications foster the need for capillary networks collecting data from sensors monitoring the cities. Among the multiple networking technologies considered for this task, cellular networks, such as LTE-A, bring an ubiquitous coverage of most cities. It is therefore necessary to understand how to adapt LTE-A, and what should be the future 5G architecture, in order to provide efficient connectivity to Machine-to-Machine (M2M) devices alongside the main target of mobile networks, Human-to-Human devices. Indeed, cellular random access procedures are known to suffer from congestion in presence of a large number of devices, while smart cities scenarios expect huge density of M2M devices. Several solutions have been investigated for the enhancement of the current LTE-A access management strategy. In [14], we contribute to the modeling and computation of the capacity of the LTE-A Random Access Channel (RACH) in terms of simultaneous successful access. In particular, we investigate the hypothesis of piggybacking the payload of Machine Type Communications from M2M devices within the RACH, and show that M2M densities considered realistic for smart cities applications are difficult to sustain by the current LTE-A architecture.

7.3.3. Visible Light Communications in the Internet of Things

The Internet of Things connects devices, such as everyday consumer objects, enabling information gathering and improved user experience. Also, this growing and dynamic market makes that consumers nowadays expect electronic products, even the cheapest, to include wireless connectivity. However, despite the fact that radio based solutions exist, such as Bluetooth Low Energy, the manufacturing costs introduced by these radio

technologies are non-negligible compared to the initial product price. As most of the home electronics already integrate small light emitting diodes, Visible Light Communication appears as a competitive alternative. However, its broad adoption is suffering from a lack of integration with smartphones, which represent the communication hubs for most of the users. To overcome this issue, in [16], we propose a line of sight LED-to-camera communication system based on a small color LED and a smartphone. We design a cheap prototype as proof of concept of a near communication framework for the Internet of Things. We evaluate the system performance, its reliability and the environment influence on the LED-to-camera communication, highlighting that a throughput of a few kilobits per second is reachable. Finally, we design a real time, efficient LED detection and image processing algorithm to leverage the specific issues encountered in the system.

7.3.4. Radio Frequency Identification in Dense Environments

Radio Frequency Identification (RFID) is another cheap technology shaping the Internet of Things. The rapid development of RFID has allowed its large adoption and led to increasing deployments of RFID solutions in diverse environments under varying scenarios and constraints. The nature of these constraints ranges from the amount to the mobility of the readers deployed, which in turn highly affects the quality of the RFID system, causing reading collisions. However, the technology suffers from a recurring issue: the reader-to-reader collisions. Numerous protocols have been proposed to attempt to reduce them, but remaining reading errors still heavily impact the performance and fairness of dense RFID deployments.

In order to ensure collision-free reading, a scheduling scheme is needed to read tags in the shortest possible time. In [25], we study this scheduling problem in a stationary setting and the reader minimization problem in a mobile setting. We show that the optimal schedule construction problem is NP-complete and provide an approximation algorithm that we evaluate our techniques through simulation. Moving closer to practical solutions, [20] introduces a new Distributed Efficient & Fair Anticollision for RFID (DEFAR) protocol. DEFAR reduces both monochannel and multichannel collisions, as well as interference, by a factor of almost 90% in comparison with the best state of the art protocols. The fairness of the medium access among the readers is improved to a 99% level. Such improvements are achieved by applying a TDMA-based "server-less" approach and assigning different priorities to readers depending on their behavior over precedent rounds. A distributed reservation phase is organized between readers with at least one winning reader afterwards. Then, multiple reading phases occur within a single frame in order to obtain fast coverage and high throughput. The use of different reader priorities based on reading behaviors of previous frames also contributes to improve both fairness and efficiency.

Another type of collisions appears when the RFID tags are not only dense, but also mobile. mDEFAR [21] is an adaptation of DEFAR, while CORA [7] 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 the state of the art, 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.