RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2019

Section New Results

Edition: 2020-03-21

DISTRIBUTED SYSTEMS AND MIDDLEWARE

DISTRIBUTED AND HIGH PERFORMANCE COMPUTING

DISTRIBUTED PROGRAMMING AND SOFTWARE ENGINEERING

NETWORKS AND TELECOMMUNICATIONS

COAST Project-Team

7. New Results

7.1. Trustworthy Collaboration

Participants: Claudia-Lavinia Ignat, Hoang Long Nguyen, Olivier Perrin.

In order to test user acceptance of a collaboration model based on automatic trust assessment, we designed an experiment relying on the trust game. In the trust game money exchange is entirely attributable to the existence of trust between users. Our experimental design [7] tested variations of the trust game: with and without showing the partner identity and with and without explicit computation of partner trust values based on the computational trust model we previously proposed. We organized a user study with 30 participants that confirmed that the availability of this trust metric improves user cooperation and that it predicts participants future behavior. We showed that trust score availability has the same effect as an identity to improve cooperation. Our study suggests that trust score could function as an enhancement or even replacement of traditional identity systems and has the advantage of scalability.

In the scope of Hoang Long Nguyen's PhD thesis, we proposed the architecture of ÔBlock, an open ecosystem for quick development of transparent applications based on consortium blockchain.

7.2. Undo in Collaborative Editing

Participants: Victorien Elvinger, Claudia-Lavinia Ignat.

In collaborative editors a selective undo allows a user to undo an earlier operation, regardless of when, where and by which user the operation was generated. In most existing collaborative editors such as GoogleDrive, selective undo is not integrated and users can only undo their own operations but not the ones generated by the other users. There is currently no generally applicable undo support as stated in the manifesto on CRDTs [17]. We presented a generic support of selective undo for CRDTs by proposing an abstraction that captures the semantics of concurrent undo and redo operations through equivalence classes. The abstraction is a natural extension of undo and redo in sequential applications and is straightforward to implement in practice [9].

7.3. Mitigating the Cost of Identifiers in Sequence CRDT

Participants: Matthieu Nicolas, Gérald Oster, Olivier Perrin.

To achieve high availability, large-scale distributed systems have to replicate data and to minimise coordination between nodes. The literature and industry increasingly adopt Conflict-free Replicated Data Types (CRDTs) to design such systems. CRDTs are data types which behave as traditional ones, e.g. the Set or the Sequence. However, compared to traditional data types, they are designed to support natively concurrent modifications. To this end, they embed in their specification a conflict-resolution mechanism.

To resolve conflicts in a deterministic manner, CRDTs usually attach identifiers to elements stored in the data structure. Identifiers have to comply with several constraints such as uniqueness or being densely ordered according to the kind of CRDT. These constraints may prevent the identifiers' size from being bounded. As the number of the updates increases, the size of identifiers grows. This leads to performance issues, since the efficiency of the replicated data structure decreases over time.

To address this issue, we propose a new CRDT for Sequence which embeds a renaming mechanism. It enables nodes to reassign shorter identifiers to elements in an uncoordinated manner. Obtained experiment results demonstrate that this mechanism decreases the overhead of the replicated data structure and eventually limits it.

To validate the proposed renaming mechanism, we performed an experimental evaluation to measure its performances on several aspects: (i) the size of the data structure ; (ii) the integration time of the rename operation ; (iii) the integration time of insert and remove operations. In cases (i) and (iii), we use LogootSplit as the baseline data structure to compare results. The results we obtained are very encouraging, as the integration time is far shorter with the renaming mechanism, even with the time spent to apply the rename operation.

7.4. Social Networks as Collaboration Support

Participants: Quentin Laporte Chabasse, Gérald Oster, François Charoy.

Safe peer to peer collaborative services requires a trusted peer to peer network in order to be effective. We started to investigate how to leverage social networks underlying inter organizational collaboration to support such collaboration. To reach this goal, we need to analyze collaborative graphs. They are a relevant sources of information to understand behavioural tendencies of groups of individuals. Exponential Random Graph Models (ERGMs) are commonly used to analyze such social processes including dependencies between members of the group. Our approach considers a modified version of ERGMs, modeling the problem as an edge labelling one. The main difficulty is inference since the normalizing constant involved in classical Markov Chain Monte Carlo approaches is not available in an analytic closed form.

The main contribution is to use the recent ABC Shadow algorithm [20]. This algorithm is built to sample from posterior distributions while avoiding the previously mentioned drawback. The proposed method is illustrated on real data sets provided by the HAL [0] platform and provides new insights on self-organized collaborations among researchers[11]

7.5. Secure Collaborative Editing

Participants: Mohammed Riyadh Abdmeziem, François Charoy.

Collaborative edition allows a group of entities to simultaneously edit and share the content of a document in real time. To provide the required keying materials, group key management protocols are usually considered in order to secure and encrypt the exchanged data. Indeed, existing fully distributed protocols induce significant overhead. Instead, centralized solutions are preferred for their high efficiency. Nevertheless, these centralized solutions present two main issues. The first issue is related to the broken end-to-end property, considering the central entity has access to the established credentials. The second issue is related to the single point of failure problem. In fact, if the central entity fails, the key establishment process fails too. To address these challenges, we proposed a simple, and yet efficient approach which enhances central-based protocols with both fault tolerance and end-to-end properties. To do so, we considered the group key as composed of two sub-keys. The first sub-key is only known to the members of the group, excluding the central entity, while the second sub-key is distributed and updated by the central entity following membership changes[3], [4]. Our initial assessment shows that the overall complexity of rekeying operations is not negatively impacted. In addition, our approach is backward compatible with existing solutions in the literature.

7.6. Trust and Data Sharing in Crisis Management

Participants: François Charoy, Béatrice Linot.

Sharing information between responders is important during crisis management response. Tools and platforms are eagerly developed for that purpose. They are supposed to support people and help them to build a shared situation awareness. However as the scale of crisis increases and as more and more organizations are involved, people get reluctant to use them to share their data. They prefer to rely on one to one communication tools like phones or text. This is why we are studying how these collaborative platforms impact the work of responders positively or negatively. We want to know why most of the time they don't want to use them for their original purpose. We studied reports on past incidents and conducted extensive analysis of the use of existing systems (e.g. the French platform CRISORSEC) through interviews, observation and data analysis. Early results show that participant have problems sharing written information for different kind of reason including its persistence, the time taken to produce the message and the lack of knowledge regarding who may access this information. This informs us on the requirement for future collaborative platforms.

[0] https://hal.inria.fr/

7.7. Identification and Selection of Services from Cloud Providers

Participants: Anis Ahmed Nacer, François Charoy, Olivier Perrin.

We continued our work on providing a framework to compare plans for services from cloud providers in order to help architects to select the best composition given the required criteria (both both functional and non-functional requirements) for an application. This year, we have made progress in two directions: the first is how to identify the key elements to be considered when architects want to compare the different plans, and the second one is a methodology to compute the best composition, given partial information provided in service description (based on the WOWA method).

In order to gather the key elements of the comparison that met the architects' requirements and the relationship between these key elements of the comparison, we reviewed the service providers' plans and previous works on benchmarks. Finally, to ensure that the list of key elements of the comparison and their relationship was complete for the service selection process, we conducted an empirical study with the architects.

Regarding the second part, we use the WOWA (Weighted Ordered Weighted Averaging) operator to solve this decision problem. This operator provides an aggregation function that uses both the simultaneous advantage of the OWA method to allow compensation between high and low values and the weighted average method to consider the importance of the suppliers who provide the information. WOWA uses two sets of weights: one corresponds to source significance, and the other corresponds to value significance.

Our evaluations are encouraging, and we are now ready to submit our proposals to conferences.

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

Participants: Amina Ahmed Nacer, Claude Godart, Guillaume Rosinosky, Samir Youcef.

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

The contribution is threefold and consists in a method, a model and a framework. The method categorizes techniques to transform an existing business process into a risk-aware process model that takes into account security risks related to cloud environments. These techniques are partially described in the form of constraints to automatically support process transformation. The model formalizes the relations and the responsibilities between the different actors of the cloud. This allows users to identify the different pieces of 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, the paper [5] synthesizes our trust-aware deployment method.

7.9. Priority based events management in IoT-BPM architecture

Participants: Khalid Benali, Abir Ismaili-Alaoui.

BPM allows organizations to evolve their performance and achieve their goals, as it helps them to have a clear vision of their business. Several research works have been done in this area and aimed at improving business processes, by focusing on the optimization of business processes issues at build-time and at run-time, from different perspectives: control-flow perspective, data and event data perspective, and scheduling and event management perspective. Business process instances scheduling and event management are considered as a crucial step in the journey of business process improvement. However, this step becomes more challenging especially when the events are triggered by IoT devices. The main objective of our research consists on scheduling business process instances based on the priority of events that trigger these instances, taking into consideration historical data gathered from previous business process instances. We proposed a clustering approach based on the K-Means algorithm that we apply on a set of event sources, as to classify these sources on different clusters using a score calculated for each event source. This score is based on the frequency and the critically of previous events. The main objective of this approach was to create clusters of priorities. These clusters are used to estimate the criticality level of incoming events, and then the priority level of incoming process instances. However, there is always a degree of uncertainty regarding the criticality/priority level of events generated from sources that belong to the same cluster. This issue can be addressed by using fuzzy logic. In fact, the integration of a Fuzzy Inference System (FIS) in our IoT-BPM architecture, helps us to handle uncertainties regarding the criticality level of events, especially when these events are generated by sources that may have the same characteristics [8].

CTRL-A Project-Team

7. New Results

7.1. Programming support for Autonomic Computing

7.1.1. Reactive languages

Participants: Gwenaël Delaval, Lucie Muller, Eric Rutten.

Our work in reactive programming for autonomic computing systems is focused on the specification and compilation of declarative control objectives, under the form of contracts, enforced upon classical mode automata as defined in synchronous languages. The compilation involves a phase of Discrete Controller Synthesis, integrating the tool ReaX, in order to obtain an imperative executable code. The programming language Heptagon / BZR (see Section Software and Platforms) integrates our research results [5].

An ongoing topic is on abstraction methods for compilation using discrete controller synthesis (needed for example, in order to program the controllers for systems where the useful data for control can be of arbitrary types (integer, real, ...), or also for systems which are naturally distributed, and require a decentralized controller).

Recent work concerns compilation and diagnosis for discrete controller synthesis. The compilation involving a phase of controller synthesis can fail to find a solution, if the problem is overconstrained. The compiler does notify so to the programmer, but the latter would need a diagnosis in order to understand where and how to debug the program. Such diagnosis is made especially difficult by the declarative nature of the synthesis.

This was the object of the M1 TER internship of Lucie Muller [19].

7.1.2. Domain-specific languages

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

Our work in Domain-specific languages (DSLs) is founded on our work in component-based programming for autonomic computing systems as examplified by e.g., FRACTAL. We consider essentially the problem of specifying the control of components assembly reconfiguration, with an approach based on the integration within such a component-based framework of a reactive language as in Section 7.1.1 [4]. In recent work, we proposed an extension of a classical Software Architecture Description Languages (ADL) with Ctrl-F, DSL for the specification of dynamic reconfiguration behavior in a [1]. Based on this experience, we also proposed a DSL called Ctrl-DPR [6], allowing designers to easily generate Autonomic Managers for DPR FPGA systems (see Section 7.2.3).

Ongoing work involves a generalization from our past experiences in software components, DPR FPGA, as well as IoT [8], and Cyberphysical Systems. As we observed a similarity in objects and structures (e.g., tasks, implementation versions, resources, and upper-level application layer), we are considering a more general DSL, which could be specialized towards such different target domains, and where the compilation towards reactive models could be studied and improved, especially considering the features of Section 7.1.1 . This direction will also lead us to study the definition of software architecture patterns for multiple loops Autonomic Managers, particularly hierarchical, with lower layers autonomy alleviating management burden from the upper layers as in Section 7.2 .

7.2. Design methods for reconfiguration controller design in computing systems

We apply the results of the previous axes of the team's activity, as well as other control techniques, 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 [7], [9].

7.2.1. Self-adaptative distributed systems

Participants: Quang Pham Tran Anh, Eric Rutten, Hamza Sahli.

Complex Autonomic Computing Systems, as found typically in distributed systems, must involve multiple management loops, addressing different subproblems of the general management, and using different modeling, decision and control approaches (discrete [3], continuous, stochastic, machine-learning based, ...) They are generally addressing deployment and allocation of computations on resources w.r.t. QoS, load, faults, ... but following different, complementary approaches. The similarities and recurring patterns are considered as in Section 7.1.2 . Their execution needs to be distributed w.r.t. different characteristics such as latency (as in Fog and Edge Computing) or load. We are studying Software Architectures to address the design of such complex systems.

7.2.1.1. Self-adaptation of micro-services in Fog/Edge and Cloud computing

Fog systems are a recent trend of distributed computing having vastly ubiquitous architectures and distinct requirements making their design difficult and complex. Fog computing is based on leveraging both resource-scarce computing nodes around the Edge to perform latency and delay sensitive tasks and Cloud servers for the more intensive computation.

In this work, we present a formal model defining spatial and structural aspects of Fog-based systems using Bigraphical Reactive Systems, a fully graphical process algebraic formalism. The model is extended with reaction rules to represent the dynamic behavior of Fog systems in terms of self-adaptation. The notion of bigraph patterns is used in conjunction with boolean and temporal operators to encode spatio-temporal properties inherent to Fog systems and applications. The feasibility of the modelling approach is demonstrated via a motivating case study and various self-adaptation scenarios.

This work is done in cooperation with the Inria team Stack in Nantes, and published in the FOCLASA workshop, co-located with the SFEM conference [13].

7.2.1.2. Autonomic management in Software Defined Networks

In the framework of our cooperation with Nokia Bell-labs (See Section 8.1.2), and the Dyonisos team at Inria Rennes, we are considering the management of Software Defined Networks (SDN), involving Data-Centers and accelerators.

The main approach AI / Machine Learning approaches, developed in Rennes. An ongoing topic is to consider that these reinforcement learning based approaches involve questions of trust, and we are beginning to consider their composition with controllers based e.g. on Control Theory, in order to maintain guarantees on the behaviors of the managed system.

7.2.2. High-Performance Grid Computing

Cloud and HPC (High-Performance Computing) systems have increasingly become more varying in their behavior, in particular in aspects such as performance and power consumption, and the fact that they are becoming less predictable demands more runtime management [10].

7.2.2.1. A Control-Theory based approach to minimize cluster underuse

Participants: Abdul Hafeez Ali, Raphaël Bleuse, Bogdan Robu, Eric Rutten.

One such problem is found in the context of CiGri, a simple, lightweight, scalable and fault tolerant grid system which exploits the unused resources of a set of computing clusters. In this work, we consider autonomic administration in HPC systems for scientific workflows management through a control theoretical approach. We propose a model described by parameters related to the key aspects of the infrastructure thus achieving a deterministic dynamical representation that covers the diverse and time-varying behaviors of the real computing system. We propose a model-predictive control loop to achieve two different objectives: maximize cluster utilization by best-effort jobs and control the file server's load in the presence of external disturbances. The accuracy of the prediction relies on a parameter estimation scheme based on the EKF (Extended Kalman Filter) to adjust the predictive-model to the real system, making the approach adaptive to parametric variations in the infrastructure. The closed loop strategy shows performance improvement and consequently a reduction in the total computation time. The problem is addressed in a general way, to allow the implementation on similar HPC platforms, as well as scalability to different infrastructures.

This work is done in cooperation with the Datamove team of Inria/LIG, and Gipsa-lab. Some results were published in the CCTA conference [14]. It was the topic of the Master's thesis of Abdul Hafeez Ali [16].

7.2.2.2. *Combining Scheduling and Autonomic Computing for Parallel Computing Resource Management*

Participants: Raphaël Bleuse, Eric Rutten.

This research topic aims at studying the relationships between scheduling and autonomic computing techniques to manage resources for parallel computing platforms. The performance of such platforms has greatly improved (149 petaflops as of November 2019 [20]) at the cost of a greater complexity: the platforms now contain several millions of computing units. While these computation units are diverse, one has to consider other constraints such as the amount of free memory, the available bandwidth, or the energetic envelope. The variety of resources to manage builds complexity up on its own. For example, the performance of the platforms depends on the sequencing of the operations, the structure (or lack thereof) of the processed data, or the combination of application running simultaneously.

Scheduling techniques offer great tools to study/guaranty performances of the platforms, but they often rely on complex modeling of the platforms. They furthermore face scaling difficulties to match the complexity of new platforms. Autonomic computing manages the platform during runtime (on-line) in order to respond to the variability. This approach is structured around the concept of feedback loops.

The scheduling community has studied techniques relying on autonomic notions, but it has failed to link the notions up. We are starting to address this topic.

7.2.3. *High-Performance Embedded Computing*

Participants: Soguy Mak Kare Gueye, Stéphane Mocanu, Eric Rutten.

This topics build upon our experience in reconfiguration control in DPR FPGA [2].

Implementing self-adaptive embedded systems, such as UAV drones, involves an offline provisioning of the several implementations of the embedded functionalities with different characteristics in resource usage and performance in order for the system to dynamically adapt itself under uncertainties. We propose an autonomic control architecture for self-adaptive and self-reconfigurable FPGA-based embedded systems. The control architecture is structured in three layers: a mission manager, a reconfiguration manager and a scheduling manager. This work is in the framework of the ANR project HPeC (see Section 9.2.1).

7.2.3.1. *DPR FPGA and discrete control for reconfiguration*

In this work we focus on the design of the reconfiguration manager. We propose a design approach using automata-based discrete control. It involves reactive programming that provides formal semantics, and discrete controller synthesis from declarative objectives.

Ongoing work concerns experimental validation, where upon the availability of hardware implementations of vision, detection and tracking tasks, a demonstrator is being built integrating our controller.

7.2.3.2. *Mission management and stochastic control*

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

This work was performed in the framework of the PhD of Chabha Hireche, defended in nov. 2019 [17].

7.2.4. IoT and Cyberphysical Systems

Participants: Neil Ayeb, Ayan Hore, Fabien Lefevre, Stéphane Mocanu, Jan Pristas, Eric Rutten, Gaetan Sorin, Mohsen Zargarani.

7.2.4.1. Device management

The research topic is targeting an adaptative and decentralized management for the IoT. It will contribute design methods for processes in virtualized gateways in order to enhance IoT infrastructures. More precisely, it concerns Device Management (DM) in the case of large numbers of connected sensors and actuators, as can be found in Smart Home and Building, Smart Electricity grids, and industrial frameworks as in Industry 4.0.

Device Management is currently industrially deployed for LAN devices, phones and workstation management. Internet of Things (IoT) devices are massive, dynamic, heterogeneous, and inter-operable. Existing solutions are not suitable for IoT management. This work in an industrial environment addresses these limitations with a novel autonomic and distributed approach for the DM.

This work is in the framework of the Inria/Orange labs joint laboratory (see Section 8.1.1), and supported by the CIFRE PhD thesis grant of Neïl Ayeb, starting dec. 2017. It was awarded a best paper distinction at the Doctoral Symposium of ICAC 2019 [12].

7.2.4.2. Security in SCADA industrial systems

We focus mainly on vulnerability search, automatic attack vectors synthesis and intrusion detection [11]. Model checking techniques are used for vulnerability search and automatic attack vectors construction. Intrusion detection is mainly based on process-oriented detection with a technical approach from run-time monitoring. The LTL formalism is used to express safety properties which are mined on an attack-free dataset. The resulting monitors are used for fast intrusion detections. A demonstrator of attack/defense scenario in SCADA systems has been built on the existing G-ICS lab (hosted by ENSE3/Grenoble-INP). This work is in the framework of the ANR project Sacade on cybersecurity of industrial systems (see Section 9.2.2).

One of important results is the realization of a Hardware-in-the-loop SCADA Cyberange based on a electronic interface card that allows to interface real-world PLC with a software simulation [21]. The entire system is available in open-source including the electronic card fabrication files (http://gics-hil.gforge.inria.fr/). Interfacing system allow connection with various commercial simulation software but also with “home made” simulators [15]. The work is also supported by Grenoble Alpes Cybersecurity Institute (see Section 9.1.1) and Pulse program of IRT NANOELEC.

Ongoing work concerns the complementary topic of analysis and identification of reaction mechanisms for self-protection in cybersecurity, where, beyond classical defense mechanisms that detect intrusions and attacks or assess the kind of danger that is caused by them, we explore models and control techniques for the automated reaction to attacks, in order to use detection information to take the appropriate defense and repair actions. A first approach was developed in the M2R internship by Ayan Hore [18]

DELYS Project-Team

5. New Results

5.1. Distributed Algorithms for Dynamic Networks and Fault Tolerance

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

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

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

In 2019, we obtained the following results.

5.1.1. Failure detectors

Mutual exclusion is one of the fundamental problems in distributed computing but existing mutual exclusion algorithms are unadapted to the dynamics and lack of membership knowledge of current distributed systems (e.g., mobile ad-hoc networks, peer-to-peer systems, etc.). Additionally, in order to circumvent the impossibility of solving mutual exclusion in asynchronous message passing systems where processes can crash, some solutions include the use of $(\mathcal{T}+\Sigma^l)$, which is the weakest failure detector to solve mutual exclusion in known static distributed systems. In [28], we define a new failure detector $\mathcal{T}\Sigma^{lr}$ which is equivalent to $(\mathcal{T}+\Sigma^l)$ in known static systems, and prove that $\mathcal{T}\Sigma^{lr}$ is the weakest failure detector to solve mutual exclusion in unknown dynamic systems with partial memory losses. We consider that crashed processes may recover.

Assuming a message-passing environment with a majority of correct processes, the necessary and sufficient information about failures for implementing a general state machine replication scheme ensuring consistency is captured by the Ω failure detector. We show in [19] that in such a message-passing environment, Ω is also the weakest failure detector to implement an eventually consistent replicated service, where replicas are expected to agree on the evolution of the service state only after some (a priori unknown) time.

5.1.2. Scheduler Tolerant to Temporal Failures in Clouds

Cloud platforms offer different types of virtual machines which ensure different guarantees in terms of availability and volatility, provisioning the same resource through multiple pricing models. For instance, in Amazon EC2 cloud, the user pays per hour for on-demand instances while spot instances are unused resources available for a lower price. Despite the monetary advantages, a spot instance can be terminated or hibernated by EC2 at any moment. Using both hibernation prone spot instances (for cost sake) and on-demand instances, we propose in [31] a static scheduling for applications which are composed of independent tasks (bag-of-task) with deadline constraints. However, if a spot instance hibernates and it does not resume within a time which guarantees the application's deadline, a temporal failure takes place. Our scheduling, thus, aims at minimizing monetary costs of bag-of-tasks applications in EC2 cloud, respecting its deadline and avoiding temporal failures. Performance results with task execution traces, configuration of Amazon EC2 virtual machines, and EC2 market history confirms the effectiveness of our scheduling and that it tolerates temporal failures. In [30], we extend our approach for dynamic scheduling.

5.1.3. Gathering of Mobile Agents

Gathering a group of mobile agents is a fundamental task in the field of distributed and mobile systems. It consists of bringing agents that initially start from different positions to meet all together in finite time. In the case when there are only two agents, the gathering problem is often referred to as the rendezvous problem.

In [14] we show that rendezvous under the strong scenario is possible for agents with asynchrony restricted in the following way: agents have the same measure of time but the adversary can impose, for each agent and each edge, the speed of traversing this edge by this agent. The speeds may be different for different edges and different agents but all traversals of a given edge by a given agent have to be at the same imposed speed. We construct a deterministic rendezvous algorithm for such agents, working in time polynomial in the size of the graph, in the length of the smaller label, and in the largest edge traversal time.

5.1.4. Perpetual self-stabilizing exploration of dynamic environments

In [15], 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.

5.1.5. Torus exploration by oblivious robots

In [17], we deal with a team of autonomous robots that are endowed with motion actuators and visibility sensors. Those robots are weak and evolve in a discrete environment. By weak, we mean that they are anonymous, uniform, unable to explicitly communicate, and oblivious. We first show that it is impossible to solve the terminating exploration of a simple torus of arbitrary size with less than 4 or 5 such robots, respectively depending on whether the algorithm is probabilistic or deterministic. Next, we propose in the SSYNC model a probabilistic solution for the terminating exploration of torus-shaped networks of size $\ell \times L$, where $7 \leq \ell \leq L$, by a team of 4 such weak robots. So, this algorithm is optimal *w.r.t.* the number of robots.

5.1.6. Explicit communication among stigmergic robots

In [18], we investigate avenues for the exchange of information (explicit communication) among deaf and mute mobile robots scattered in the plane. We introduce the use of movement-signals (analogously to flight signals and bees waggle) as a mean to transfer messages, enabling the use of distributed algorithms among robots. We propose one-to-one deterministic movement protocols that implement explicit communication among semi-synchronous robots. Our protocols enable the use of distributing algorithms based on message exchanges among swarms of stigmergic robots. They also allow robots to be equipped with the means of communication to tolerate faults in their communication devices.

5.1.7. Gradual stabilization

In [13], we introduce the notion of *gradual stabilization under* (τ, ρ)*-dynamics* (gradual stabilization, for short). A gradually stabilizing algorithm is a self-stabilizing algorithm with the following additional feature: after up to τ*dynamic steps* of a given type ρ occur starting from a legitimate configuration, it first quickly recovers to a configuration from which a specification offering a minimum quality of service is satisfied.

It then gradually converges to specifications offering stronger and stronger safety guarantees until reaching a configuration (1) from which its initial (strong) specification is satisfied again, and (2) where it is ready to achieve gradual convergence again in case of up to τ new dynamic steps of type ρ. A gradually stabilizing algorithm being also self-stabilizing, it still recovers within finite time (yet more slowly) after any other finite number of transient faults, including for example more than τ arbitrary dynamic steps or other failure patterns such as memory corruptions. We illustrate this new property by considering three variants of a synchronization problem respectively called *strong*, *weak*, and *partial* unison. We propose a self-stabilizing unison algorithm which achieves gradual stabilization in the sense that after one dynamic step of a certain type *BULCC* (such a step may include several topological changes) occurs starting from a configuration which is legitimate for the strong unison, it maintains clocks almost synchronized during the convergence to strong unison: it satisfies partial unison immediately after the dynamic step, then converges in at most one round to weak unison, and finally re-stabilizes to strong unison.

5.2. Distributed systems and Large-scale data distribution

Participants: Guillaume Fraysse, Saalik Hatia, Mesaac Makpangou, Sreeja Nair, Jonathan Sid-Otmane, Pierre Sens, Marc Shapiro, Ilyas Toumlilt, Dimitrios Vasilas.

5.2.1. Proving the safety of highly-available distributed objects

To provide high availability in distributed systems, object replicas allow concurrent updates. Although replicas eventually converge, they may diverge temporarily, for instance when the network fails. This makes it difficult for the developer to reason about the object's properties , and in particular, to prove invariants over its state. For the sub-class of state-based distributed systems, we propose a proof methodology for establishing that a given object maintains a given invariant, taking into account any concurrency control. Our approach allows reasoning about individual operations separately. We demonstrate that our rules are sound, and we illustrate their use with some representative examples. We automate the rule using Boogie, an SMT-based tool.

This work is accepted for publication at the 29th European Symposium on Programming (ESOP), April 2020, Dublin, Ireland [34]. Preliminary results were presented at the Workshop on Principles and Practice of Consistency for Distributed Data (PaPoC), March 2019, Dresden, Germany [29].

5.3. Resource management in system software

Participants: Jonathan Lejeune, Marc Shapiro, Julien Sopena, Francis Laniel.

5.3.1. MemOpLight: Leveraging applicative feedback to improve container memory consolidation

The container mechanism supports consolidating several servers on the same machine, thus amortizing cost. To ensure performance isolation between containers, Linux relies on memory limits. However these limits are static, but application needs are dynamic; this results in poor performance. To solve this issue, MemOpLight reallocates memory to containers based on dynamic applicative feedback. MemOpLight rebalances physical memory allocation, in favor of under-performing ones, with the aim of improving overall performance. Our research explores the issues, addresses the design of MemOpLight, and validates it experimentally. Our approach increases total satisfaction by 13% compared to the default.

It is standard practice in Infrastructure as a Service to *consolidate* several logical servers on the same physical machine, thus amortizing cost. However, the execution of one logical server should not disturb the others: the logical servers should remain *isolated* from one another.

To ensure both consolidation and isolation, a recent approach is "containers," a group of processes with sharing and isolation properties. To ensure *memory performance isolation*, *i.e.*, guaranteeing to each container enough memory for it to perform well, the administrator limits the total amount of physical memory that a container may use at the expense of others. In previous work, we showed that these limits impede memory consolidation [26]. Furthermore, the metrics available to the kernel to evaluate its policies (*e.g.*, frequency of page faults, I/O requests, use of CPU cycles, *etc.*), are not directly relevant to performance as experienced from the application perspective, which is better characterized by, for instance, response time or throughput measured at application level.

To solve these problems, we propose a new approach, called the Memory Optimization Light (MemOpLight). It is based on application-level feedback from containers. Our mechanism aims to rebalance memory allocation in favor of unsatisfied containers, while not penalizing the satisfied ones. By doing so, we guarantee application satisfaction, while consolidating memory; this also improves overall resource consumption.

Our main contributions are the following:

- An experimental demonstration of the limitations of the existing Linux mechanisms.
- The design of a simple feedback mechanism from application to the kernel.
- An algorithm for adapting container memory allocation.
- And implementation in Linux and experimental confirmation.

This work is currently under submission at a major conference. Some preliminary results are published at NCA 2019 [26].

MIMOVE Project-Team

7. New Results

7.1. Automated Synthesis of Mediators for Middleware-Layer Protocol Interoperability in the IoT

Participants: Georgios Bouloukakis, Nikolaos Georgantas, Patient Ntumba, Valérie Issarny (MiMove)
To enable direct Internet connectivity of Things, complete protocol stacks need to be deployed on resource-constrained devices. Such protocol stacks typically build on lightweight IPv6 adaptations and may even include a middleware layer supporting high-level application development. However, the profusion of IoT middleware-layer interaction protocols has introduced technology diversity and high fragmentation in the IoT systems landscape with siloed vertical solutions. To enable the interconnection of heterogeneous Things across these barriers, advanced interoperability solutions at the middleware layer are required. In this paper, we introduce a solution for the automated synthesis of protocol mediators that support the interconnection of heterogeneous Things. Our systematic approach relies on the Data eXchange (DeX) connector model, which comprehensively abstracts and represents existing and potentially future IoT middleware protocols. Thanks to DeX, Things seamlessly interconnect through lightweight mediators. We validate our solution with respect to: (i) the support to developers when developing heterogeneous IoT applications; (ii) the runtime performance of the synthesized mediators.

7.2. Probabilistic Event Dropping for Intermittently Connected Subscribers over Pub/Sub Systems

Participants: Georgios Bouloukakis, Nikolaos Georgantas (MiMove), Ioannis Moscholios (Univ of Peloponnese)
Internet of Things (IoT) aim to leverage data from multiple sensors, actuators and devices for improving peoples' daily life and safety. Multiple data sources must be integrated, analyzed from the corresponding application and notify interested stakeholders. To support the data exchange between data sources and stakeholders, the publish/subscribe (pub/sub) middleware is often employed. Pub/sub provides additional mechanisms such as reliable messaging, event dropping, prioritization, etc. The event dropping mechanism is often used to satisfy Quality of Service (QoS) requirements and ensure system stability. To enable event dropping, basic approaches apply finite buffers or data validity periods and more sophisticated ones are information-aware. In this paper, we introduce a pub/sub mechanism for probabilistic event dropping by considering the stakeholders' intermittent connectivity and QoS requirements. We model the pub/sub middleware as a network of queues which includes a novel ON/OFF queueing model that enables the definition of join probabilities. We validate our analytical model via simulation and compare our mechanism with existing ones. Experimental results can be used as insights for developing hybrid dropping mechanisms.

7.3. Adaptive Mediation for Data Exchange in IoT Systems

Participants: Georgios Bouloukakis (MiMove & Univ of California, Irvine), Andrew Chio, Sharad Mehrotra, Nalini Venkatasubramanian (Univ of California, Irvine), Cheng-Hsin Hsu (National Tsing Hua Univ)
Messaging and communication is a critical aspect of next generation Internet-of-Things (IoT) systems where interactions among devices, software systems/services and end-users is the expected mode of operation. Given the diverse and changing communication needs of entities, the data exchange interactions may assume different protocols (MQTT, CoAP, HTTP) and interaction paradigms (point to point, multicast, unicast). In this paper, we address the issue of supporting adaptive communications in IoT systems through a mediation-based architecture for data exchange. Here, components called mediators support protocol translation to bridge the heterogeneity gap. Aiming to provide a placement of mediators to nodes, we introduce an integer linear programming solution that takes as input: a set of Edge nodes, IoT devices, and networking semantics. Our proposed solution achieves adaptive placement resulting in timely interactions between IoT devices for larger topologies of IoT spaces.

7.4. Universal Social Network Bus: Toward the Federation of Heterogeneous Online Social Network Services

Participants: Valérie Issarny, Nikolaos Georgantas, Ehsan Ahvar, Bruno Lefèvre, Shohreh Ahvar (MiMove), Rafael Angarita (ISEP Paris)

Online Social Network Services (OSNSs) are changing the fabric of our society, impacting almost every aspect of it. Over the past few decades, an aggressive market rivalry has led to the emergence of multiple competing, "closed" OSNSs. As a result, users are trapped in the walled gardens of their OSNS, encountering restrictions about what they can do with their personal data, the people they can interact with, and the information they get access to. As an alternative to the platform lock-in, "open" OSNSs promote the adoption of open, standardized APIs. However, users still massively adopt closed OSNSs to benefit from the services' advanced functionalities and/or follow their "friends," although the users' virtual social sphere is ultimately limited by the OSNSs they join. Our work aims at overcoming such a limitation by enabling users to meet and interact beyond the boundary of their OSNSs, including reaching out to "friends" of distinct closed OSNSs. We specifically introduce *Universal Social Network Bus* (USNB), which revisits the "service bus" paradigm that enables interoperability across computing systems to address the requirements of "social interoperability." USNB features synthetic profiles and personae for interaction across the boundaries of closed and open and profile- and non-profile-based OSNSs through a reference social interaction service. We ran a 1-day workshop with a panel of users who experimented with the USNB prototype to assess the potential benefits of social interoperability for social network users. Results show the positive evaluation of users for USNB, especially as an enabler of applications for civic participation. This further opens up new perspectives for future work, among which includes enforcing security and privacy guarantees.

7.5. Social Middleware for Civic Engagement

Participants: Valérie Issarny, Nikolas Georgantas, Grigoris Piperagkas (MiMove), Rafael Angarita (ISEP Paris)

Civic engagement refers to any collective action towards the identification and solving of public issues. Current civic technologies are traditional Web- or mobile-based platforms that make difficult, or just impossible, the participation of citizens via different communication technologies. Moreover, connected objects sensing physical-world data can nourish participatory processes by providing physical evidence to citizens; however, leveraging these data is not direct and still a time-consuming process for civic technologies developers. We introduce the concept of *social middleware* for civic engagement. Social middleware allows citizens to engage in participatory processes -supported by civic technologies- via their favorite communication tools, and to interact not only with other citizens but also with relevant connected objects and software platforms. The mission of social middleware goes beyond the connection of all these heterogeneous entities. It aims at easing the implementation of distributed applications oriented toward civic engagement by featuring dedicated built-in services.

7.6. Mobile Crowd-Sensing as a Resource for Contextualized Urban Public Policies

Participants: Valŕie Issarny, Bruno Lefèvre, Rachit Agarwal (MiMove), Vivien Mallet (Inria Ange)

Environmental noise is a major pollutant in contemporary cities and calls for the active monitoring of noise levels to spot the locations where it most affects the people's health and well-being. However, due to the complex relationship between environmental noise and its perception by the citizens, it is not sufficient to quantitatively measure environmental noise. We need to collect and aggregate contextualized –both quantitative and qualitative– data about the urban environmental noise so as to be able to study the objective and subjective relationships between sound and living beings. This complex knowledge is a prerequisite for making efficient territorial public policies for soundscapes that are inclined towards living beings welfare. In this paper, we investigate how Mobile Phone Sensing (MPS) –*aka* crowdsensing– enables the gathering of such knowledge, provided the implementation of sensing protocols that are customized according to the context of use and the intended exploitation of the data. Through three case studies that we carried out in France and Finland, we show that MPS is not solely a tool that contributes to sensitizing citizens and decision-makers about noise pollution; it also contributes to increasing our knowledge about the impact of the environmental noise on people's health and well-being in relation to its physical and subjective perception.

7.7. Multi-Sensor Calibration Planning in IoT-Enabled Smart Spaces

Participants: Valérie Issarny (MiMove), Françoise Sailhan (CNAM), Qiuxi Zhu, Md Yusuf Sarwar Uddin, Nalini Venkatasubramanian (University of California, Irvine)

Emerging applications in smart cities and communities require massive IoT deployments using sensors/actuators (things) that can enhance citizens' quality of life and public safety. However, budget constraints often lead to limited instrumentation and/or the use of low-cost sensors that are subject to drift and bias. This raises concerns of robustness and accuracy of the decisions made on uncertain data. To enable effective decision making while fully exploiting the potential of low-cost sensors, we propose to send mobile units (e.g., trained personnel) equipped with high-quality (more expensive) and freshly-calibrated reference sensors so as to carry out calibration in the field. We design and implement an efficient cooperative approach to solve the calibration planning problem, which aims at minimizing the cost of the recurring calibration of multiple sensor types in the long-term operation. We propose a two-phase solution that consists of a sensor selection phase that minimizes the average cost of recurring calibration, and a path planning phase that minimizes the travel cost of multiple calibrators which have load constraints. We provide fast and effective heuristics for both phases. We further build a prototype that facilitates the mapping of the deployment field and provides navigation guidance to mobile calibrators. Extensive use-case-driven simulations show that our proposed approach significantly reduces the average cost compared to naive approaches: up to 30% in a moderate-sized indoor case, and higher in outdoor cases depending on scale

7.8. User-Centric Context Inference for Mobile Crowdsensing

Participants: Yifan Du, Valérie Issarny (MiMove), Françoise Sailhan (CNAM)

Mobile crowdsensing is a powerful mechanism to aggregate hyperlocal knowledge about the environment. Indeed, users may contribute valuable observations across time and space using the sensors embedded in their smartphones. However, the relevance of the provided measurements depends on the adequacy of the sensing context with respect to the phenomena that are analyzed. Our research concentrates more specifically on assessing the sensing context when gathering observations about the physical environment beyond its geographical position in the Euclidean space, i.e., whether the phone is in-/out-pocket, in-/out-door and on-/under-ground. We introduce an online learning approach to the local inference of the sensing context so as to overcome the disparity of the classification performance due to the heterogeneity of the sensing devices as well as the diversity of user behavior and novel usage scenarios. Our approach specifically features a hierarchical algorithm for inference that requires few opportunistic feedbacks from the user, while increasing the accuracy of the context inference per user.

7.9. Let Opportunistic Crowdsensors Work Together for Resource-efficient, Quality-aware Observations

Participants: Yifan Du, Valérie Issarny (MiMove), Françoise Sailhan (CNAM)

Opportunistic crowdsensing empowers citizens carrying hand-held devices to sense physical phenomena of common interest at a large and fine-grained scale without requiring the citizens' active involvement. However, the resulting uncontrolled collection and upload of the massive amount of contributed raw data incur significant resource consumption, from the end device to the server, as well as challenge the quality of the collected observations. Our research tackles both challenges raised by opportunistic crowdsensing, that is, enabling the resource-efficient gathering of relevant observations. To achieve so, we introduce the *BeTogether* middleware fostering context-aware, collaborative crowdsensing at the edge so that co-located crowdsensors operating in the same context, group together to share the work load in a cost- and quality-effective way. Our implementation-driven evaluation of the proposed solution, which leverages a dataset embedding nearly one million entries contributed by 550 crowdsensors over a year, shows that *BeTogether* increases the quality of the collected data while reducing the overall resource cost compared to the cloud-centric approach.

7.10. Detecting Mobile Crowdsensing Context in the Wild

Participants: Rachit Agarwal, Shaan Chopra, Vassilis Christophides, Nikolaos Georgantas, Valérie Issarny (MiMove)

Understanding the sensing context of raw data is crucial for assessing the quality of large crowdsourced spatio-temporal datasets and supporting context-augmented personal trajectories. Detecting sensing contexts in the wild is a challenging task and requires features from smartphone sensors that are not always available. In this paper, we propose three heuristic algorithms for detecting sensing contexts such as in/out-pocket, under/over-ground, and in/out-door for crowdsourced spatio-temporal datasets. These are unsupervised binary classifiers with a small memory footprint and execution time. Using a segment of the Ambiciti real dataset-a feature-limited crowdsourced dataset-we report that our algorithms perform equally well in terms of balanced accuracy (within 4.3%) when compared to machine learning (ML) models reported by an AutoML tool.

7.11. Inferring Streaming Video Quality from Encrypted Traffic: Practical Models and Deployment Experience

Participants: Francesco Bronzino, Sara Ayouibi, Renata Teixeira (MiMove), Paul Schmitt (Princeton), Guilherme Martins, Nick Feamster (University of Chicago)

Inferring the quality of streaming video applications is important for Internet service providers, but the fact that most video streams are encrypted makes it difficult to do so. We develop models that infer quality metrics (i.e., startup delay and resolution) for encrypted streaming video services. Our paper builds on previous work, but extends it in several ways. First, the models work in deployment settings where the video sessions and segments must be identified from a mix of traffic and the time precision of the collected traffic statistics is more coarse (e.g., due to aggregation). Second, we develop a single composite model that works for a range of different services (i.e., Netflix, YouTube, Amazon, and Twitch), as opposed to just a single service. Third, unlike many previous models, our models perform predictions at finer granularity (e.g., the precise startup delay instead of just detecting short versus long delays) allowing to draw better conclusions on the ongoing streaming quality. Fourth, we demonstrate the models are practical through a 16-month deployment in 66 homes and provide new insights about the relationships between Internet "speed" and the quality of the corresponding video streams, for a variety of services; we find that higher speeds provide only minimal improvements to startup delay and resolution. This work was accepted for publication at the ACM SIGMETRICS conference. The models we developed in this work and the findings were the basis for a first-page story published on The Wall Street Journal (“The Truth About Faster Internet: It’s Not Worth It”). [0]

7.12. Implications of User Perceived Page Load Time Multi-Modality on Web QoE Measurement

Participants: Renata Teixeira, Vassilis Christophides (MiMove), Flavia Salutari, Diego Da Hora (Telecom Paris Tech), Matteo Varvello (Brave Software), Dario Rossi (Huawei)

[0] The article is available online at: https://www.wsj.com/graphics/faster-internet-not-worth-it/?mod=article_inline&mod=hp_lead_pos5.

Web browsing is one of the most popular applications for both desktop and mobile users. A lot of effort has been devoted to speedup the Web, as well as in designing metrics that can accurately tell whether a webpage loaded fast or not. An often implicit assumption made by industrial and academic research communities is that a *single* metric is sufficient to assess whether a webpage loaded fast. In this work we collect and make publicly available a unique dataset which contains webpage features (e.g., number and type of embedded objects) along with both *objective* and *subjective* Web quality metrics. This dataset was collected by crawling over 100 websites—representative of the top 1 M websites in the Web—while crowdsourcing 6,000 user opinions on *user perceived page load time* (uPLT). In contrast to related work, we show that the uPLT distribution is often multimodal and that, in practice, no more than three modes are present. The main conclusion drawn from our analysis is that, for complex webpages, each of the different objective QoE metrics proposed in the literature (such as AFT, TTI, PLT, etc.) is suited to approximate one of the different uPLT modes.

7.13. The News We Like Are Not the News We Visit: News Categories Popularity in Usage Data

Participants: Renata Teixeira (MiMove), Giuseppe Scavo (MiMove, Nokia Bell Labs), Zied Ben-Houidi (Nokia Bell-Labs), Stefano Traverso, Marco Mellia (Politecnico di Torino)
Most of our knowledge about online news consumption comes from survey-based news market reports, partial usage data from a single editor, or what people publicly share on social networks. Our work published on the 13th International AAAI Conference on Web and Social Media (ICWSM-2019) complements these sources by presenting the first holistic study of visits across online news outlets that a population uses to read news. We monitored the entire network traffic generated by Internet users in four locations in Italy. Together these users generated 80 million visits to 5.4 million news articles in about one year and a half. This unique view allowed us to evaluate how usage data complements existing data sources. We find for instance that only 16% of news visits in our datasets came from online social networks. In addition, the popularity of news categories when considering all visits is quite different from the one when considering only news discovered on social media, or visits to a single major news outlet. Interestingly, a substantial mismatch emerges between self-reported news-category preferences (as measured by Reuters Institute in the same year and same country) and their actual popularity in terms of visits in our datasets. In particular, unlike self-reported preferences expressed by users in surveys that put “Politics”, “Science” and “International” as the most appreciated categories, “Tragedies and Weird news" and “Sport” are by far the most visited. Our paper discusses two possible causes of this mismatch and conjecture that the most plausible reason is the disassociation that may occur between individuals’ cognitive values and their cue-triggered attraction.

7.14. Classification of Load Balancing in the Internet

Participants: Renata Teixeira (MiMove), Rafael Almeida, Ítalo Cunha (Universidade Federal de Minas Gerais), Darryl Veitch (University of Technology Sydney), Christophe Diot (Google)

Recent advances in programmable data planes, software-defined networking, and the adoption of IPv6, support novel, more complex load balancing strategies. We introduce the Multipath Classification Algorithm (MCA), a probing algorithm that extends traceroute to identify and classify load balancing in Internet routes. MCA extends existing formalism and techniques to consider that load balancers may use arbitrary combinations of bits in the packet header for load balancing. We propose optimizations to reduce probing cost that are applicable to MCA and existing load balancing measurement techniques. Through large-scale measurement campaigns, we characterize and study the evolution of load balancing on the IPv4 and IPv6 Internet with multiple transport protocols. Our results that will appear in the IEEE INFOCOM 2020 conference show that load balancing is more prevalent and that load balancing strategies are more mature than previous characterizations have found.

7.15. MinoanER: Schema-Agnostic, Non-Iterative, Massively Parallel Resolution of Web Entities

Participants: Vassilis Christophides (MiMove), Vasilis Efthymiou (IBM Almaden Research Center), George Papadakis (Univ of Athens), Kostas Stefanidis (Univ of Tampere)

Entity Resolution (ER) aims to identify different descriptions in various Knowledge Bases (KBs) that refer to the same entity. ER is challenged by the Variety, Volume and Veracity of entity descriptions published in the Web of Data. To address them, we propose the MinoanER framework that simultaneously fulfills full automation, support of highly heterogeneous entities, and massive parallelization of the ER process. MinoanER leverages a token-based similarity of entities to define a new metric that derives the similarity of neighboring entities from the most important relations, as they are indicated only by statistics. A composite blocking method is employed to capture different sources of matching evidence from the content, neighbors, or names of entities. The search space of candidate pairs for comparison is compactly abstracted by a novel disjunctive blocking graph and processed by a non-iterative, massively parallel matching algorithm that consists of four generic, schema-agnostic matching rules that are quite robust with respect to their internal configuration. We demonstrate that the effectiveness of MinoanER is comparable to existing ER tools over real KBs exhibiting low Variety, but it outperforms them significantly when matching KBs with high Variety.

Myriads Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Efficient Docker container deployment in fog environments

Participants: Arif Ahmed, Lorenzo Civolani, Guillaume Pierre, Paulo Rodrigues de Souza Junior.

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

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

We investigated the challenges of efficiently deploying Docker containers in fog platforms composed of tiny single-board computers such as Raspberry PIs. Significant improvements in the Docker image cache hit rate can be obtained by sharing the caches of multiple co-located servers rather than letting them operate independently [9]. In the case when an image must be downloaded and locally installed, large performance gains can be obtained with relatively simple modifications in the way Docker imports container images [3]. Finally, we showed (in collaboration with Prof. Paolo Bellavista from the University of Bologna) that it is possible to let a container start producing useful work even before its image has been fully downloaded [14]. Another paper in this direction of work is in preparation about the way to speedup the boot phase of Docker containers. We are also exploring innovative techniques to improve the performance of live container migration in fog computing environments.

7.1.2. Fog computing platform design

Participants: Ali Fahs, Ayan Mondal, Nikos Parlavantzas, Guillaume Pierre, Mulugeta Tamiru.

There does not yet exist any reference platform for fog computing platforms. We therefore investigated how Kubernetes could be adapted to support the specific needs of fog computing platforms. In particular we focused on the problem of redirecting end-user traffic to a nearby instance of the application. When different users impose various load on the system, any traffic routing system must necessarily implement a tradeoff between proximity and fair load-balancing between the application instances. We demonstrated how such customizeable traffic routing policies can be integrated in Kubernetes to help transform it in a suitable platform for fog computing [15]. We extended this work to let the platform automatically choose (and maintain over time) the best locations where application replicas should be deployed. A paper on this topic is currently under submission. We finally started addressing the topic of application autoscaling such that the system can enforce performance guarantees despite traffic variations. We expect one or two publications on this topic next year.

In collaboration with Prof. Misra from IIT Kharagpur (India), and thanks to the collaboration established by the FogCity associate team, we developed mechanisms based on game theory to assign resources to competing applications in a fog computing platform. The objective of those mechanisms is to satisfy user preferences while maximizing resource utilisation. We evaluated the mechanisms using an emulated fog platform built on Kubernetes and Grid'5000, and showed that they significantly outperform baseline algorithms. A paper on this topic is in preparation.

7.1.3. Edgification of micro-service applications

Participants: Genc Tato, Cédric Tedeschi, Marin Bertier.

Last year, we investigated in collaboration with Etienne Riviere from UC Louvain the feasibility and possible benefits brought about by the *edgification* of a legacy micro-service-based application [35]. In other words, we devised a method to classify services composing the application as *edgifiable* or not, based on several criteria. We applied this method to the particular case of the ShareLatex application which enables the collaborative edition of LaTeX documents. Recently, we continue this work by automate the localization and the migration of microservices. Our middleware, based on Koala [36], a lightweight Distributed Hash Table, allows adapting compatible legacy microservices applications for hybrid core/edge deployments [21].

7.1.4. Community Clouds

Participants: Jean-Louis Pazat, Bruno Stevant.

Small communities of people who need to share data and applications can now buy inexpensive devices in order to use only "on premise" resources instead of public Clouds. This "self-hosting-and-sharing" solution provides a better privacy and does not need people to pay any monthly fee to a resource provider. We have implemented a prototype based on micro-services in order to be able to distribute the load of applications among devices.

However, such a distributed platform needs to rely on a very good distribution of the computing and communication load over the devices. Using an emulator of the system, we have shown that, thanks to well known optimization techniques (Particle Swarm Optimization), it is possible to quickly find a service placement resulting in a response time close to the optimal one.

This year we evaluated the results of the optimization algorithm on a prototype (5 "boxes" installed in different home locations connected by fiber or ADSL). Results shown that due to the variation of the network available bandwidth it is necessary to dynamically modify the deployment of applications. This was not a big surprise, but we were not able to find any predictive model of this variation during a day. So, we developed and experimented a dynamic adaptation of the placement of micro-services based applications based on a regular monitoring of the response time of applications. We plan to submit a paper on this topic in early 2020.

7.1.5. Geo-distributed data stream processing

Participants: Hamidreza Arkian, Davaadorj Battulga, Mehdi Belkhiria, Guillaume Pierre, Cédric Tedeschi.

We investigated a decentralized scaling mechanism for stream processing applications where the different operators composing the processing topology are able to take their own scaling decisions independently, based on local information. We built a simulation tool to validate the ability of our algorithm to react to load variation. Then, we started the development of a software prototype of a decentralized Stream Processing Engine including this autoscaling mechanism, and deployed it over the Grid'5000 platform. Two papers have been accepted in 2019 about this work [11], [12].

Although data stream processing platforms such as Apache Flink are widely recognized as an interesting paradigm to process IoT data in fog computing platforms, the existing performance model to capture of stream processing in geo-distributed environments are theoretical works only, and have not been validated against empirical measurements. We developed and experimentally validated such a model to represent the performance of a single stream processing operator [10]. This model is very accurate with predictions $\pm 2\%$ of the actual values even in the presence of heterogeneous network latencies. Individual operator models can be composed together and, after the initial calibration of a first operator, a reasonably accurate model for other operators can be derived from a single measurement only.

7.1.6. QoS-aware and energy-efficient resource management for Function-as-a-Service

Participants: Yasmina Bouizem, Christine Morin, Nikos Parlavantzas.

Recent years have seen the widespread adoption of serverless computing, and in particular, Function-as-a-Service (FaaS) systems. These systems enable users to execute arbitrary functions without managing underlying servers. However, existing FaaS frameworks provide no quality of service guarantees to FaaS users in terms of performance and availability. Moreover, they provide no support for FaaS providers to reduce energy consumption. The goal of this work is to develop an automated resource management solution for FaaS plaforms that takes into account performance, availability, and energy efficiency in a coordinated manner. This work is performed in the context of the thesis of Yasmina Bouizem. In 2019, we integrated a fault-tolerance mechanism into Fission, an open-source FaaS framework based on Kubernetes, and are currently evaluating its impact on performance, availability, and energy consumption.

7.2. Greening Clouds

7.2.1. Energy Models

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

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

The costs of current data centers are mostly driven by their energy consumption (specifically by the air conditioning, computing and networking infrastructures). Yet, current pricing models are usually static and rarely consider the facilities' energy consumption per user. The challenge is to provide a fair and predictable model to attribute the overall energy costs per virtual machine and to increase energy-awareness of users. We aim at proposing such energy cost models without heavily relying on physical wattmeters that may be costly to install and operate. These results have been published in [24].

Another goal consists in better understanding the energy consumption of computing and networking resources of Clouds in order to provide energy cost models for the entire infrastructure including incentivizing cost models for both Cloud providers and energy suppliers. These models should be based on experimental measurement campaigns on heterogeneous devices. As hardware architectures become more complex, measurement campains are required to better understand their energy consumption and to identify potential sources of energy waste. These results, conducted with Amina Guermouche (IMT Telecom SudParis), have been presented in [30].

Similarly, software stacks add complexity in the identification of energy inefficiencies. For HPC applications, precise measurements are required to determine the most efficient options for the runtime, the resolution algorithm and the mapping on physical resources. An example of such a study has been published in collaboration with HiePACS (Bordeaux) and NACHOS (Sophia) teams in [8].

The fine-grain measurements lead us to propose models that have been used to compare different Cloud architectures (from fog and edge to centralized clouds) in terms of energy consumption on a given scenario. These results have been published in [4].

Inferring a cost model from energy measurements is an arduous task since simple models are not convincing, as shown in our previous work. We aim at proposing and validating energy cost models for the heterogeneous Cloud infrastructures in one hand, and the energy distribution grid on the other hand. These models will be integrated into simulation frameworks in order to validate our energy-efficient algorithms at larger scale. In particular, this year we implemented in SimGrid a flow-based energy model for wired network devices [17].

7.2.2. End-to-end energy models for the Internet of Things

Participants: Anne-Cécile Orgerie, Loic Guegan.

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

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

We investigate the end-to-end energy consumption of IoT platforms. Our aim is to evaluate, on concrete use-cases, the benefits of edge computing platforms for IoT regarding energy consumption. We aim at proposing end-to-end energy models for estimating the consumption when offloading computation from the objects to the Cloud, depending on the number of devices and the desired application QoS. This work has been published in [18].

7.2.3. *Exploiting renewable energy in distributed clouds*

Participants: Benjamin Camus, Anne-Cécile Orgerie.

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

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

7.2.4. *Smart Grids*

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

Smart grids allow to efficiently perform demand-side management in electrical grids in order to increase the integration of fluctuating and/or intermittent renewable energy sources in the energy mix. In this work, we consider the computing infrastructure that controls the smart grid. This infrastructure comprises communication and computing resources to allow for a smart management of the electrical grid. In particular, we study the influence of communication latency over a shedding scenario on a small-scale electrical network. We show that depending on the latency some shedding strategies are not feasible [13].

7.3. Securing Clouds

7.3.1. *Security monitoring in Cloud computing platforms*

Participants: Clément Elbaz, Christine Morin, 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 the past years we proposed a verification method for security monitoring SLOs [37] and we have then studied a methodology to define security monitoring SLOs that are at the same time relevant for the tenant, achievable for the provider, and verifiable. The methodology is based on metrics benchmarks that a cloud service provider runs on a set of basic setups of an NIDS (Network Intrusion Detection), the basic setups covering together the variety of NIDS rules that may interest tenants. In order to make it achievable for a cloud service provider to run such benchmarks despite thousands of rules that could be chosen individually by tenants, we proposed a rule clustering strategy to lower the number of sets of rules that should be benchmarked and thus the number of benchmarks run. Finally we proposed extensions to an existing cloud SLA language to define security monitoring SLOs. These results were published in a technical report [27] as well as in Amir Teshome Wonjiga's thesis (to appear) and were submitted for publication in an international conference.

In a side project with Dr Sean Peisert at LBNL, the work on security SLO verification was extended to the use case of data integrity, where tenants outsource data to a cloud storage provider. This work allowed us to tackle a challenge in SLO verification because, in this use case as well as in the security monitoring use case, tenants cannot verify SLOs without a minimal trust in providers involvment in the verification process. We proposed a strategy based on blockchains that allows tenants as well as providers to do SLO verification without having to trust any individual entity. This work was published in the CIFS security workshop [22].

To make security monitoring SLOs adaptable to context changes like the evolution of threats and updates to the tenants' software, we have worked on automating the mitigation of new threats during the time window in which no intrusion detection rule exist and no security patch is applied yet (if available). This time winwow is critical because newly published vulnerabilities get exploited up to five orders of magnitude right after they are published and the time window may last several days or weeks. We have worked on a first step of mitigation, which consists in deciding if a newly published vulnerabiliy impacts a given information system. A major challenge in automating this step is that newly published vulnerabilities do not contain machine-readable data and this data only appears up to several weeks later. For this reason we designed and evaluated a keyword extraction process from the free-form text description of a vulnerability to map a given vulnerability to product names. This keyword exctraction process was first published at the RESSI French security conference [23] and will appear in the NOMS 2020 international conference. In future work this mapping should be combined with a knowledge base of the information system to automatically score the impact of a new vulnerability on the information system.

Our results were published in [27], [28], [22], [23], [25].

7.3.2. *Privacy monitoring in Fog computing platforms*

Participants: Mozhdeh Farhadi, Guillaume Pierre.

IoT devices are integrated in our daily lives, and as a result they often have access to lots of private information. For example many digital assistants (Alexa, Amazon Echo...) were shown to have violated the privacy policy they had established themselves. To increase the level of confidence that end users may have in these devices and the applications which process their data, we started designing monitoring mechanisms such that the fog or the cloud platform can certify whether an application actually follows its own privacy policy or not. A survey paper on security of fog computing platforms is under submission, and we expect another paper on privacy monitoring in 2020.

7.4. Experimenting with Clouds

7.4.1. *Simulating distributed IT systems*

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

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

This year, we pursued our effort to ensure that SimGrid becomes a *de facto* standard for the simulation of distributed IT platforms. We further polished the new interface to ensure that it correctly captures the concepts needed by the experimenters, and provided a Python binding to smooth the learning curve. To that extend, we also continued our rewriting of the documentation.

The work on SimGrid is fully integrated to the other research efforts of the Myriads team. This year, we added the ability to co-simulate IT systems with SimGrid and physical systems modeled with equational systems [13]. This work, developed to study the co-evolution of thermal systems or of the electic grid with the IT system, is now distributed as an official plugin of the SimGrid framework.

7.4.2. Formal methods for IT systems

Participants: Ehsan Azimi, The Anh Pham, Martin Quinson.

The SimGrid framework also provide a state of the art Model-Checker for MPI applications. This can be used to formally verify whether the application entails synchronization issues such as deadlocks or livelocks [32]. This year, we pursued our effort on this topic, in collaboration with Thierry Jéron (EPI SUMO).

The Anh Pham defended his thesis this year on techniques to mitigate the state space explosion while verifying asynchronous distributed applications. He adapted an algorithm leveraging event folding structures to this context. This allows to efficiently compute how to not explore equivalent execution traces more than once. This work was published this year[19]. This work, co-advised by Martin Quinson with Thierry Jéron (team SUMO, formal methods), was important to bridge the gap between the involved communities.

Ehsan Azimi joined the Myriads team as an engineer in December to integrate the results of this thesis into the SimGrid framework.

7.4.3. Executing epidemic simulation applications in the Cloud

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

In the context of the DiFFuSE ADT and in collaboration with INRA researchers, we transformed a legacy application for simulating the spread of Mycobacterium avium subsp. paratuberculosis (MAP) to a cloud-enabled application based on the DiFFuSE framework (Distributed framework for cloud-based epidemic simulations). This is the second application to which the DiFFuSE framework is applied. The first application was a simulator of the spread of the bovine viral diarrhea virus, developed within the MIHMES project (2012-2017). Using both the MAP and BVDV applications, we performed extensive experiments showing the advantages of the DiFFuSE framework. Specifically, we showed that DiFFuSE enhances application performance and allows exploring different cost-performance trade-offs while supporting automatic failure handling and elastic resource acquisition from multiple clouds [7].

7.4.4. Tools for experimentation

Participant: Matthieu Simonin.

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

In this spirit we discuss in [31] a possible solution to the above desiderata. We illustrate its use in a real world use case study which has been completed in [34]. We show that an experimenter can express their experimental workflow and execute it in a safe manner (side effects are controlled) which increases the repeatability of the experiments.

The outcome is a library (EnOSlib) target reusability in experiment driven research in distributed systems. The library can be found in https://bil.inria.fr/fr/software/view/3589/tab.

SPIRALS Project-Team

7. New Results

7.1. Browser fingerprinting

We obtained new results on the concept of browser fingerprinting. This is a major technique of Internet security that is widely used for many purposes such as tracking activities, enhancing authentication, detecting bots, just to name a few. These results contribute to the enhancement of security for distributed software systems.

Our contributions to browser fingerprinting include the following three elements. First, we collected 122K fingerprints from 2 346 browsers and studied their stability over more than 2 years. We showed that, despite frequent changes in the fingerprints, a significant fraction of browsers can be tracked over a long period of time. Second, we designed a test suite to evaluate fingerprinting countermeasures. We applied our test suite to 7 countermeasures, some of them claiming to generate consistent fingerprints, and show that all of them can be identified, which can make their users more identifiable. Third, we explored the use of browser fingerprinting for crawler detection. We measured its use in the wild, as well as the main detection techniques. Since fingerprints are collected on the client-side, we also evaluated its resilience against an adversarial crawler developer that tries to modify its crawler fingerprints to bypass security checks.

These results have been obtained in the context of the PhD thesis of Antoine Vastel [14] defended in October 2019.

7.2. Test amplification

With respect to self-healing, we proposed a new algorithm for test amplification. Test amplification consists of exploiting the knowledge of test methods, in which developers embed input data and expected properties, in order to enhance these tests [22].

We proposed a new approach based on test inputs transformation and assertions generation to amplify test suites, and implemented this approach in the DSpot software tool that we created [21]. By evaluating DSpot on open-source projects from GitHub, we showed that we improve the mutation score of test suites. These improvements have been proposed to developers through pull requests: their feedbacks show that they value the output of DSpot by accepting to integrate amplified test methods into their test suite. This proves that DSpot can improve the quality of the test suite of real projects. We also showed that DSpot can generate amplified test methods that specify behavioral changes, and can generate amplified test methods to improve the ability to detect potential regressions.

These results have been obtained in the context of the STAMP H2020 project and in the context of the PhD thesis of Benjamin Danglot [11] defended in November 2019.

7.3. Understanding mobile-specific code smells

With respect to self-healing, we obtained new results in the domain of code smells for mobile software systems. Code smells are well-known concepts in software engineering. They refer to bad design and development practices commonly observed in software systems.

We obtained three new results that contribute to a better understanding of mobile code smells. First, we studied the expansion of code smells in different mobile platforms. Then, we conducted a large-scale study to analyze the change history of mobile apps and discern the factors that favor the introduction and survival of code smells. To consolidate these studies, we also performed a user study to investigate developers' perception of code smells and the adequacy of static analyzers as a solution for coping with them. Finally, we performed a qualitative study to question the established foundation about the definition and detection of mobile code smells. The results of these studies revealed important research findings. Notably, we showed that pragmatism, prioritization, and individual attitudes are not relevant factors for the accrual of mobile code smells. The problem is rather caused by ignorance and oversight, which are prevalent among mobile developers. Furthermore, we highlighted several flaws in the code smell definitions that are currently adopted by the research community. These results allowed us to elaborate some recommendations for researchers and tool makers willing to design detection and refactoring tools for mobile code smells [33], [34]. On top of that, our results opened perspectives for research works about the identification of mobile code smells and development practices in general.

These results have been obtained in the context of the PhD thesis of Sarra Habchi [12] defended in December 2019.

7.4. Towards privacy-sensitive mobile crowdsourcing

We obtained new results in the domain of data privacy for crowdsourced data.

We proposed an anonymous data collection library for mobile apps, a software library that improves the user's privacy without compromising the overall quality of the crowdsourced dataset. In particular, we proposed a decentralized approach, named FOUGERE, to convey data samples from user devices using peer-to-peer (P2P) communications to third-party servers, thus introducing an a priori data anonymization process that is resilient to location-based attacks. To validate the approach, we proposed a testing framework to test this P2P communication library, named PeerFleet. Beyond the identification of P2P-related errors, PeerFleet also helps to tune the discovery protocol settings to optimize the deployment of P2P apps. We validated FOUGERE using 500 emulated devices that replay a mobility dataset and use FOUGERE to collect location data. We evaluated the overhead, the privacy and the utility of FOUGERE. We showed that FOUGERE defeats the state-of-the-art location-based privacy attacks with little impact on the quality of the collected data [38], [5].

These results have been obtained in the context of the PhD thesis of Lakhdar Meftah [13] defended in December 2019.

STACK Project-Team

7. New Results

7.1. Resource Management

Participants: Mohamed Abderrahim, Adwait Jitendra Bauskar, Emile Cadorel, Hélène Coullon, Jad Darrous, David Espinel, Shadi Ibrahim, Thomas Lambert, Adrien Lebre, Jean-Marc Menaud, Alexandre Van Kempen.

In 2019, we achieved several contributions regarding the management of resources and data of cloud infrastructures, especially in a geo-distributed context (*e.g.*, Fog and Edge computing).

The first contributions are related to improvements of low-level building blocks. The following ones deal with geo-distributed considerations. Finally the last ones are related to capacity and placement strategies of distributed applications and scientific workflows.

In [15], we discuss how to improve I/O fairness and SSDs' utilization through the introduction of a NCQ-aware I/O scheduling scheme, NASS. The basic idea of NASS is to elaborately control the request dispatch of workloads to relieve NCQ conflict and improve NCQ utilization at the same time. To do so, NASS builds an evaluation model to quantify important features of the workload. In particular, the model first finds aggressive workloads, which cause NCQ conflict, based on the request size and the number of requests of the workloads. Second, it evaluates merging tendency of each workload, which may affect the bandwidth and cause NCQ conflict indirectly, based on request merging history. Third, the model identifies workloads with deceptive idleness, which cause low NCQ utilization, based on historical requests in I/O scheduler. Then, based on the model, NASS sets the request dispatch of each workload to guarantee fairness and improve device utilization: (1) NASS limits aggressive workloads to relieve NCQ conflict; (2) it adjusts merging of sequential workloads to improve bandwidth of the workloads while relieving NCQ conflict; and (3) it restricts request dispatch of I/O scheduler, rather than stopping request dispatch to improve NCQ utilization. We integrate NASS into four state-of-the-art I/O schedulers including CFQ, BFQ, FlashFQ, and FIOPS. The experimental results show that with NASS, I/O schedulers can achieve 11-23% better fairness and at the same time improve device utilization by 9-29%.

In [16], [28], we address the challenge related to the boot duration of virtual machines and containers in high consolidated cloud scenarios.This time, which can last up to minutes, is critical as it defines how an application can react w.r.t. demands' fluctuations (horizontal elasticity). Our contribution is the YOLO proposal (You Only Load Once). YOLO reduces the number of I/O operations generated during a boot process by relying on a boot image abstraction, a subset of the VM/container image that contains data blocks necessary to complete the boot operation. Whenever a VM or a container is booted, YOLO intercepts all read accesses and serves them directly from the boot image, which has been locally stored on fast access storage devices (e.g., memory, SSD, etc.). In addition to YOLO, we show that another mechanism is required to ensure that files related to VM/container management systems remain in the cache of the host OS. Our results show that the use of these two techniques can speed up the boot duration 2–13 times for VMs and 2 times for containers. The benefit on containers is limited due to internal choices of the docker design. We underline that our proposal can be easily applied to other types of virtualization (e.g., Xen) and containerization because it does not require intrusive modifications on the virtualization/container management system nor the base image structure.

Complementary to the previous contribution and in an attempt to demonstrate the importance of container image placement across edge servers, we propose and evaluate through simulation two novel container image placement algorithms based on k-Center optimization in [14]. In particular, we introduce a formal model to tackle down the problem of reducing the maximum retrieval time of container images, which we denote as MaxImageRetrievalTime. Based on the model, we propose KCBP and KCBP-WC, two placement algorithms which target reducing the maximum retrieval time of container images from any edge server. While KCBP is based on a k-Center solver (i.e., placing k facilities on a set of nodes to minimize the distance from any node to the closet facility) which is applied on each layer and its replicas (taking into account the storage capacities of the nodes), KCBP-WC uses the same principle but it tries to avoid simultaneous downloads from the same node. More precisely, if two layers are part of the same image, then they cannot be placed on the same nodes. We have implemented our proposed algorithms alongside two other state-of-the-art placement algorithms (i.e., Best-Fit and Random) in a simulator written in Python. Simulation results show that the proposed algorithms can outperform state-of- the-art algorithms by a factor of 1.1x to 4x depending on the characteristics of the networks.

In [13], we conduct experiments to thoroughly understand the performance of data-intensive applications under replication and EC. We use representative benchmarks on the Grid'5000 testbed to evaluate how analytic workloads, data persistency, failures, the back-end storage devices, and the network configuration impact their performances. While some of our results follow our intuition, others were unexpected. For example, disk and network contentions caused by chunks distribution and the unawareness of their functionalities are the main factor affecting the performance of data-intensive applications under EC, not data locality. An important outcome of our study is that it illustrates in practice the potential benefits of using EC in data-intensive clusters, not only in reducing the storage cost – which is becoming more critical with the wide adoption of high-speed storage devices and the explosion of generated and to be processed data – but also in improving the performance of data-intensive applications. We extended our work to Fog infrastructures in [31]. In particular, we empirically demonstrate the impact of network heterogeneity on the execution time of MR applications when running in the Fog.

In [5], we propose a first approach to deal with the data location challenges in geo-distribtued object stores. Existing solutions, relying on a distributed hash table to locate the data, are not efficient because location record may be placed far away from the object replicas. In this work, we propose to use a tree-based approach to locate the data, inspired by the Domain Name System (DNS) protocol. In our protocol, servers look for the location of an object by requesting successively their ancestors in a tree built with a modified version of the Dijkstra's algorithm applied to the physical topology. Location records are replicated close to the object replicas to limit the network traffic when requesting an object. We evaluate our approach on the Grid'5000 testbed using micro experiments with simple network topologies and a macro experiment using the topology of the French National Research and Education Network (RENATER). In this macro benchmark, we show that the time to locate an object in our approach is less than 15 ms on average which is around 20% shorter than using a traditional Distributed Hash Table (DHT).

In [20], we present the design, implementation, and evaluation of F-Storm, an FPGA-accelerated and general-purpose distributed stream processing system in the Edge. By analyzing current efforts to enable stream data processing in the Edge and to exploit FPGAs for data-intensive applications, we derive the key design aspects of F-Storm. Specifically, F-Storm is designed to: (1) provide a light-weight integration of FPGA with a DSP system in Edge servers, (2) make full use of FPGA resources when assigning tasks, (3) relieve the high overhead when transferring data between Java Virtual Machine (JVM) and FPGAs, and importantly (4) provide programming interface for users that enable them to leverage FPGA accelerators easily while developing their stream data applications. We have implemented F-Storm based on Storm. Evaluation results show that F-Storm reduces the latency by 36% and 75% for matrix multiplication and grep application compared to Storm. Furthermore, F-Storm obtains 1.4x, 2.1x, and 3.4x throughput improvement for matrix multiplication, grep application, and vector addition, respectively.

In [30], we discuss the main challengres related to the design and development of inter-site services for operating a massively distributed Cloud-Edge architecture deployed in different locations of the Internet backbone (i.e, network point of presences). More precisely, we discuss challenges related to the establishment of connectivity among several virtual infrastructure managers in charge of operating each site. Our goal is to initiate the discussion about the research directions on this field providing some interesting points to promote future work.

In [7], we focus on how to reduce the costly cross-rack data transferring in MapReduce systems. We observe that with high Map locality, the network is mainly saturated in Shuffling but relatively free in the Map phase. A little sacrifice in Map locality may greatly accelerate Shuffling. Based on this, we propose a novel scheme called Shadow for Shuffle-constrained general applications, which strikes a trade-off between Map locality and

Shuffling load balance. Specifically, Shadow iteratively chooses an original Map task from the most heavily loaded rack and creates a duplicated task for it on the most lightly loaded rack. During processing, Shadow makes a choice between an original task and its replica by efficiently pre-estimating the job execution time. We conduct extensive experiments to evaluate the Shadow design. Results show that Shadow greatly reduces the cross-rack skewness by 36.6% and the job execution time by 26% compared to existing schemes.

In [6], we consider a complete framework for straggler detection and mitigation. We start with a set of metrics that can be used to characterize and detect stragglers including 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 efficiencies. For example, our results indicate that the default Hadoop straggler detector could be made more effective. In certain case, Precision is low and only 55% of those detected are actual stragglers and the Recall, i.e., percent of actual detected stragglers, 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 99% and a Recall of 29%. This increase in Precision can be translated to achieve lower execution time and energy consumption, and thus higher performance and energy efficiency; compared to the default Hadoop mechanism, the energy consumption is reduced by almost 31%. 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 MapReduce systems.

In [21], we provide a general solution for workflow performance optimizations considering system variations. Specifically, we model system variations as time-dependent random variables and take their probability distributions as optimization input. Despite its effectiveness, this solution involves heavy computation overhead. Thus, we propose three pruning techniques to simplify workflow structure and reduce the probability evaluation overhead. We implement our techniques in a runtime library, which allows users to incorporate efficient probabilistic optimization into existing resource provisioning methods. Experiments show that probabilistic solutions can improve the performance by 51% compared to state-of-the-art static solutions while guaranteeing budget constraint, and our pruning techniques can greatly reduce the overhead of probabilistic optimization.

In [11], we propose a new strategy to schedule heteregeneous scientific workflows while minimizing the energy consumption of the cloud provider by introducing a deadline sensitive algorithm. Scheduling workflows in a cloud environment is a difficult optimization problem as capacity constraints must be fulfilled additionally to dependencies constraints between tasks of the workflows. Usually, work around the scheduling of scientific workflows focuses on public clouds where infrastructure management is an unknown black box. Thus, many works offer scheduling algorithms designed to select the best set of virtual machines over time, so that the cost to the end user is minimized. This paper presents the new v-HEFT-*deadline* algorithm that takes into account users deadlines to minimize the number of machines used by the cloud provider. The results show the real benefits of using our algorithm for reducing the energy consumption of the cloud provider.

In [9],we investigate how a monitoring service for Edge infrastructures should be designed in order to mitigate as much as possible its footprint in terms of used resources. Monitoring functions tend to become compute-, storage-and network-intensive, in particular because they will be used by a large part of applications that rely on real-time data. To reduce as much as possible the footprint of the whole monitoring service, we propose to mutualize identical processing functions among different tenants while ensuring their quality-of-service (QoS) expectations. We formalize our approach as a constraint satisfaction problem and show through micro-benchmarks its relevance to mitigate compute and network footprints.

In [22], we propose a generalization of the previous work. More precisely, weinvestigates whether the use of Constraint Programming (CP) could enable the development of a generic and easy-to-upgrade placement service for Fog/Edge Computing infrastructures. Our contribution is a new formulation of the placement problem, an implementation of this model leveraging Choco-solver and an evaluation of its scalability in comparison to recent placement algorithms. To the best of our knowledge, our study is the first one to evaluate the relevance of CP approaches in comparison to heuristic ones in this context. CP interleaves inference and systematic exploration to search for solutions, letting users on what matters: the problem description. Thus, our service placement model not only can be easily enhanced (deployment constraints/objectives) but also shows a competitive tradeoff between resolution times and solutions quality.

In [27], we present the first building blocks of a simulator to investigate placement challenges in Edge infrastructures. Efficiently scheduling computational jobs with data-sets dependencies is one of the most important challenges of fog/edge computing infrastructures. Although several strategies have been proposed, they have been evaluated through ad-hoc simulator extensions that are, when available, usually not maintained. This is a critical problem because it prevents researchers to-easily-conduct fair evaluations to compare each proposal. We propose to address this limitation throught the design and development of a common simulator. More precisely, in this research report, we describe an ongoing project involving academics and a high-tech company that aims at delivering a dedicated tool to evaluate scheduling policies in edge computing infrastructures. This tool enables the community to simulate various policies and to easily customize researchers/engineers' use-cases, adding new functionalities if needed. The implementation has been built upon the Batsim/SimGrid toolkit, which has been designed to evaluate batch scheduling strategies in various distributed infrastructures. Although the complete validation of the simulation toolkit is still ongoing, we demonstrate its relevance by studying different scheduling strategies on top of a simulated version of the Qarnot Computing platform, a production edge infrastructure based on smart heaters.

In [8], we propose an efficient graph partitioning method named Geo-Cut, which takes both the cost and performance objectives into consideration for large graph processing in geo-distributed DCs.Geo-Cut adopts two optimization stages. First, we propose a cost-aware streaming heuristic and utilize the one-pass streaming graph partitioning method to quickly assign edges to different DCs while minimizing inter-DC data communication cost. Second, we propose two partition refinement heuristics which identify the performance bottlenecks of geo-distributed graph processing and refine the partitioning result obtained in the first stage to reduce the inter-DC data transfer time while satisfying the budget constraint. Geo-Cut can be also applied to partition dynamic graphs thanks to its lightweight runtime overhead. We evaluate the effectiveness and efficiency of Geo-Cut using real-world graphs with both real geo-distributed DCs and simulations. Evaluation results show that Geo-Cut can reduce the inter-DC data transfer time by up to 79% (42% as the median) and reduce the monetary cost by up to 75% (26% as the median) compared to state-of-the-art graph partitioning methods with a low overhead.

7.2. Programming Support

Participants: Maverick Chardet, Hélène Coullon, Thomas Ledoux, Jacques Noyé, Dimitri Pertin, Simon Robillard, Hamza Sahli, Charlène Servantie.

Our contributions regarding programming support are divided in two topics. First, we focused on one specific challenge related to distributed software deployment: distributed software commissioning. We have proposed a useful approach for introducing model checking to help system operators design their parallel distributed software commissioning. Then, we focused on Fog formalization and we have proposed a fully graphical process algebraic formalism to design a Fog system.

In [12], MADA, a deployment approach to facilitate the design of efficient and safe distributed software commissioning is presented. MADA is built on top of the Madeus formal model that focuses on the efficient execution of installation procedures. Madeus puts forward more parallelism than other commissioning models, which implies a greater complexity and a greater propensity for errors. MADA provides a new specific language on top of Madeus that allows the developer to easily define the properties that should be ensured during the commissioning process. Then, MADA automatically translates the description to a time Petri net and a set of TCTL formulae. MADA is evaluated on the OpenStack commissioning.

About Fog formalization, we present a novel formal model defining spatial and structural aspects of Fog-based systems using Bigraphical Reactive Systems, a fully graphical process algebraic formalism [17]. The model is extended with reaction rules to represent the dynamic behavior of Fog systems in terms of self-adaptation.

The notion of bigraph patterns is used in conjunction with boolean and temporal operators to encode spatio-temporal properties inherent to Fog systems and applications. The feasibility of the modelling approach is demonstrated via a motivating case study and various self-adaptation scenarios.

Overall, the number of contributions we made this year on the programming support topic is less significative than the previous one. However, we would like to underline that it does not reflect the recent efforts we put. In particular, the team has strongly developed the field of dynamic reconfiguration of distributed software systems and expects to get important results during 2020.

7.3. Energy-aware computing

Participants: Emile Cadorel, Hélène Coullon, Adrien Lebre, Thomas Ledoux, Jean-Marc Menaud, Jonathan Pastor, Dimitri Saingre, Yewan Wang.

Energy consumption is one of the major challenges of modern datacenters and supercomputers. Our works in Energy-aware computing can be categorized into two subdomains: Software level (SaaS, PaaS) and Infrastructure level (IaaS). At Software level, we worked on the general Cloud applications architectures and more recently on BlockChain-based solutions. At Infrastructure level, we worked this year on two directions: (i) investigating the thermal aspects in datacenters, and (ii) analyzing the energy footprint of geo-distributed plateforms.

In [11], the scheduling of heterogeneous scientific workflows while minimizing the energy consumption of the cloud provider is tackled by introducing a deadline sensitive algorithm. Scheduling workflows in a cloud environment is a difficult optimization problem as capacity constraints must be fulfilled additionally to dependencies constraints between tasks of the workflows. Usually, work around the scheduling of scientific workflows focuses on public clouds where infrastructure management is an unknown black box. Thus, many works offer scheduling algorithms designed to select the best set of virtual machines over time, so that the cost to the end user is minimized. This paper presents the new v-HEFT-*deadline* algorithm that takes into account users deadlines to minimize the number of machines used by the cloud provider. The results show the real benefits of using our algorithm for reducing the energy consumption of the cloud provider.

In [25], over the last year, both academic and industry have increase their work on blockchain technologies. Despite the potential of blockchain technologies in many areas, several obstacles are slowing down their development. In addition to the legal and social obstacles, technical limitations now prevent them from imposing themselves as a real alternative to centralised services. For example, several problems dealing with the scalability or the energy cost have been identified. That's why, a significant part of this research is focused on improving the performances (latency, throughput, energy footprint, etc.) of such systems. Unfortunately, Those projects are often evaluated with ad hoc tools and experimental environment, preventing reproducibility and easy comparison of new contribution to the state of the art. As a result, we notice a clear lack of tooling concerning the benchmarking of blockchain technologies. To the best of our knowledge only a few tools address such issues. Those tools often relies on the load generation aspect and omit some other important aspect of benchmark experiments such as reproducibility and the network emulation. We introduce BCTMark, a general framework for benchmarking blockchain technologies in an emulated environment in a reproductible way.

In [18], we present a deep evaluation about the power models based on CPU utilization. The influence of inlet temperature on models has been especially discussed. According to the analysis, one regression formula by using CPU utilization as the only indicator is not adequate for building reliable power models. First of all, Workloads have different behaviors by using CPU and other hardware resources in server platforms. Therefore, power is observed to have high dispersion for a fixed CPU utilization, especially at full workload. At the same time, we also find that, power is well proportional to CPU utilization within the execution of one single workload. Hence, applying workload classifications could be an effective way to improve model accuracy. Moreover, inlet temperature can cause surprising influence on model accuracy. The model reliability can be questioned without including inlet temperature data. In a use case, after including inlet temperature data, we have greatly improved the precision of model outputs while stressing server under three different ambient temperatures.

In [18], our physical experiments have shown that even under the same conditions, identical processors consume different amount of energy to complete the same task. While this manufacturing variability has been observed and studied before, there is lack of evidence supporting the hypotheses due to limited sampling data, especially from the thermal characteristics. In this article, we compare the power consumption among identical processors for two Intel processors series with the same TDP (Thermal Design Power) but from different generations. The observed power variation of the processors in newer generation is much greater than the older one. Then, we propose our hypotheses for the underlying causes and validate them under precisely controlled environmental conditions. The experimental results show that, with the increase of transistor densities, difference of thermal characteristics becomes larger among processors, which has non-negligible contribution to the variation of power consumption for modern processors. This observation reminds us of re-calibrating the precision of the current energy predictive models. The manufacturing variability has to be considered when building energy predictive models for homogeneous clusters.

In [3], we propose a model and a first implementation of a simulator in order to compare the energy footprint of different cloud architectures (single sites vs fully decentrlaized). Despite the growing popularity of Fog/Edge architectures, their energy consumption has not been well investigated yet. To move forward on such a critical question, we first introduce a taxonomy of different Cloud-related architectures. From this taxonomy, we then present an energy model to evaluate their consumption. Unlike previous proposals, our model comprises the full energy consumption of the computing facilities, including cooling systems, and the energy consumption of network devices linking end users to Cloud resources. Finally, we instantiate our model on different Cloud-related architectures, ranging from fully centralized to completely distributed ones, and compare their energy consumption. The results validates that a completely distributed architecture, because of not using intra-data center network and large-size cooling systems, consumes less energy than fully centralized and partly distributed architectures respectively. To the best of our knowledge, our work is the first one to propose a model that enables researchers to analyze and compare energy consumption of different Cloud-related architectures.

7.4. Security and Privacy

Participants: Mohammad-Mahdi Bazm, Fatima Zahra Boujdad, Wilmer Edicson Garzon Alfonso, Jean-Marc Menaud, Sirine Sayadi, Mario Südholt.

This year the team has provided two major contributions on security and privacy challenges in distributed systems. First, we have extended our model for secure and privacy-aware biomedical analyses, as well as started to explore the impact of the big-data analyses in this context. Second, we have contributed mitigation methods for Cloud-based side-channel attacks.

In [24], we have developed a methodology for the development of secure and privacy-aware biomedical analyses we motivate the need for real distributed biomedical analyses in the context of several ongoing projects, including the I-CAN project that involves 34 French hospitals and affiliated research groups. We present a set of distributed architectures for such analyses that we have derived from discussions with different medical research groups and a study of related work. These architectures allow for scalability, security/privacy and reproducibility properties to be taken into account. A predefined set of architectures allows medecins and biomedical engineers to define high-level distributed architectures for biomedical analyses that ensure strong security and constraints on private data. Architectures from this set can then be implemented with ease because of detailed, also predefined, detailed implementation templates. Finally, we illustrate how these architectures can serve as the basis of a development method for biomedical distributed analyses.

In [10] and [23], we presented a new taxonomy for container security with a particular focus on data transmitted through the virtualization boundary. Containerization is a lightweight virtualization technique reducing virtualization overhead and deployment latency compared to full VM; its popularity is quickly increasing. However, due to kernel sharing, containers provide less isolation than full VM. Thus, a compromised container may break out of its isolated context and gain root access to the host server. This is a huge concern, especially in multi-tenant cloud environments where we can find running on a single server containers serving very different purposes, such as banking microservices, compute nodes or honeypots. Thus, containers with specific security needs should be able to tune their own security level. Because OS-level defense approaches inherited from time-sharing OS generally requires administrator rights and aim to protect the entire system, they are not fully suitable to protect usermode containers. Research recently made several contributions to deliver enhanced security to containers from host OS level to (partially) solve these challenges. In this survey, we propose a new taxonomy on container defense at the infrastructure level with a particular focus on the virtualization boundary, where interactions between kernel and containers take place. We then classify the most promising defense frameworks into these categories.

Finally, we have leveraged an approach based on Moving Target Defense (MTD) theory to interrupt a cache-based side-channel attack between two Linux containers in the context of the Mohammad Mahdi's PhD thesis [1]. MTD allows us to make the configuration of system more dynamic and consequently more harder to attack by an adversary, by using shuffling at different level of system and cloud. Our approach does not need to carrying modification neither into the guest OS or the hypervisor. Experimental results show that our approach imposes very low performance overhead. We have also provided a survey on the isolation challenge and on the cache-based side-channel attacks in cloud computing infrastructures. We have developed different approaches to detect/mitigate cross-VM/cross-containers cache-based side-channel attacks. Regarding the detection of cache-based side-channel attacks, we have enabled their detection by leveraging Hardware performance Counters (HPCs) and Intel Cache Monitoring Technology (CMT) with anomaly detection approaches to identify a malicious virtual machine or a Linux container. Our experimental results show a high detection rate.

WHISPER Project-Team

7. New Results

7.1. Software engineering for infrastructure software

Data races are often hard to detect in device drivers, due to the non-determinism of concurrent execution. With colleagues from Tsinghua University, we have addressed this issue using dynamic analysis. According to our study of Linux driver patches that fix data races, more than 38% of patches involve a pattern that we call inconsistent lock protection. Specifically, if a variable is accessed within two concurrently executed functions, the sets of locks held aroundeach access are disjoint, at least one of the locksets is non-empty, and at least one of the involved accesses is a write, then a datarace may occur. In a paper published at SANER 2019 [17], we present a runtime analysis approach, named DILP, to detect data races caused by inconsistent lock protection in device drivers. By monitoring driver execution, DILP collects the information about runtime variable accesses and executed functions. Then after driver execution, DILP analyzes the collected information to detect and report data races caused by inconsistent lock protection. We evaluate DILP on 12 device drivers in Linux 4.16.9, and find 25 real data races.

For waiting, the Linux kernel offers both sleep-able and non-sleep operations. However, only non-sleep operations can be used in atomic context. Detecting the possibility of execution in atomic context requires a complete inter-procedural flow analysis, often involving function pointers. Developers may thus conservatively use non-sleep operations even outside of atomic context, which may damage system performance, as such operations unproductively monopolize the CPU. Until now, no systematic approach has been proposed to detect such conservative non-sleep (CNS) defects. In a paper published at ASPLOS 2019 [14] with colleagues from Tsinghua University, we propose a practical static approach, named DCNS, to automatically detect conservative non-sleep defects in the Linux kernel. DCNS uses a summary-based analysis to effectively identify the code in atomic context and a novel file-connection-based alias analysis to correctly identify the set of functions referenced by a function pointer. We evaluate DCNS on Linux 4.16, and in total find 1629 defects. We manually check 943 defects whose call paths are not so difficult to follow, and find that 890 are real. We have randomly selected 300 of the real defects and sent them to kernel developers, and 251 have been confirmed.

In Linux device drivers, use-after-free (UAF) bugs can cause system crashes and serious security problems. We have addressed this issue in work with colleagues at Tsinghua University. According to our study of Linux kernel commits, 42% of the driver commits fixing use-after-free bugs involve driver concurrency. We refer to these use-after-free bugs as concurrency use-after-free bugs. Due to the non-determinism of concurrent execution, concurrency use-after-free bugs are often more difficult to reproduce and detect than sequential use-after-free bugs. In a paper published at USENIX ATC 2019 [13], we propose a practical static analysis approach named DCUAF, to effectively detect concurrency use-after-free bugs in Linux device drivers. DCUAF combines a local analysis analyzing the source code of each driver with a global analysis statistically analyzing the local results of all drivers, forming a local-global analysis, to extract the pairs of driver interface functions that may be concurrently executed. Then, with these pairs, DCUAF performs a summary-based lockset analysis to detect concurrency use-after-free bugs. We have evaluated DCUAF on the driver code of Linux 4.19, and found 640 real concurrency use-after-free bugs. We have randomly selected 130 of the real bugs and reported them to Linux kernel developers, and 95 have been confirmed.

Linux kernel stable versions serve the needs of users who value stability of the kernel over new features. The quality of such stableversions depends on the initiative of kernel developers and maintainers to propagate bug fixing patches to the stable versions. Thus, it is desirable to consider to what extent this process can be automated. A previous approach relies on words from commit messages and a small set of manually constructed code features. This approach, however, shows only moderate accuracy. In a tool paper published ICSE 2019 [11], in the context of the ANR-NRF ITrans project with colleagues from Singapore Management

University, paper, we investigate whether deep learning can provide a more accurate solution. We propose PatchNet, a hierarchical deep learning-based approach capable of automatically extracting features from commit messages and commit code and usingthem to identify stable patches. PatchNet contains a deep hierarchical structure that mirrors the hierarchical and sequential structure of commit code, making it distinctive from the existing deep learning models on source code. Experiments on 82,403 recent Linux patches confirm the superiority of PatchNet against various state-of-the-art baselines, including the one recently-adopted by Linux kernel maintainers.

Developing software often requires code changes that are widespread and applied to multiple locations. Previously, the Whisper team has addressed this problem with the tool Coccinelle. In a recent experience paper, published at ECOOP 2019 [21], in the context of the ANR-NRF ITrans project with colleagues from Singapore Management University, we have considered the benefits of extending Coccinelle to Java code. There are tools for Java that allow developers to specify patterns for program matching and source-to-source transformation. However, to our knowledge, none allows for transforming code based on its control-flow context. We prototype Coccinelle4J, an extension to Coccinelle, which is a program transformation tool designed for widespread changes in C code, in order to work on Java source code. We adapt Coccinelle to be able to apply scripts written in the Semantic Patch Language (SmPL), a language provided by Coccinelle, to Java source files. As a case study, we demonstrate the utility of Coccinelle4J with the task of API migration. We show 6 semantic patches to migrate from deprecated Android API methods on several open source Android projects. We describe how SmPL can be used to express several API migrations and justify several of our design decisions. This paper was accompanied by a tool demo.

A challenge in designing cooperative distributed systems is to develop feasible and cost-effective mechanisms to foster cooperation among selfish nodes, i.e., nodes that strategically deviate from the intended specification to increase their individual utility. Finding a satisfactory solution to this challenge may be complicated by the intrinsic characteristics of each system, as well as by the particular objectives set by the system designer. In a previous work we addressed this challenge by proposing RACOON, a general and semi-automatic framework for designing selfishness-resilient cooperative systems. RACOON relies on classical game theory and a custom built simulator to predict the impact of a fixed set of selfish behaviours on the designer's objectives. In a paper published in IEEE Transactions on Dependable and Secure Computing [12], we present RACOON++, which extends the previous framework with a declarative model for defining the utility function and the static behaviour of selfish nodes, along with a new model for reasoning on the dynamic interactions of nodes, based on evolutionary game theory. We illustrate the benefits of using RACOON++ by designing three cooperative systems: a peer-to-peer live streaming system, a load balancing protocol, and an anonymous communication system. Extensive experimental results using the state-of-the-art PeerSim simulator verify that the systems designed using RACOON++ achieve both selfishness-resilience and high performance.

7.2. Programming after the end of Moore's law

The end of Moore's law is a wake-up call that resonates across Computer Science at large. We are now firmly in an era of custom hardware design, as witnessed by the diversity of system-on-chip (SoC) and specialized processing units – such as graphics processing units (GPUs), tensor processing unit (TPUs) or programmable network adapters, to name but a few. This trend is justified by the existence of niche application domains (graphic processing, linear algebra, packet processing, etc.) that greatly benefit from specialized hardware. Faced with the imminent explosion of the number of niche applications and niche architectures, we are still grasping for a programming model that would accommodate this diversity.

The Usuba project is an exploratory effort in that direction. We chose a niche application domain (symmetric cryptographic algorithms), a specialized execution platform (Single Instruction Multiple Data, SIMD) processors and we set out to design a programming language faithfully describing our application domain as well as an optimizing compiler efficiently exploiting our target execution platform.

Indeed, cryptographic primitives are subject to diverging imperatives. Functional correctness and auditability pushes for the use of a high-level programming language. Performance and the threat of timing attacks push for directly programming in assembler to exploit (or avoid!) the micro-architectural features of a given machine.

In a paper published at PLDI 2019 [23], we have demonstrated that a suitable programming language could reconcile both views and actually improve on the state of the art of both.

USUBA is a dataflow programming language in which block ciphers become so simple as to be "obviously correct" and whose types document and enforce valid parallelization strategies at the granularity of individual bits. Its optimizing compiler, USUBAC, produces high-throughput, constant-time implementations performing on par with hand-tuned reference implementations. The cornerstone of our approach is a systematization and generalization of *bitslicing*, an implementation trick frequently used by cryptographers. We have shown that USUBA can produce code that executes between 5% slower to 22% faster than hand-tuned reference implementations while gracefully scaling across a wide range of architectures and automatically exploiting Single Instruction Multiple Data (SIMD) instructions whenever the cipher's structure allows it.

7.3. Support for multicore machines

The complexity of computer architectures has risen since the early years of the Linux kernel: Simultaneous Multi-Threading (SMT), multicore processing, and frequency scaling with complex algorithms such as Intel Turbo Boost have all become omnipresent. In order to keep up with hardware innovations, the Linux scheduler has been rewritten several times, and many hardware-related heuristics have been added. Despite this, we have shown in a PLOS paper [16] that a fundamental problem was never identified: the POSIX process creation model, i.e., fork/wait, can behave inefficiently on current multicore architectures due to frequency scaling. We investigate this issue through a simple case study: the compilation of the Linux kernel source tree. To do this, we have developed SchedLog, a low-overhead scheduler tracing tool, and SchedDisplay, a scriptable tool to graphically analyze SchedLog's traces efficiently. We implement two solutions to the problem at the scheduler level which improve the speed of compiling part of the Linux kernel by up to 26%, and the whole kernel by up to 10%.

In an Eurosys paper [15], we address the problem of efficiently virtualizing NUMA architectures. The major challenge comes from the fact that the hypervisor regularly reconfigures the placement of a virtual machine (VM) over the NUMA topology. However, neither guest operating systems (OSes) nor system runtime libraries (e.g., Hotspot) are designed to consider NUMA topology changes at runtime, leading end user applications to unpredictable performance. We present eXtended Para-Virtualization (XPV), a new principle to efficiently virtualize a NUMA architecture. XPV consists in revisiting the interface between the hypervisor and the guest OS, and between the guest OS and system runtime libraries (SRL) so that they can dynamically take into account NUMA topology changes. We introduce a methodology for systematically adapting legacy hypervisors, OSes, and SRLs. We have applied our approach with less than 2k line of codes in two legacy hypervisors (Xen and KVM), two legacy guest OSes (Linux and FreeBSD), and three legacy SRLs (Hotspot, TCMalloc, and jemalloc). The evaluation results showed that XPV outperforms all existing solutions by up to 304%.

Memory interferences may introduce important slowdowns in applications running on COTS multi-core processors. They are caused by concurrent accesses to shared hardware resources of the memory system. The induced delays are difficult to predict, making memory interferences a major obstacle to the adoption of COTS multi-core processors in real-time systems. In an RTSS paper[18], we propose an experimental characterization of applications' memory consumption to determine their sensitivity to memory interferences. Thanks to a new set of microbenchmarks, we show the lack of precision of a purely quantitative characterization. To improve accuracy, we define new metrics quantifying qualitative aspects of memory consumption and implement a profiling tool using the VALGRIND framework. In addition, our profiling tool produces high resolution profiles allowing us to clearly distinguish the various phases in applications' behavior. Using our microbenchmarks and our new characterization, we train a state-of-the-art regressor. The validation on applications from the MIBENCH and the PARSEC suites indicates significant gain in prediction accuracy compared to a purely quantitative characterization.

WIDE Project-Team

6. New Results

6.1. Recommender Systems

6.1.1. A Biclustering Approach to Recommender Systems

Participants: Florestan de Moor, Davide Frey.

Recommendation systems are a core component of many e-commerce industries and online services since they ease the discovery of relevant products. Because catalogs are huge, it is impossible for an individual to manually search for an item of interest, hence the need for some automatic filtering process. Many approaches exist, from content-based ones to collaborative filtering that include neighborhood and model-based techniques. Despite these intensive research activities, numerous challenges remain to be addressed, particularly under real-time settings or regarding privacy concerns, which motivates further work in this area. We focus on techniques that rely on biclustering, which consists in simultaneously building clusters over the two dimensions of a data matrix. Although it was little considered by the recommendation system community, it is a well-known technique in other domains such as genomics. In work [42] we present the different biclustering-based approaches that were explored. We then are the first to perform an extensive experimental evaluation to compare these approaches with one another, but also with the current state-of-the-art techniques from the recommender field. Existing evaluations are often restrained to a few algorithms and consider only a limited set of metrics. We then expose a few ideas to improve existing approaches and address the current challenges in the design of highly efficient recommendation algorithms, along with some preliminary results.

This work was done in collaboration with Antonio Mucherino (University of Rennes 1).

6.1.2. Unified and Scalable Incremental Recommenders with Consumed Item Packs

Participant: Erwan Le Merrer.

Recommenders personalize the web content by typically using collaborative filtering to relate users (or items) based on explicit feedback, e.g., ratings. The difficulty of collecting this feedback has recently motivated to consider implicit feedback (e.g., item consumption along with the corresponding time). In this work [39], we introduce the notion of consumed itempack (CIP) which enables to link users (or items) based on their implicit analogous consumption behavior. Our proposal is generic, and we show that it captures three novel implicit recommenders: a user-based (CIP-U), an item-based (CIP-I),and a word embedding-based (DEEPCIP), as well as a state-of-art technique using implicit feedback (FISM). We show that our recommenders handle incremental updates incorporating freshly consumed items. We demonstrate that all three recommenders provide a recommendation quality that is competitive with state-of-the-art ones, including one incorporating both explicit and implicit feedback

This work was done in collaboration with Rachid Guerraoui (EPFL), Rhicheek Patra (Oracle) and Jean-Ronan Vigouroux (Technicolor).

6.2. Systems for the Support of Privacy

6.2.1. Robust Privacy-Preserving Gossip Averaging

Participants: Amaury Bouchra-Pilet, Davide Frey, François Taïani.

This contribution aims to address the privacy risks inherent in decentralized systems by considering the emblematic problem of privacy-preserving decentralized averaging. In particular, we propose a novel gossip protocol that exchanges noise for several rounds before starting to exchange actual data. This makes it hard for an honest but curious attacker to know whether a user is transmitting noise or actual data. Our protocol and analysis do not assume a lock-step execution, and demonstrate improved resilience to colluding attackers. In a paper, publishing this work at SSS 2019 [26], we prove the correctness of this protocol as well as several privacy results. Finally, we provide simulation results about the efficiency of our averaging protocol.

6.2.2. A Collaborative Strategy for Mitigating Tracking through Browser Fingerprinting.

Participants: David Bromberg, Davide Frey, Alejandro Gomez-Boix.

Browser fingerprinting is a technique that collects information about the browser configuration and the environment in which it is running. This information is so diverse that it can partially or totally identify users online. Over time, several countermeasures have emerged to mitigate tracking through browser fingerprinting. However, these measures do not offer full coverage in terms of privacy protection, as some of them may introduce inconsistencies or unusual behaviors, making these users stand out from the rest.

In this work, we address these limitations by proposing a novel approach that minimizes both the identifiability of users and the required changes to browser configuration. To this end, we exploit clustering algorithms to identify the devices that are prone to share the same or similar fingerprints and to provide them with a new non-unique fingerprint. We then use this fingerprint to automatically assemble and run web browsers through virtualization within a docker container. Thus all the devices in the same cluster will end up running a web browser with an indistinguishable and consistent fingerprint.

We carried out this work in collaboration with Benoit Baudry from KTH Sweden and published our results at the 2019 Moving-Target Defense Workshop [30].

6.3. Distributed Algorithms

6.3.1. One for All and All for One:Scalable Consensus in a Hybrid Communication Model

Participant: Michel Raynal.

This work [34] addresses consensus in an asynchronous model where the processes are partitioned into clusters. Inside each cluster, processes can communicate through a shared memory, which favors efficiency. Moreover, any pair of processes can also communicate through a message-passing communication system, which favors scalability. In such a “hybrid communication” context, the work presents two simple binary consensus algorithms (one based on local coins,the other one based on a common coin). These algorithms are straightforward extensions of existing message-passing randomized round-based consensus algorithms. At each round, the processes of each cluster first agree on the same value (using an underlying shared memory consensus algorithm), and then use a message-passing algorithm to converge on the same decided value. The algorithms are such that, if all except one processes of a cluster crash, the surviving process acts as if all the processes of its cluster were alive (hence the motto “one for all and all for one”). As a consequence, the hybrid communication model allows us to obtain simple, efficient, and scalable fault-tolerant consensus algorithms. As an important side effect, according to the size of each cluster, consensus can be obtained even if a majority of processes crash.

This work was done in collaboration with Jiannong Cao (Polytechnic University, Hong Kong).

6.3.2. Optimal Memory-Anonymous Symmetric Deadlock-Free Mutual Exclusion

Participant: Michel Raynal.

The notion of an anonymous shared memory (recently introduced in PODC 2017) considers that processes use different names for the same memory location. Hence, there is permanent disagreement on the location names among processes. In this context, the PODC paper presented -among other results- a symmetric deadlock-free mutual exclusion (mutex) algorithm for two processes and a necessary condition on the size m of the anonymous memory for the existence of a symmetric deadlock-free mutex algorithm in an n-process system. This work [22] answers several open problems related to symmetric deadlock-free mutual exclusion in an n-process system where the processes communicate through m registers. It first presents two algorithms. The first considers that the registers are anonymous read/write atomic registers and works for any m greater than 1 and belonging to the set M(n). It thus shows that this condition on m is both necessary and sufficient. The second algorithm considers anonymous read/modify/write atomic registers. It assumes that $m \in M(n)$. These algorithms differ in their design principles and their costs (measured as the number of registers which must contain the identity of a process to allow it to enter the critical section). The work also shows that the condition

$m \in M(n)$ is necessary for deadlock-free mutex on top of anonymous read/modify/write atomic registers. It follows that, when m > 1, $m \in M(n)$ is a tight characterization of the size of the anonymous shared memory needed to solve deadlock-free mutex, be the anonymous registers read/write or read/modify/write.

This work was done in collaboration with Zahra Aghazadeh (University of Calgary), Damien Imbs (LIS, Université d'Aix-Marseille,CNRS, Université de Toulon), Gadi Taubenfeld (The Interdisciplinary Center of Herzliya) and Philipp Woelfel (University of Calgary).

6.3.3. Merkle Search Trees

Participants: Alex Auvolat, François Taïani.

Most recent CRDT (Conflict-free Replicated Data Type) techniques rely on a causal broadcast primitive to provide guarantees on the delivery of operation deltas. Such a primitive is unfortunately hard to implement efficiently in large open networks, whose membership is often difficult to track. As an alternative, we argue that pure state-based CRDTs can be efficiently implemented by encoding states as specialized Merkle trees, and that this approach is well suited to open networks where many nodes may join and leave. Indeed, Merkle trees enable efficient remote comparison and reconciliation of data sets, which can be used to implement the CRDT merge operator between two nodes without any prior information. This approach also does not require vector clock information, which would grow linearly with the number of participants.

At the core of our contribution [24] lies a new kind of Merkle tree, called Merkle Search Tree (MST), that implements a balanced search tree while maintaining key ordering. This latter property makes it particularly efficient in the case of updates on sets of sequential keys, a common occurrence in many applications. We use this new data structure to implement a distributed event store, and show its efficiency in very large systems with low rates of updates. In particular, we show that in some scenarios our approach is able to achieve both a 66% reduction of bandwidth cost over a vector-clock approach, as well as a 34% improvement in consistency level. We finally suggest other uses of our construction for distributed databases in open networks.

6.3.4. Dietcoin: Hardening Bitcoin Transaction Verification Process For Mobile Devices

Participants: Davide Frey, François Taïani.

Distributed ledgers are among the most replicated data repositories in the world. They offer data consistency, immutability, and auditability, based on the assumption that each participating node locally verifies their entire content. Although their content, currently extending up to a few hundred gigabytes, can be accommodated by dedicated commodity hard disks, downloading it, processing it, and storing it in general-purpose desktop and laptop computers can prove largely impractical. Even worse, this becomes a prohibitive restriction for smartphones, mobile devices, and resource-constrained IoT devices.

We thus proposed Dietcoin, a Bitcoin protocol extension that allows nodes to perform secure local verification of Bitcoin transactions with small bandwidth and storage requirements. We carried out an extensive evaluation of the features of Dietcoin that are important for today's cryptocurrency and smart-contract systems, but are missing in the current state-of-the-art. These include (i) allowing resource-constrained devices to verify the correctness of selected blocks locally without having to download the complete ledger; (ii) enabling devices to join a blockchain quickly yet securely, dropping bootstrap time from days down to a matter of seconds; (iii) providing a generic solution that can be applied to other distributed ledgers secured with Proof-of-Work. We showcased our results in a demo at VLDB 2019 [15], and we are currently preparing a full paper submission.

We carried out this work in collaboration with Pierre-Louis Roman, now at University of Lugano (Switzerland), as well as with Mark Makke from Vrije Universiteit, Amsterdam (the Netherlands), and Spyros Voulgaris from Athens University of Economics and Business (Greece).

6.3.5. Byzantine-Tolerant Set-Constrained Delivery Broadcast

Participants: Alex Auvolat, François Taïani, Michel Raynal.

Set-Constrained Delivery Broadcast (SCD-broadcast), recently introduced at ICDCN 2018, is a high-level communication abstraction that captures ordering properties not between individual messages but between sets of messages. More precisely, it allows processes to broadcast messages and deliver sets of messages, under the constraint that if a process delivers a set containing a message m before a set containing a message m', then no other process delivers first a set containing m' and later a set containing m. It has been shown that SCD-broadcast and read/write registers are computationally equivalent, and an algorithm implementing SCD-broadcast is known in the context of asynchronous message passing systems prone to crash failures.

We introduce a Byzantine-tolerant SCD-broadcast algorithm in [23], which we call BSCD-broadcast. Our proposed algorithm assumes an underlying basic Byzantine-tolerant reliable broadcast abstraction. We first introduce an intermediary communication primitive, Byzantine FIFO broadcast (BFIFO-broadcast), which we then use as a primitive in our final BSCD-broadcast algorithm. Unlike the original SCD-broadcast algorithm that is tolerant to up to $t < n/2$ crashing processes, and unlike the underlying Byzantine reliable broadcast primitive that is tolerant to up to $t < n/3$ Byzantine processes, our BSCD-broadcast algorithm is tolerant to up to $t < n/4$ Byzantine processes. As an illustration of the high abstraction power provided by the BSCD-broadcast primitive, we show that it can be used to implement a Byzantine-tolerant read/write snapshot object in an extremely simple way.

6.3.6. PnyxDB: a Lightweight Leaderless Democratic Byzantine Fault Tolerant Replicated Datastore

Participants: Loïck Bonniot, François Taïani.

Byzantine-Fault-Tolerant (BFT) systems are rapidly emerging as a viable technology for production-grade systems, notably in closed consortia deployments for financial and supply-chain applications. Unfortunately, most algorithms proposed so far to coordinate these systems suffer from substantial scalability issues, mainly due to the requirement of a single leader node. We observed that many application workloads offer little concurrency, and proposed PnyxDB, an eventually-consistent BFT replicated datastore that exhibits both high scalability and low latency. Our approach (proposed in [40]) is based on conditional endorsements, that allow nodes to specify the set of transactions that must *not* be committed for the endorsement to be valid.

Additionally, although most of prior art rely on internal voting or quorum mechanisms, these mechanisms are not exposed to applications as first-class primitives. As a result, individual nodes cannot implement application-defined policies without additional effort, costs, and complexity. This is problematic, as application-level voting capabilities are key to a number of emerging decentralized BFT applications involving independent participants who need to balance conflicting goals and shared interests. In addition to its high scalability, PnyxDB supports application-level voting by design. We provided a comparison against BFTS-MaRt and Tendermint, two competitors with different design aims, and demonstrated that our implementation speeds up commit latencies by a factor of 11, remaining below 5 seconds in a worldwide geodistributed deployment of 180 nodes.

PnyxDB's source code is freely available [0]. This work has also been done in collaboration with Christoph Neumann at InterDigital.

6.3.7. Vertex Coloring with Communication Constraints in Synchronous Broadcast Networks

Participants: Hicham Lakhlef, Michel Raynal, François Taïani.

In this work [17], we consider distributed vertex-coloring in broadcast/receive networks suffering from conflicts and collisions. (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 specifically, our work focuses on multi-channel networks, in which a process may either broadcast a message to its neighbors or receive a message from at most γ of them. The work first provides a new upper bound on the corresponding graph coloring problem (known as frugal coloring) in general graphs, proposes an exact bound for the problem in trees, and presents a deterministic, parallel, color-optimal, collision- and conflict-free distributed coloring algorithm for trees, and proves its correctness.

[0] https://github.com/technicolor-research/pnyxdb

6.3.8. *Efficient Randomized Test-and-Set Implementations*

Participant: George Giakkoupis.

In [16], we study randomized test-and-set (TAS) implementations from registers in the asynchronous shared memory model with n processes. We introduce the problem of *group election*, a natural variant of leader election, and propose a framework for the implementation of TAS objects from group election objects. We then present two group election algorithms, each yielding an efficient TAS implementation. The first implementation has expected maxstep complexity $O(\log^* k)$ in the location-oblivious adversary model, and the second has expected maxstep complexity $O(\log\log k)$ against any read/write-oblivious adversary, where $k \leq n$ is the contention. These algorithms improve the previous upper bound by Alistarh and Aspnes (2011) of $O(\log\log n)$ expected maxstep complexity in the oblivious adversary model.

We also propose a modification to a TAS algorithm by Alistarh, Attiya, Gilbert, Giurgiu, and Guerraoui (2010) for the strong adaptive adversary, which improves its space complexity from super-linear to linear, while maintaining its $O(\log n)$ expected maxstep complexity. We then describe how this algorithm can be combined with any randomized TAS algorithm that has expected maxstep complexity $T(n)$ in a weaker adversary model, so that the resulting algorithm has $O(\log n)$ expected maxstep complexity against any strong adaptive adversary and $O(T(n))$ in the weaker adversary model.

Finally, we prove that for any randomized 2-process TAS algorithm, there exists a schedule determined by an oblivious adversary such that with probability at least $1/4^t$ one of the processes needs at least t steps to finish its TAS operation. This complements a lower bound by Attiya and Censor-Hillel (2010) on a similar problem for $n \geq 3$ processes.

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

6.4. Machine Learning and Security

6.4.1. *Adversarial Frontier Stitching for Remote Neural Network Watermarking*

Participant: Erwan Le Merrer.

The state-of-the-art performance of deep learning models comes at a high cost for companies and institutions, due to the tedious data collection and the heavy processing requirements. Recently, Nagai et al. proposed to watermark convolutional neural networks for image classification, by embedding information into their weights. While this is a clear progress toward model protection, this technique solely allows for extracting the watermark from a network that one accesses locally and entirely. Instead, we aim at allowing the extraction of the watermark from a neural network (or any other machine learning model) that is operated remotely, and available through a service API. To this end, we propose in this work [18] to mark the model's action itself, tweaking slightly its decision frontiers so that a set of specific queries convey the desired information. In this work, we formally introduce the problem and propose a novel zero-bit watermarking algorithm that makes use of adversarial model examples. While limiting the loss of performance of the protected model, this algorithm allows subsequent extraction of the watermark using only few queries. We experimented the approach on three neural networks designed for image classification, in the context of the MNIST digit recognition task.

This work was done in collaboration with Gilles Trédan (LAAS/CRNS) and Patrick Pérez (Valéo AI).

6.4.2. *TamperNN: Efficient Tampering Detection of Deployed Neural Nets*

Participant: Erwan Le Merrer.

Neural networks are powering the deployment of embedded devices and Internet of Things. Applications range from personal assistants to critical ones such as self-driving cars. It has been shown recently that models obtained from neural nets can be trojaned ; an attacker can then trigger an arbitrary model behavior facing crafted inputs. This has a critical impact on the security and reliability of those deployed devices. In this work [33], we introduce novel algorithms to detect the tampering with deployed models, classifiers in particular. In the remote interaction setup we consider, the proposed strategy is to identify markers of the model input space that are likely to change class if the model is attacked, allowing a user to detect a possible tampering. This setup makes our proposal compatible with a wide range of scenarios, such as embedded models, or models exposed through prediction APIs. We experiment those tampering detection algorithms on the canonical MNIST dataset, over three different types of neural nets, and facing five different attacks (trojaning, quantization, fine-tuning, compression and watermarking). We then validate over five large models (VGG16, VGG19, ResNet, MobileNet, DenseNet) with a state of the art dataset (VGGFace2), and report results demonstrating the possibility of an efficient detection of model tampering.

This work was done in collaboration with Gilles Trédan (LAAS/CRNS).

6.4.3. MD-GAN: Multi-Discriminator Generative Adversarial Networks for Distributed Datasets

Participant: Erwan Le Merrer.

A recent technical breakthrough in the domain of machine learning is the discovery and the multiple applications of Generative Adversarial Networks (GANs). Those generative models are computationally demanding, as a GAN is composed of two deep neural networks, and because it trains on large datasets. A GAN is generally trained on a single server. In this work, we address the problem of distributing GANsso that they are able to train over datasets that are spread on multiple workers. In this work [31] MD-GAN is exposed as the first solution for this problem: we propose a novel learning procedure for GANs so that they fit this distributed setup. We then compare the performance of MD-GAN to an adapted version of Federated Learning to GANs, using the MNIST and CIFAR10 datasets.MD-GAN exhibits a reduction by a factor of two of the learning complexity on each worker node, while providing better performances than federated learning on both datasets. We finally discuss the practical implications of distributing GANs.

This work was done in collaboration with Bruno Sericola (Inria) and Corentin Hardy (Technicolor).

6.5. Network and Graph Algorithms

6.5.1. Multisource Rumor Spreading with Network Coding

Participants: David Bromberg, Quentin Dufour, Davide Frey.

The last decade has witnessed a rising interest in Gossip protocols in distributed systems. In particular, as soon as there is a need to disseminate events, they become a key functional building block due to their scalability, robustness and fault tolerance under high churn. However, Gossip protocols are known to be bandwidth intensive. A huge amount of algorithms has been studied to limit the number of exchanged messages using different combinations of push/pull approaches. In this work we revisited the state of the art by applying Random Linear Network Coding to further increase performance. In particular, the originality of our approach consists in combining sparse-vector encoding to send our network-coding coefficients and Lamport timestamps to split messages in generations in order to provide efficient gossiping. Our results demonstrate that we are able to drastically reduce bandwidth overhead and dissemination delay compared to the state of the art. We published our results at INFOCOM 2019 [27].

6.5.2. DiagNet: towards a generic, Internet-scale root cause analysis solution

Participants: Loïck Bonniot, François Taïani.

Internet content providers and network operators allocate significant resources to diagnose and troubleshoot problems encountered by end-users, such as service quality of experience degradations. Because the Internet is decentralized, the cause of such problems might lie anywhere between an end-user's device and the service datacenters. Further, the set of possible problems and causes cannot be known in advance, making it impossible to train a classifier with all combinations of faults, causes and locations. We explored how machine learning can be used for Internet-scale root cause analysis using measurements taken from end-user devices: our solution, DiagNet, is able to build generic models that (i) do not make any assumption on the underlying network topology, (ii) do not require to define the full set of possible causes during training, and (iii) can be quickly adapted to diagnose new services.

DiagNet adapts recent image analysis tactics for system and network metrics, collected from a large and dynamic set of landmark servers. In details, it applies non-overlapping convolutions and global pooling to extract generic information about the analyzed network. This genericness allows to build a general model, that can later be generalized to any Internet service with minimal effort. DiagNet leverages backpropagation attention mechanisms to extend the possible root causes to the set of available metrics, making the model fully extensible. We evaluated DiagNet on geodistributed mockup web services and automated users running in 6 AWS regions, and demonstrated promising root cause analysis capabilities. While this initial work is being reviewed, we are deploying DiagNet for real web services and users to evaluate its performance in a more realistic setup.

Christoph Neumann (InterDigital) actively participated in this work.

6.5.3. Application-aware adaptive partitioning for graph processing systems

Participant: Erwan Le Merrer.

Modern online applications value real-time queries over fresh data models. This is the case for graph-based applications, such as social networking or recommender systems,running on front-end servers in production. A core problem in graph processing systems is the efficient partitioning of the input graph over multiple workers. Recent advances over Bulk Synchronous Parallel processing systems (BSP) enabled computations over partitions on those workers, independently of global synchronization supersteps. A good objective partitioning makes the understanding of the load balancing and communication trade-off mandatory for performance improvement. This work [32] addresses this trade-off through the proposal of an optimization problem, that is to be solved continuously to avoid performance degradation over time. Our simulations show that the design of the software module we propose yields significant performance improvements over the BSP processing model.

This work was done in collaboration with Gilles Trédan (LAAS/CRNS).

6.5.4. How to Spread a Rumor: Call Your Neighbors or Take a Walk?

Participant: George Giakkoupis.

In [28], we study the problem of randomized information dissemination in networks. We compare the now standard push-pull protocol, with agent-based alternatives where information is disseminated by a collection of agents performing independent random walks. In the visit-exchange protocol, both nodes and agents store information, and each time an agent visits a node, the two exchange all the information they have. In the meet-exchange protocol, only the agents store information, and exchange their information with each agent they meet.

We consider the broadcast time of a single piece of information in an n-node graph for the above three protocols, assuming a linear number of agents that start from the stationary distribution. We observe that there are graphs on which the agent-based protocols are significantly faster than push-pull, and graphs where the converse is true. We attribute the good performance of agent-based algorithms to their inherently fair bandwidth utilization, and conclude that, in certain settings, agent-based information dissemination, separately or in combination with push-pull, can significantly improve the broadcast time.

The graphs considered above are highly non-regular. Our main technical result is that on any regular graph of at least logarithmic degree, push-pull and visit-exchange have the same asymptotic broadcast time. The proof uses a novel coupling argument which relates the random choices of vertices in push-pull with the random walks in visit-exchange. Further, we show that the broadcast time of meet-exchange is asymptotically at least as large as the other two's on all regular graphs, and strictly larger on some regular graphs.

As far as we know, this is the first systematic and thorough comparison of the running times of these very natural information dissemination protocols.

This work was done in collaboration with Frederik Mallmann-Trenn (MIT) and Hayk Saribekyan (University of Cambridge, UK).

ALPINES Project-Team

7. New Results

7.1. Adaptive Domain Decomposition Method for Saddle Point Problem

In [37], we introduce an adaptive domain decomposition (DD) method for solving saddle point problems defined as a block two by two matrix. The algorithm does not require any knowledge of the constrained space. We assume that all sub matrices are sparse and that the diagonal blocks are the sum of positive semi definite matrices. The latter assumption enables the design of adaptive coarse space for DD methods.

7.2. A Class of Efficient Locally Constructed Preconditioners Based on Coarse Spaces

In [24] we present a class of robust and fully algebraic two-level preconditioners for SPD matrices. We introduce the notion of algebraic local SPSD splitting of an SPD matrix and we give a characterization of this splitting. It helps construct *algebraically and locally* a class of efficient coarse subspaces which bound the spectral condition number of the preconditioned system by a number defined a priori. Some PDEs-dependant preconditioners correspond to a special case of the splitting. The examples of the algebraic coarse subspaces in this paper are not practical due to expensive construction. We propose an heuristic approximation that is not costly. Numerical experiments illustrate the efficiency of the proposed method.

7.3. A Multilevel Schwarz Preconditioner Based on a Hierarchy of Robust Coarse Spaces

In [32] we present a multilevel preconditioner for SPD matrices. Robust two-level additive Schwarz preconditioners guarantee a fast convergence of the Krylov method. To maintain the robustness each subdomain contributes a small number of vectors to construct a basis for the second level (the coarse space). As long as the dimension of the coarse space is reasonable i.e., direct solvers can be used efficiently, the two-level method scales well. However, the bottleneck arises when factoring the coarse space matrix becomes costly. Using an iterative Krylov method on the second level might be the right choice. Nevertheless, the condition number of the coarse space matrix is typically larger than the one of the first level. One of the difficulties of using two-level methods to solve the coarse problem is that the matrix does not arise from a PDE anymore. We introduce in this paper a practical method of applying a multilevel additive Schwarz preconditioner efficiently. This multilevel preconditioner is implemented in HPDDM and the code for reproducing the results from the paper is available here.

7.4. Inverse scattering problems without knowing the source term

The solution of inverse scattering problems always presupposed knowledge of the incident wave-field and require repeated computations of the forward problem, for which knowing the source term is crucial. In [26], we present a three-step strategy to solve inverse scattering problems when the time signature of the source is unknown. The proposed strategy combines three recent techniques: (i) wave splitting to retrieve the incident and the scattered wavefields, (ii) time-reversed absorbing conditions (TRAC) for redatuming the data inside the computational domain, (iii) adaptive eigenspace inversion (AEI) to solve the inverse problem. Numerical results illustrate step-by-step the feasibility of the proposed strategy.

7.5. Envelope following methods

One difficulty when solving problems in plasma physics is the behaviour at several scales in time and space of the solutions of equations. For example, central equations in this domain of application are highly oscillatory in time. The multiscale aspect makes the models difficult to tackle when we aim at avoiding a high computational cost. A solution to this problem is to solve the models by designing adapted numerical methods with a low computational cost and which are able to deal efficiently with rapid and slow scales in time. In this direction, we worked on envelope following methods, which have been efficiently applied in the community of oscillators in RF circuits. The method has (at least) two variants: in a first place, it is based on the concept of using extra variables to represent the changing rapid period and the cumulative effect of changing periods and then, use of Newton iterations allows to find these unknowns. In a second place, we adopt a similar strategy except that the rapid period is not an extra variable but a direct outcome of the numerical integration by the use of the Poincaré map. We implemented and tested both approaches for equations of interest in plasma physics and we observed that these methods didn't perform accurate results.

7.6. Domain decomposition preconditioning for high frequency wave propagation problems

The work about domain decomposition preconditioning for Maxwell equations has been published in [21]. It studies two-level preconditioners where the coarse space is based on the discretisation of the PDE on a coarse mesh. The PDE is discretised using finite-element methods of fixed, arbitrary order. The theoretical part of this work is the Maxwell analogue of a previous work for Helmholtz equation, and shows that for Maxwell problems with absorption, if the absorption is large enough and if the subdomain and coarse mesh diameters are chosen appropriately, then classical two-level overlapping Additive Schwarz Domain Decomposition preconditioning performs optimally – in the sense that GMRES converges in a wavenumber-independent number of iterations. The theory is also illustrated by various numerical experiments.

Ongoing studies are being conducted on recursive one-level optimized Schwarz methods for the high frequency Helmholtz and Maxwell equations. The method consists in solving the subdmain problems in a one-level optimized Schwarz preconditioner only approximately, using inner GMRES iterations preconditioned again by a one-level method, with smaller subdomains. The asymptotic behaviour and parallel scalability of the method are being investigated. Exhaustive numerical experiments are being conducted to compare the efficiency of this method with two-level preconditioners, including cavity problems and benchmarks in seismic imaging.

7.7. The boundary element method in FreeFEM

The BemTool and HTOOL libraries developed by the team, implementing respectively the Boundary Element Method and Hierarchical Matrices, have been interfaced with FreeFEM to allow FreeFEM users to use the Boundary Element Method (BEM) in their FreeFEM scripts. New additions to the Domain Specific Language (DSL) of FreeFEM allows the user to define and manipulate curved (1D) and surface (2D) meshes, as well as define and solve BEM variational problems in a high-level manner, similarly to FEM problems. The parallelization of the HTOOL library allows the user to assemble and solve their BEM problems in parallel in a transparent way.

Ongoing work consists in finalizing the BEM DSL to propose complete and documented features to the FreeFEM user in the next release, as well as investigating FEM-BEM coupling.

7.8. New Optimised Schwarz Method for dealing with cross-points

We consider a scalar wave propagation in harmonic regime modelled by Helmholtz equation with heterogeneous coefficients. Using the Multi-Trace Formalism (MTF), we propose a new variant of the Optimized Schwarz Method (OSM) that can accomodate the presence of cross-points in the subdomain partition. This leads to the derivation of a strongly coercive formulation of our Helmholtz problem posed on the union of all interfaces. The corresponding operator takes the form "identity + contraction".

7.9. Two-level preconditioning for h-version boundary element approximation of hypersingular operator with GenEO

We consider symmetric positive definite operators stemming from boundary integral equation (BIE), and we analysed a two-level preconditioner where the coarse space is built using local generalized eigenproblems in the overlap. We will refer to this coarse space as the GenEO coarse space. We obtained bounds on the condition number of the preconditionned system. In this work package, we also performed large scale numerical experiments for testing the scalability of our approach. We relied on parallel implementation of our algorithm.

7.10. Adaptive resolution of linear systems based on a posteriori error estimators

In [18] we discuss a new adaptive approach for iterative solution of sparse linear systems arising from partial differential equations (PDEs) with self-adjoint operators. The idea is to use the a posteriori estimated local distribution of the algebraic error in order to steer and guide the solve process in such way that the algebraic error is reduced more efficiently in the consecutive iterations. We first explain the motivation behind the proposed procedure and show that it can be equivalently formulated as constructing a special combination of preconditioner and initial guess for the original system. We present several numerical experiments in order to identify when the adaptive procedure can be of practical use.

7.11. Adaptive hierarchical subtensor partitioning for tensor compression

In [33] a numerical method is proposed to compress a tensor by constructing a piece-wise tensor approximation. This is defined by partitioning a tensor into sub-tensors and by computing a low-rank tensor approximation (in a given format) in each sub-tensor. Neither the partition nor the ranks are fixed a priori, but, instead, are obtained in order to fulfill a prescribed accuracy and optimize, to some extent, the storage. The different steps of the method are detailed and some numerical experiments are proposed to assess its performances.

7.12. Frictionless contact problem for hyper-elastic materials with interior point optimizer

In [35] we present a method to solve the mechanical problems undergoing finite deformations and the unilateral contact problems without friction for hyperelastic materials. We apply it to an industrial application: contact between a mechanical gasket and an obstacle. The main idea is to formulate the contact problem into an optimization one, in or- der to use the Interior Point OPTimizer (IPOPT) to solve it. Finally, the FreeFEM software is used to compute and solve the contact problem. Our method is validated against several benchmarks and used on an industrial application example.

7.13. A posteriori error estimates for Darcy's problem coupled with the heat equation

In [25] we derive a posteriori error estimates, in two and three dimensions, for the heat equation coupled with Darcy's law by a nonlinear viscosity depending on the temperature. We introduce two variational formulations and discretize them by finite element methods. We prove optimal a posteriori errors with two types of computable error indicators. The first one is linked to the linearization and the second one to the discretization. Then we prove upper and lower error bounds under regularity assumptions on the solutions. Finally, numerical computations are performed to show the effectiveness of the error indicators.

AVALON Project-Team

6. New Results

6.1. Energy Efficiency in HPC and Large Scale Distributed Systems

Participants: Laurent Lefèvre, Dorra Boughzala, Thierry Gautier.

6.1.1. Performance and Energy Analysis of OpenMP Runtime Systems with Dense Linear Algebra Algorithms

In the article [4], we analyze performance and energy consumption of five OpenMP runtime systems over a non-uniform memory access (NUMA) platform. We also selected three CPU-level optimizations or techniques to evaluate their impact on the runtime systems: processors features Turbo Boost and C-States, and CPU Dynamic Voltage and Frequency Scaling through Linux CPUFreq governors. We present an experimental study to characterize OpenMP runtime systems on the three main kernels in dense linear algebra algorithms (Cholesky, LU, and QR) in terms of performance and energy consumption. Our experimental results suggest that OpenMP runtime systems can be considered as a new energy leverage, and Turbo Boost, as well as C-States, impacted significantly performance and energy. CPUFreq governors had more impact with Turbo Boost disabled, since both optimizations reduced performance due to CPU thermal limits. An LU factorization with concurrent-write extension from libKOMP achieved up to 63% of performance gain and 29% of energy decrease over original PLASMA algorithm using GNU C compiler (GCC) libGOMP runtime. This paper was first published online in 2018-08-09.

6.1.2. Building and Exploiting the Table of Leverages in Large Scale HPC Systems

Large scale distributed systems and supercomputers consume huge amounts of energy. To address this issue, an heterogeneous set of capabilities and techniques that we call leverages exist to modify power and energy consumption in large scale systems. This includes hardware related leverages (such as Dynamic Voltage and Frequency Scaling), middleware (such as scheduling policies) and application (such as the precision of computation) energy leverages. Discovering such leverages, benchmarking and orchestrating them, remains a real challenge for most of the users. We have formally defined energy leverages, and we proposed a solution to automatically build the table of leverages associated with a large set of independent computing resources. We have shown that the construction of the table can be parallelized at very large scale with a set of independent nodes in order to reduce its execution time while maintaining precision of observed knowledge. In 2019 we have explored the leverage energy-efficient non-lossy compression for data-intensive applications [9].

6.2. HPC Component Models and Runtimes

Participants: Thierry Gautier, Christian Perez, Laurent Turpin, Marie Durand, Philippe Virouleau.

6.2.1. Fine-Grained MPI+OpenMP Plasma Simulations: Communication Overlap with Dependent Tasks

In the article [15], we demonstrate how OpenMP 4.5 tasks can be used to efficiently overlap computations and MPI communications based on a case-study conducted on multi-core and many-core architectures. The paper focuses on task granularity, dependencies and priorities, and also identifies some limitations of OpenMP. Results on 64 Skylake nodes show that while 64% of the wall-clock time is spent in MPI communications, 60% of the cores are busy in computations, which is a good result. Indeed, the chosen dataset is small enough to be a challenging case in terms of overlap and thus useful to assess worst-case scenarios in future simulations. Two key features were identified: by using task priority we improved the performance by 5.7% (mainly due to an improved overlap), and with recursive tasks we shortened the execution time by 9.7%. We also illustrate the need to have access to tools for task tracing and task visualization. These tools allowed a fine understanding and a performance increase for this task-based OpenMP+MPI code.

6.2.2. Patches to LLVM compiler

We propose two source code patches to LLVM https://reviews.llvm.org/D63196 and https://reviews.llvm.org/D67447 in order to improve performance of application using numerous fine grain tasks such as [15]. Patches were accepted in 2019.

6.3. Modeling and Simulation of Parallel Applications and Distributed Infrastructures

Participants: Eddy Caron, Zeina Houmani, Frédéric Suter.

6.3.1. Bridging Concepts and Practice in eScience via Simulation-driven Engineering

The CyberInfrastructure (CI) has been the object of intensive research and development in the last decade, resulting in a rich set of abstractions and interoperable software implementations that are used in production today for supporting ongoing and breakthrough scientific discoveries. A key challenge is the development of tools and application execution frameworks that are robust in current and emerging CI configurations, and that can anticipate the needs of upcoming CI applications. In [14] we presented WRENCH, a framework that enables simulation-driven engineering for evaluating and developing CI application execution frameworks. WRENCH provides a set of high-level simulation abstractions that serve as building blocks for developing custom simulators. These abstractions rely on the scalable and accurate simulation models that are provided by the SimGrid simulation framework. Consequently, WRENCH makes it possible to build, with minimum software development effort, simulators that can accurately and scalably simulate a wide spectrum of large and complex CI scenarios. These simulators can then be used to evaluate and/or compare alternate platform, system, and algorithm designs, so as to drive the development of CI solutions for current and emerging applications.

6.3.2. Accurately Simulating Energy Consumption of I/O-intensive Scientific Workflows

While distributed computing infrastructures can provide infrastructure-level techniques for managing energy consumption, application-level energy consumption models have also been developed to support energy-efficient scheduling and resource provisioning algorithms. In [7], we analyze the accuracy of a widely-used application-level model that have been developed and used in the context of scientific workflow executions. To this end, we profile two production scientific workflows on a distributed platform instrumented with power meters. We then conduct an analysis of power and energy consumption measurements. This analysis shows that power consumption is not linearly related to CPU utilization and that I/O operations significantly impact power, and thus energy consumption. We then propose a power consumption model that accounts for I/O operations, including the impact of waiting for these operations to complete, and for concurrent task executions on multi-socket, multi-core compute nodes. We implement our proposed model as part of a simulator that allows us to draw direct comparisons between real-world and modeled power and energy consumption. We find that our model has high accuracy when compared to real-world executions. Furthermore, our model improves accuracy by about two orders of magnitude when compared to the traditional models used in the energy-efficient workflow scheduling literature.

6.4. Cloud Resource Management

Participants: Eddy Caron, Jad Darrous, Christian Perez.

6.4.1. On the Importance of Container Image Placement for Service Provisioning in the Edge

Edge computing promises to extend Clouds by moving computation close to data sources to facilitate short-running and low-latency applications and services. Providing fast and predictable service provisioning time presets a new and mounting challenge, as the scale of Edge-servers grows and the heterogeneity of networks between them increases. Our work [6] is driven by a simple question: can we place container images across Edge-servers in such a way that an image can be retrieved to any Edge-server fast and in a predictable time. To this end, we present KCBP and KCBP-WC, two container image placement algorithms which aim to reduce the maximum retrieval time of container images. KCBP and KCBP-WC are based on k-Center optimization. However, KCBP-WC tries to avoid placing large layers of a container image on the same Edge-server. Evaluations using trace-driven simulations show that KCBP and KCBP-WC can be applied to various network configurations and reduce the maximum retrieval time of container images by 1.1x to 4x compared to state-of-the-art placements (*i.e.,* Best-Fit and Random).

Data-intensive clusters are heavily relying on distributed storage systems to accommodate the unprecedented growth of data. Hadoop distributed file system (HDFS) is the primary storage for data analytic frameworks such as Spark and Hadoop. Traditionally, HDFS operates under replication to ensure data availability and to allow locality-aware task execution of data-intensive applications. Recently, erasure coding (EC) is emerging as an alternative method to replication in storage systems due to the continuous reduction in its computation overhead. We have conducted an extensive experimental study to understand the performance of data-intensive applications under replication and EC [5], [23]. We use representative benchmarks on the Grid'5000 testbed to evaluate how analytic workloads, data persistency, failures, the back-end storage devices, and the network configuration impact their performances. Our study sheds the light not only on the potential benefits of erasure coding in data-intensive clusters but also on the aspects that may help to realize it effectively.

6.5. Data Stream Processing on Edge Computing

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

6.5.1. Operator Placement for Data Stream Processing on Fog/Edge Computing

DSP (Data Stream Processing) frameworks are often employed to process the large amount of data generated by the increasing number of IoT devices. A DSP application is commonly structured as a directed graph, or dataflow, whose vertices are operators that perform transformations over the incoming data and edges representing the data dependencies between operators. Such applications are often deployed on the Cloud in order to explore the large number of available resources and its pay-as-you-go business model. Fog computing enables offloading operators from the cloud by placing them close to where the data is generated, whereby reducing the time to process data events. However, fog computing resources often have lower capacity than those available in the Cloud. When offloading operators from the Cloud, the scheduler needs to adjust their level of parallelism and hence decides on the number of operator instances to create during placement in order to achieve a given throughput. This gives rise to two interrelated issues, namely deciding the operators parallelism and computing their placement onto available resources [16].

While addressing the placement problem [8], we proposed an approach consisting of a programming model and real-world implementation of an IoT application. The results show that our approach can minimise the end-to-end latency by at least 38% by pushing part of the IoT application to edge computing resources. Meanwhile, the edge-to-cloud data transfers are reduced by at least 38%, and the messaging costs are reduced by at least 50% when using the existing commercial edge cloud cost models.

In addition, we have designed and validated a discrete event simulation for modelling and simulation of DSP applications on edge computing environments [3].

6.5.2. Multi-Objective Reinforcement Learning for Reconfiguring Data Stream Analytics on Edge Computing

As DSP applications are often long-running, their workload and the infrastructure conditions can change over time. When changes occur, the application must be reconfigured. The operator reconfiguration consists of changing the initial placement by reassigning operators to different devices given target performance metrics. We modelled the operator reconfiguration as a Reinforcement Learning (RL) problem and defined a multi-objective reward considering metrics regarding operator reconfiguration, and infrastructure and application improvement [11]. We also use Monte Carlo Tree Search to organise the episodes generated during simulation and training [12]. Experimental results show that reconfiguration algorithms that minimise only end-to-end processing latency can have a substantial impact on WAN traffic and communication cost. The results also demonstrate that when reconfiguring operators, RL algorithms improve by over 50% the performance of the initial placement provided by state-of-the-art approaches.

6.6. An Operational Tool for Software Asset Management Improvement

Participants: Eddy Caron, Arthur Chevalier.

6.6.1. Multi-objective algorithm that guarantees license compliance

We have developed a new feature to OpTISAM, an Orange™ software offering tools to perform Software Asset Management (SAM) much more efficiently in order to be able to ensure the full compliance with all contracts from each software and a new type of deployment taking into account these aspects and other additional parameters like energy and performance. Our new feature is a multi-objective algorithm for deploying services in the Cloud that guarantees license compliance while reducing energy consumption but maintaining reasonable performance. In both cases of use and with a significant set of 5000 servers, we were able to show our approach is close to the best values in each criterion while dropping less than 10% of performance each time while keeping a full compliance.

6.7. Platform

Participants: Thierry Gautier, Christian Perez, Simon Delamare, Laurent Lefèvre.

6.7.1. Gemini cluster based on DGX-1 high density computer

The LECO experimental platform is a new medium size scientific instrument funded by DRRT and Inria to investigate research related to BigData and HPC. It is bi-located in Grenoble as part of the the HPCDA computer managed by UMS GRICAD (deployed in 2018) and in Lyon as part of the Grid5K Gemini cluster. The Gemini cluster is composed of two DGX-1 high density computers for HPC and BigData. Each computers has 8 NVIDIA V100 GPGPU cards with 4 infiniband high speed network cards.

DATAMOVE Project-Team

7. New Results

7.1. Integration of High Performance Computing and Data Analytics

7.1.1. In Situ Processing Model

The work in [2] focuses on proposing a model for in situ analysis taking into account memory constraints. This model is used to provide different scheduling policies to determine both the number of resources that should be dedicated to analysis functions, and that schedule efficiently these functions. We evaluate them and show the importance of considering memory constraints when choosing in between in situ and in transit resource allocation.

7.1.2. I/O Characterization

I/O operations are the bottleneck of several HPC applications due to the difference between process- ing and data access speeds. Hence, it is important to understand and characterize the typical I/O behavior of these applications, so we can identify problems in HPC architectures and propose solutions. In [3], we conducted an extensive analysis to collect and analyze information about applications that run in the Santos Dumont supercomputer, deployed in the National Laboratory for Scientific Computing (LNCC), in Brazil. In [9], we propose an I/O characterization approach that uses unsupervised learning to cluster jobs with similar I/O behavior, using information from high-level aggregated traces.

7.1.3. Online adaptation of the I/O stack to applications

I/O optimization techniques such as request scheduling can improve performance mainly for the access patterns they target, or they depend on the precise tune of parameters. In [19], we propose an approach to adapt the I/O forwarding layer of HPC systems to the application access patterns by tuning a request scheduler. Our case study is the TWINS scheduling algorithm, where performance improvements depend on the time window parameter, which depends on the current workload. Our approach uses a reinforcement learning technique to make the system capable of learning the best parameter value to each access pattern during its execution, without a previous training phase. Our approach can achieve a precision of 88% on the parameter selection in the first hundreds of observations of an access pattern. After having observed an access pattern for a few minutes (not necessarily contiguously), the system will be able to optimize its performance for the rest of the life of the system (years).

Such an auto-tuning approach requires a classification of application access patterns,to separate situations where the optimization techniques will have a different performance behavior. Such a classification is not available in the stateless server-side, hence it has to be estimated from metrics on recent accesses. In [8], we evaluate three machine learning techniques to automatically detect the I/O access pattern of HPC applications at run time: decision trees, random forests, and neural networks. We also proposed in [15] a pattern matching approach for server-side access pattern detection for the HPC I/O stack. The goal is to empower the system to learn a classification during the execution of the system, by representing access patterns by all relevant metrics. We build a time series to represent accesses spatiality, and use a pattern matching algorithm, in addition to an heuristic, to compare it to known patterns.

7.1.4. Data management for workflow execution

In [11], we studied a typical scenario in research facilities. Instrumental data is generated by lab equipment such as microscopes, collected by researchers into USB devices, and analyzed in their own computers. In this scenario, an instrumental data management framework could store data in a institution-level storage infrastructure and allow to execute tasks to analyze this data in some available processing nodes. This setup has the advantages of promoting reproducible research and the efficient usage of the expensive lab equipment (in addition to increasing researchers productivity). We detailed the requirements for such a framework regarding the needs of our case study of the CEA, analyzed performance limitations of the proposed architecture, and pointed to the connection between centralized storage and the processing nodes as the critical point.

In order to alleviate this bottleneck, we investigated using the storage devices of the processing nodes as a cache for the remote storage, and replication strategies to maximize data locality for tasks. A simulator called RepliSim was developed for this research.

7.2. Data Aware Batch Scheduling

We obtained in 2018 two important results on on-line scheduling using resource augmentation. The main idea is that the algorithm is applied to a more powerful environment than that of the adversary. We focused more specifically on the mechanism of rejection based on the concept of duality for mathematical programming applied for the analysis of the algorithm's performance. More precisely, we proposed 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 analysis concerned usual sequential jobs. These results have been distinguished among the most significant ones on the annual ACM review of on-line algorithms. We extended this work on a practical side by applying the analysis to actual batch schedulers with parallel jobs, rejection was interpreted as redirecting jobs to some predefined machines.

Machine Learning is a hot topic which received recently a great attention for dealing with the huge amount of data produced by the explosion of the digital applications and for dealing with uncertainties. The members of DataMove promoted a methodology based on simulation and machine learning to obtain efficient dynamic scheduling policies. The main idea is to focus the learning scheme targeting the policies them-selves, and not the specific parameters of the problem. Today, this methodology is mature and it is applied in several project like ANR Energumen (performances and replaced by energy saving). We also launched a new project at MIAI on edge Intelligence. The idea is to propose an alternative to the high-consuming classical IA by doing most of the computations close the the place where the data are produced. We are developing both an efficient task orchestration framework and distributed learning algorithms.

We wrote a survey [20] on scheduling on heterogeneous machines where we provided a complete benchmark suite and we recoded all existing algorithms and compared them.

HIEPACS Project-Team

7. New Results

7.1. High-performance computing on next generation architectures

7.1.1. Memory optimization for the training phase of deep convolutional networks

Training Deep Neural Networks is known to be an expensive operation, both in terms of computational cost and memory load. Indeed, during training, all intermediate layer outputs (called activations) computed during the forward phase must be stored until the corresponding gradient has been computedin the backward phase. These memory requirements sometimes prevent to consider larger batch sizes and deeper networks, so that they can limit both convergence speed and accuracy. Recent works have proposed to offload some of the computed forward activations from the memory of the GPU to the memory of the CPU. This requires to determine which activations should be offloaded and when these transfers from and to the memory of the GPU should take place. In [28], We prove that this problem is NP-hard in the strong sense, and we propose two heuristics based on relaxations of the problem. We perform extensive experimental evaluation on standard Deep Neural Networks. We compare the performance of our heuristics against previous approaches from the literature, showing that they achieve much better performance in a wide variety of situations.

In [23], we also introduce a new activation checkpointing method which allows to significantly decrease memory usage when training Deep Neural Networks with the back-propagation algorithm. Similarly to checkpointing techniques coming from the literature on Automatic Differentiation, it consists in dynamically selecting the forward activations that are saved during the training phase, and then automatically recomputing missing activations from those previously recorded. We propose an original computation model that combines two types of activation savings: either only storing the layer inputs, or recording the complete history of operations that produced the outputs (this uses more memory, but requires fewer recomputations in the backward phase), and we provide an algorithm to compute the optimal computation sequence for this model. This paper also describes a PyTorch implementation that processes the entire chain, dealing with any sequential DNN whose internal layers may be arbitrarily complex and automatically executing it according to the optimal checkpointing strategy computed given a memory limit. Through extensive experiments, we show that our implementation consistently outperforms existing checkpointing approaches for a large class of networks, image sizes and batch sizes.

In [4], [24], we consider the problem of optimally scheduling the backpropagation of Deep Join Networks. Deep Learning training memory needs can prevent the user to consider large models and large batch sizes. In this work, we propose to use techniques from memory-aware scheduling and Automatic Differentiation (AD) to execute a backpropagation graph with a bounded memory requirement at the cost of extra recomputations. The case of a single homogeneous chain, i.e. the case of a network whose all stages are identical and form a chain, is well understood and optimal solutions have been proposed in the AD literature. The networks encountered in practice in the context of Deep Learning are much more diverse, both in terms of shape and heterogeneity. In this work, we define the class of backpropagation graphs, and extend those on which one can compute in polynomial time a solution that minimizes the total number of recomputations. In particular we consider join graphs which correspond to models such as Siamese or Cross Modal Networks.

7.1.2. Sizing and Partitioning Strategies for Burst-Buffers to Reduce IO Contention

Burst-Buffers are high throughput and small size storage which are being used as an intermediate storage between the PFS (Parallel File System) and the computational nodes of modern HPC systems. They can allow to hinder to contention to the PFS, a shared resource whose read and write performance increase slower than processing power in HPC systems. A second usage is to accelerate data transfers and to hide the latency to the PFS. In this work, we concentrate on the first usage. We propose a model for Burst-Buffers and application transfers. We consider the problem of dimensioning and sharing the Burst-Buffers between several applications. This dimensioning can be done either dynamically or statically. The dynamic allocation considers that any application can use any available portion of the Burst-Buffers. The static allocation considers that when a new application enters the system, it is assigned some portion of the Burst-Buffers, which cannot be used by the other applications until that application leaves the system and its data is purged from it. We show that the general sharing problem to guarantee fair performance for all applications is an NP-Complete problem. We propose a polynomial time algorithms for the special case of finding the optimal buffer size such that no application is slowed down due to PFS contention, both in the static and dynamic cases. Finally, we provide evaluations of our algorithms in realistic settings. We use those to discuss how to minimize the overhead of the static allocation of buffers compared to the dynamic allocation. More information on these results can be found in [9].

7.1.3. Efficient Ordering of Kernel Submission on GPUs

In distributed memory systems, it is paramount to develop strategies to overlap the data transfers between memory nodes with the computations in order to exploit their full potential. In [11], we consider the problem of determining the order of data transfers between two memory nodes for a set of independent tasks with the objective of minimizing the makespan. We prove that, with limited memory capacity, the problem of obtaining the optimal data transfer order is NP-complete. We propose several heuristics to determine this order and discuss the conditions that might be favorable to different heuristics. We analyze our heuristics on traces obtained by running two molecular chemistry kernels, namely, Hartree–Fock (HF) and Coupled Cluster Singles Doubles (CCSD), on 10 nodes of an HPC system. Our results show that some of our heuristics achieve significant overlap for moderate memory capacities and resulting in makespans that are very close to the lower bound.

Concurrent kernel execution is a relatively new feature in modern GPUs, which was designed to improve hardware utilization and the overall system throughput. However, the decision on the simultaneous execution of tasks is performed by the hardware with a leftover policy, that assigns as many resources as possible for one task and then assigns the remaining resources to the next task. This can lead to unreasonable use of resources. In [30], we tackle the problem of co-scheduling for GPUs with and without preemption, with the focus on determining the kernels submission order to reduce the number of preemptions and the kernels makespan, respectively. We propose a graph-based theoretical model to build preemptive and non-preemptive schedules. We show that the optimal preemptive makespan can be computed by solving a Linear Program in polynomial time, and we propose an algorithm based on this solution which minimizes the number of preemptions. We also propose an algorithm that transforms a preemptive solution of optimal makespan into a non-preemptive solution with the smallest possible preemption overhead. We show, however, that finding the minimal amount of preemptions among all preemptive solutions of optimal makespan is a NP-hard problem, and computing the optimal non-preemptive schedule is also NP-hard. In addition, we study the non-preemptive problem, without searching first for a good preemptive solution, and present a Mixed Integer Linear Program solution to this problem. We performed experiments on real-world GPU applications and our approach can achieve optimal makespan by preempting 6 to 9% of the tasks. Our non-preemptive approach, on the other side, obtains makespan within 2.5% of the optimal preemptive schedules, while previous approaches exceed the preemptive makespan by 5 to 12%.

7.1.4. Scheduling Tasks on Two Types of Resources

We consider the problem of scheduling task graphs on two types of unrelated resources, which arises in the context of task-based runtime systems on modern platforms containing CPUs and GPUs. In [10], we focus on an algorithm named HeteroPrio, which was originally introduced as an efficient heuristic for a particular application. HeteroPrio is an adaptation of the well known list scheduling algorithm, in which the tasks are picked by the resources in the order of their acceleration factor. This algorithm is augmented with a spoliation mechanism: a task assigned by the list algorithm can later on be reassigned to a different resource if it allows to finish this task earlier. We propose here the first theoretical analysis of the HeteroPrio algorithm in the presence of dependencies. More specifically, if the platform contains m and n processors of each type, we show that the worst-case approximation ratio of HeteroPrio is between $1 + \max(m/n, n/m)$ and $2 + \max(m/n, n/m)$.

Our proof structure allows to precisely identify the necessary conditions on the spoliation strategy to obtain such a guarantee. We also present an in-depth experimental analysis, comparing several such spoliation strategies, and comparing HeteroPrio with other algorithms from the literature. Although the worst case analysis shows the possibility of pathological behavior, HeteroPrio is able to produce, in very reasonable time, schedules of significantly better quality.

The evolution in the design of modern parallel platforms leads to revisit the scheduling jobs on distributed heterogeneous resources. We contribute to [31], a survey whose goal is to present the main existing algorithms, to classify them based on their underlying principles and to propose unified implementations to enable their fair comparison, both in terms of running time and quality of schedules, on a large set of common benchmarks that we made available for the community [27]. Beyond this comparison, our goal is also to understand the main difficulties that heterogeneity conveys and the shared principles that guide the design of efficient algorithms.

7.1.5. Data-Locality Aware Tasks Scheduling with Replicated Inputs

In [5], we consider the influence on data-locality of the replication of data files, as automatically performed by Distributed File Systems such as HDFS. Replication is known to have a crucial impact on data locality in addition to system fault tolerance. Indeed, intuitively, having more replicas of the same input file gives more opportunities for this task to be processed locally, i.e. without any input file transfer. Given the practical importance of this problem, a vast literature has been proposed to schedule tasks, based on a random placement of replicated input files. Our goal in this paper is to study the performance of these algorithms, both in terms of makespan minimization (minimize the completion time of the last task when non-local processing is forbidden) and communication minimization (minimize the number of non-local tasks when no idle time on resources is allowed). In the case of homogenous tasks, we are able to prove, using models based on "balls into bins" and "power of two choices" problems, that the well known good behavior of classical strategies can be theoretically grounded. Going further, we even establish that it is possible, using semi-matchings theory, to find the optimal solution in very small time. We also use known graph-orientation results to prove that this optimal solution is indeed near-perfect with strong probability. In the more general case of heterogeneous tasks, we propose heuristics solutions both in the clairvoyant and non-clairvoyant cases (i.e. task length is known in advance or not), and we evaluate them through simulations, using actual traces of a Hadoop cluster.

7.2. High performance solvers for large linear algebra problems

7.2.1. Deflation and preconditioning strategies for sequences of sampled stochastic elliptic equations

We are interested in the quantification of uncertainties in discretized elliptic partial differential equations with random coefficients. In sampling-based approaches, this relies on solving large numbers of symmetric positive definite linear systems with different matrices. In this work, we investigate recycling Krylov subspace strategies for the iterative solution of sequences of such systems. The linear systems are solved using deflated conjugate gradient (CG) methods, where the Krylov subspace is augmented with approximate eigenvectors of the previously sampled operator. These operators are sampled by Markov chain Monte Carlo, which leads to sequences of correlated matrices. First, the following aspects of eigenvector approximation, and their effect on deflation, are investigated: (i) projection technique, and (ii) restarting strategy of the eigen-search space. Our numerical experiments show that these aspects only impact convergence behaviors of deflated CG at the early stages of the sampling sequence. Second, unlike sequences with multiple right-hand sides and a constant operator, our experiments with multiple matrices show the necessity to orthogonalize the iterated residual of the linear system with respect to the deflation subspace, throughout the sampling sequence. Finally, we observe a synergistic effect of deflation and block-Jacobi (bJ) preconditioning. While the action of bJ preconditioners leaves a trail of isolated eigenvalues in the spectrum of the preconditioned operator, for 1D problems, the corresponding eigenvectors are well approximated by the recycling strategy. Then, up to a certain number of blocks, deflated CG methods with bJ preconditioners achieve similar convergence behaviors to those observed with CG when using algebraic multigrid (AMG) as a preconditioner.

This work, developed in the framework of the PhD thesis of Nicolas Venkovic in collaboration with P. Mycek (Cerfacs) and O. Le Maitre (CMAP, Ecole Polytechnique), will be presented at the next Copper Mountain conference on iterative methods.

7.2.2. Robust preconditionners via generalized eigenproblems for hybrid sparse linear solvers

The solution of large sparse linear systems is one of the most time consuming kernels in many numerical simulations. The domain decomposition community has developed many efficient and robust methods in the last decades. While many of these solvers fall into the abstract Schwarz (aS) framework, their robustness has originally been demonstrated on a case-by-case basis. In this work, we propose a bound for the condition number of all deflated aS methods provided 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 or improve 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. Numerical experiments illustrate these statements.

More information on these results can be found in [3]

7.2.3. Rank Revealing QR Methods for Sparse Block Low Rank Solvers

In the context of the ANR Sashimi project and the Phd of Esragul Korkmaz, we have investigated several compression methods of dense blocks appearing inside sparse matrix solvers to reduce the memory consumption, as well as the time to solution.

Solving linear equations of type Ax=b for large sparse systems frequently emerges in science and engineering applications, which creates the main bottleneck. In spite that the direct methods are costly in time and memory consumption, they are still the most robust way to solve these systems. Nowadays, increasing the amount of computational units for the supercomputers became trendy, while the memory available per core is reduced. Therefore, when solving these linear equations, memory reduction becomes as important as time reduction. While looking for the lowest possible compression rank, Singular Value Decomposition (SVD) gives the best result. It is however too costly as the whole factorization is computed to find the resulting rank. In this respect, rank revealing QR decomposition variants are less costly, but can introduce larger ranks. Among these variants, column pivoting or matrix rotation can be applied on the matrix A, such that the most important information in the matrix is gathered to the leftmost columns and the remaining unnecessary information can be omitted. For reducing the communication cost of the classical QR decomposition with column pivoting, blocking versions with randomization are suggested as an alternative solution to find the pivots. In these randomized variants, the matrix A is projected on a much lower dimensional matrix by using an independent and identically distributed Gaussian matrix so that the pivoting/rotational matrix can be computed on the lower dimensional matrix. In addition, to avoid unnecessary updates of the trailing matrix at each iteration, a truncated randomized method is suggested and shown to be more efficient for larger matrix sizes. Thanks to these methods, closer results to SVD can be obtained and the cost of compression can be reduced.

A comparison of all these methods in terms of complexity, numerical stability and performance have been presented at the national conference COMPAS'2019 [18], and at the international workshop SparseDay'2019 [19].

7.2.4. Accelerating Krylov linear solvers with agnostic lossy data compression

In the context of the Inria International Lab JLESC we have an ongoing collaboration with Argonne National Laboratory on the use of agnostic compression techniques to reduce the memory footprint of iterative linear solvers. Krylov methods are among the most efficient and widely used algorithms for the solution of large linear systems. Some of these methods can, however, have large memory requirements. Despite the fact that modern high-performance computing systems have more and more memory available, the memory used by applications remains a major concern when solving large scale problems. This is one of the reasons why interest in lossy data compression techniques has grown tremendously in the last two decades: it can reduce the amount of information that needs to be stored and communicated. Recently, it has also been shown that Krylov methods allow for some inexactness in the matrix-vector product that is typically required in each iteration. We showed that the loss of accuracy caused by compressing and decompressing the solution of the preconditioning step in the flexible generalized minimal residual method can be interpreted as an inexact matrix-vector product. This allowed us to find a bound on the maximum compression error in each iteration based on the theory of inexact Krylov methods. We performed a series of numerical experiment in order to validate our results. A number of “relaxed compression strategies” was also considered in order to achieve higher compression ratios.

The results of this joint effort will be presented to the next SIAM conférence on Parallel processing SIAM-PP’20.

7.2.5. Energy Analysis of a Solver Stack for Frequency-Domain Electromagnetics

High-performance computing aims at developing models and simulations for applications in numerous scientific fields. Yet, the energy consumption of these HPC facilities currently limits their size and performance, and consequently the size of the tackled problems. The complexity of the HPC software stacks and their various optimizations makes it difficult to finely understand the energy consumption of scientific applications. To highlight this difficulty on a concrete use-case, we perform in [8] an energy and power analysis of a software stack for the simulation of frequency-domain electromagnetic wave propagation. This solver stack combines a high order finite element discretization framework of the system of three-dimensional frequency-domain Maxwell equations with an algebraic hybrid iterative-direct sparse linear solver. This analysis is conducted on the KNL-based PRACE-PCP system. Our results illustrate the difficulty in predicting how to trade energy and runtime.

7.2.6. Exploiting Parameterized Task-graph in Sparse Direct Solvers

Task-based programming models have been widely studied in the context of dense linear algebra, but remains less studied for the more complex sparse solvers. In this talk [17], we have presented the use of two different programming models: Sequential Task Flow from StarPU, and Parameterized Task Graph from PaRSEC to parallelize the factorization step of the PaStiX sparse direct solver. We have presented how those programming models have been used to integrate more complex and finer parallelism to take into account new architectures with many computational units. Efficiency of such solutions on homogeneous and heterogeneous architectures with a spectrum of matrices from different applications have been shown. We also have presented how such solutions enable, without extra cost to the programmer, better performance on irregular computations such as in the block low-rank implementation of the solver.

7.2.7. Block Low-rank Algebraic Clustering for Sparse Direct Solvers

In these talks [20], [21], we adressed the Block Low-Rank (BLR) clustering problem, to cluster unknowns within separators appearing during the factorization of sparse matrices. We have shown that methods considering only intra-separators connectivity (i.e., k-way or recursive bissection) as well as methods managing only interaction between separators have some limitations. The new strategy we proposed consider interactions between a separator and its children to pre-select some interactions while reducing the number of off-diagonal blocks. We demonstrated how this method enhance the BLR strategies in the sparse direct supernodal solver PaStiX, and discuss how it can be extended to low-rank formats with more than one level of hierarchy.

7.2.8. Leveraging Task-Based Polar Decomposition Using PARSEC on Massively Parallel Systems

In paper [13], we describe how to leverage a task-based implementation of the polar decomposition on massively parallel systems using the PARSEC dynamic runtime system. Based on a formulation of the iterative QR Dynamically-Weighted Halley (QDWH) algorithm, our novel implementation reduces data traffic while exploiting high concurrency from the underlying hardware architecture. First, we replace the most time-consuming classical QR factorization phase with a new hierarchical variant, customized for the specific structure of the matrix during the QDWH iterations. The newly developed hierarchical QR for QDWH exploits not only the matrix structure, but also shortens the length of the critical path to maximize hardware occupancy. We then deploy PARSEC to seamlessly orchestrate, pipeline, and track the data dependencies of the various linear algebra building blocks involved during the iterative QDWH algorithm. PARSEC enables to overlap communications with computations thanks to its asynchronous scheduling of fine-grained computational tasks. It employs look-ahead techniques to further expose parallelism, while actively pursuing the critical path. In addition, we identify synergistic opportunities between the task-based QDWH algorithm and the PARSEC framework. We exploit them during the hierarchical QR factorization to enforce a locality-aware task execution. The latter feature permits to minimize the expensive inter-node communication, which represents one of the main bottlenecks for scaling up applications on challenging distributed-memory systems. We report numerical accuracy and performance results using well and ill-conditioned matrices. The benchmarking campaign reveals up to 2X performance speedup against the existing state-of-the-art implementation for the polar decomposition on 36,864 cores.

7.3. Parallel Low-Rank Linear System and Eigenvalue Solvers Using Tensor Decompositions

It is common to accelerate the boundary element method by compression techniques (FMM, $\mathcal{H}$-matrix/ACA) that enable a more accurate solution or a solution in higher frequency. In this article, we present a compression method based on a transformation of the linear system into the tensor-train format by the quantization technique. The method is applied to a scattering problem on a canonical object with a regular mesh and improves the performance obtained from existing methods. This method has been presented at the 22nd International Conference on the Computation of Electromagnetic Field, (COMPUMAG 2019) [12] and an extented version is accepted in IEEE TRANSACTIONS ON MAGNETICS.

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

7.4.1. StarPart Redesign

In the context of the french ICARUS project (FUI), which focuses the development of high-fidelity calculation tools for the design of hot engine parts (aeronautics & automotive), we are looking to develop new load-balancing algorithms to optimize the complex numerical simulations of our industrial and academic partners (Turbomeca, Siemens, Cerfacs, Onera, ...). Indeed, the efficient execution of large-scale coupled simulations on powerful computers is a real challenge, which requires revisiting traditional load-balancing algorithms based on graph partitioning. A thesis on this subject has already been conducted in the Inria HiePACS team in 2016, which has successfully developed a co-partitioning algorithm that balances the load of two coupled codes by taking into account the coupling interactions between these codes. This work was initially integrated into the StarPart platform. The necessary extension of our algorithms to parallel & distributed (increasingly dynamic) versions has led to a complete redesign of StarPart, which has been the focus of our efforts this year (as in the previous year). The StarPart framework provides the necessary building blocks to develop new graph algorithms in the context of HPC, such as those we are targeting. The strength of StarPart lies in the fact that it is a light runtime system applied to the issue of "graph computing". It provides a unified data model and a uniform programming interface that allows easy access to a dozen partitioning libraries, including Metis, Scotch, Zoltan, etc. Thus, it is possible, for example, to load a mesh from an industrial test case provided by our partners (or an academic graph collection as DIMACS'10) and to easily compare the results for the different partitioners integrated in StarPart.

Alongside this work, we are beginning to work on the application of learning techniques to the problem of graph partitioning. Recent work on GCNs (Graph Convolutional Networks) is an interesting approach that we will explore.

7.5. Application Domains

7.5.1. Material physics

7.5.1.1. EigenSolver

The adaptive vibrational configuration interaction algorithm has been introduced as a new eigennvalues method for large dimension problem. It is based on the construction of nested bases for the discretization of the Hamiltonian operator according to a theoretical criterion that ensures the convergence of the method. It efficiently reduce the dimension of the set of basis functions used and then we are able solve vibrationnal eigenvalue problem up to the dimension 15 (7 atoms). Beyond this molecule size, two major issues appear. First, the size of the approximation domain increases exponentially with the number of atoms and the density of eigenvalues in the target area.
This year we have worked on two main areas. First of all, not all the eigenvalues that are calculated are determined by spectroscopy and therefore do not interest chemists. Only eigenvalues with an intensity are relevant. Also, we have set up a selection of interesting eigenvalues using the intensity operator. This requires calculating the scalar product between the smallest eigenvalues and the dipole moment applied to an eigenvector to evaluate its intensity. In addition, to get closer to the experimental values, we introduced the Coriolis operator into the Hamiltonian. A document is being written on these last two points showing that we can reach for a molecule 10 atoms the area of interest (i.e. more than 2 400 eigenvalues). Moreover, we continue to extend our shared memory parallelization to distributed memory using the message exchange paradigm to speedup the eigensolver time.

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

7.5.2.1. Numerical and parallel scalable hybrid solvers in large scale calculations

We have been working with the NACHOS team on the treatment of the system of three-dimensional frequency-domain (or time-harmonic) Maxwell equations using a high order hybridizable discontinuous Galerkin (HDG) approximation method combined to domain decomposition (DD) based hybrid iterative-direct parallel solution strategies. The proposed HDG method preserves the advantages of classical DG methods previously introduced for the time-domain Maxwell equations, in particular in terms of accuracy and flexibility with regards to the discretization of complex geometrical features, while keeping the computational efficiency at the level of the reference edge element based finite element formulation widely adopted for the considered PDE system. We study in details the computational performances of the resulting DD solvers in particular in terms of scalability metrics by considering both a model test problem and more realistic large-scale simulations performed on high performance computing systems consisting of networked multicore nodes. More information on these results can be found in [2].

In the context of a parallel plasma physics simulation code, we perform a qualitative performance study between two natural candidates for the parallel solution of 3D Poisson problems that are multigrid and domain decomposition. We selected one representative of each of these numerical techniques implemented in state of the art parallel packages and show that depending on the regime used in terms of number of unknowns per computing cores the best alternative in terms of time to solution varies. Those results show the interest of having both types of numerical solvers integrated in a simulation code that can be used in very different configurations in terms of problem sizes and parallel computing platforms. More information on these results will be shortly available in an Inria scientific report.

7.5.2.2. Efficient Parallel Solution of the 3D Stationary Boltzmann Transport Equation for Diffusive Problems

In the context of a collaboration with EDF-Lab with the Phd of Salli Moustafa, we present an efficient parallel method for the deterministic solution of the 3D stationary Boltzmann transport equation applied to diffusive problems such as nuclear core criticality computations. Based on standard MultiGroup-Sn-DD discretization schemes, our approach combines a highly efficient nested parallelization strategy with the PDSA parallel acceleration technique applied for the first time to 3D transport problems. These two key ingredients enable us to solve extremely large neutronic problems involving up to 10^{12} degrees of freedom in less than an hour using 64 super-computer nodes.

These contributions have been published in Journal of Computational Physics (JCP) [7].

7.5.2.3. Bridging the Gap Between $\mathcal{H}$-Matrices and Sparse Direct Methods for the Solution of Large Linear Systems

For the sake of numerical robustness in aeroacoustics simulations, the solution techniques based on the factorization of the matrix associated with the linear system are the methods of choice when affordable. In that respect, hierarchical methods based on low-rank compression have allowed a drastic reduction of the computational requirements for the solution of dense linear systems over the last two decades. For sparse linear systems, their application remains a challenge which has been studied by both the community of hierarchical matrices and the community of sparse matrices. On the one hand, the first step taken by the community of hierarchical matrices most often takes advantage of the sparsity of the problem through the use of nested dissection. While this approach benefits from the hierarchical structure, it is not, however, as efficient as sparse solvers regarding the exploitation of zeros and the structural separation of zeros from non-zeros. On the other hand, sparse factorization is organized so as to lead to a sequence of smaller dense operations, enticing sparse solvers to use this property and exploit compression techniques from hierarchical methods in order to reduce the computational cost of these elementary operations. Nonetheless, the globally hierarchical structure may be lost if the compression of hierarchical methods is used only locally on dense submatrices. In [1], we have reviewed the main techniques that have been employed by both those communities, trying to highlight their common properties and their respective limits with a special emphasis on studies that have aimed to bridge the gap between them. With these observations in mind, we have proposed a class of hierarchical algorithms based on the symbolic analysis of the structure of the factors of a sparse matrix. These algorithms rely on a symbolic information to cluster and construct a hierarchical structure coherent with the non-zero pattern of the matrix. Moreover, the resulting hierarchical matrix relies on low-rank compression for the reduction of the memory consumption of large submatrices as well as the time to solution of the solver. We have also compared multiple ordering techniques based on geometrical or topological properties. Finally, we have opened the discussion to a coupling between the Finite Element Method and the Boundary Element Method in a unified computational framework.

7.5.2.4. Design of a coupled MUMPS - $\mathcal{H}$-Matrix solver for FEM-BEM applications

In that approach, the FEM matrix is eliminated by computing a Schur complement using MUMPS. Given the size of the BEM matrix, this can not be done in one operation, so it is done block by block, and added in the $\mathcal{H}$-matrix, which is then factorized to complete the process. The overall process yields an interesting boost in performance when compared to the previously existing approach that coupled MUMPS with a classical dense solver. However, a full comparison with all the other existing methods must still be performed (full $\mathcal{H}$-matrix solver with [22] or without nested dissection, iterative approaches, etc.).

7.5.2.5. Metabarcoding

Distance Geometry Problem (DGP) and Nonlinear Mapping (NLM) are two well established questions: DGP is about finding a Euclidean realization of an incomplete set of distances in a Euclidean space, whereas Nonlinear Mapping is a weighted Least Square Scaling (LSS) method. We show how all these methods (LSS, NLM, DGP) can be assembled in a common framework, being each identified as an instance of an optimization problem with a choice of a weight matrix. In [6], we studied the continuity between the solutions (which are point clouds) when the weight matrix varies, and the compactness of the set of solutions (after centering). We finally studied a numerical example, showing that solving the optimization problem is far from being simple and that the numerical solution for a given procedure may be trapped in a local minimum.

We are involved in the ADT Gordon ((partners: TADAAM (coordinator), STORM, HIEPACS, PLEIADE). The objectives of this ADT is to scale our slolver stack on a PLEIADE dimensioning metabarcoding application (multidimensional scaling method). Our goal is to be able to handle a problem leading to a distance matrix around 100 million individuals. Our contribution concerns the the scalability of the multidimensional scaling method and more particulary the random projection methods to speed up the SVD solver. Experiments son PlaFRIM and MCIA CURTA plateforms have allowed us to show that the solver stack was able to solve efficiently a large problem up to 300,000 individuals in less than 10 minutes on 25 nodes. This has highlighted that for these problem sizes the management of I/O, inputs and outputs with the disks, becomes critical and dominates calculation times.

KERDATA Project-Team

7. New Results

7.1. Convergence HPC and Big Data

7.1.1. Convergence at the data-processing level

Participants: Gabriel Antoniu, Alexandru Costan, Daniel Rosendo.

Traditional data-driven analytics relies on Big Data processing techniques, consisting of batch processing and real-time (stream) processing, potentially combined in a so-called *Lambda architecture*. This architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data.

On the other side, simulation-driven analytics is based on computational (usually physics-based) simulations of complex phenomena, which often leverage HPC infrastructures. The need to get fast and relevant insights from massive amounts of data generated by extreme-scale simulations led to the emergence of in situ and in transit processing approaches: they allow data to be visualized and processed interactively in real-time as data are produced, while the simulation is running.

To support hybrid analytics and continuous model improvement, we propose to combine the above data processing techniques in what we will call the *Sigma architecture*, a HPC-inspired extension of the Lambda architecture for Big Data processing [17]. Its instantiation in specific application settings depends of course of the specific application requirements and of the constraints that may be induced by the underlying infrastructure. Its main conceptual strength consists in the ability to leverage in a unified, consistent framework, data processing techniques that became reference in HPC in the Big Data communities respectively, without however being combined so far for joint usage in converged environments.

The given framework will integrate previously-validated approaches developed in our team, such as Damaris, a middleware system for efficient I/O management and large-scale in situ data processing, and KerA, a unified system for data flow ingestion and storage. The overall objective is to enable the usage of a large spectrum of Big Data analytics and Intelligence techniques at extreme scales in the Cloud and Edge, to support continuous intelligence (from streaming and historical data) and precise insights/predictions in real-time and fast decision making.

7.1.2. Pufferscale: Elastic storage to support dynamic hybrid workflows systems

Participants: Nathanaël Cheriere, Gabriel Antoniu.

User-space HPC data services are emerging as an appealing alternative to traditional parallel file systems, because of their ability to be tailored to application needs while eliminating unnecessary overheads incurred by POSIX compliance. Such services may need to be rescaled up and down to adapt to changing workloads, in order to optimize resource usage. This can be useful, for instance, to better support complex workflows that mix on-demand simulations and data analytics.

We formalized the operation of rescaling a distributed storage system as a multi objective optimization problem considering three criteria: load balance, data balance, and duration of the rescaling operation. We proposed a heuristic for rapidly finding a good approximate solution, while allowing users to weight the criteria as needed. The heuristic is evaluated with Pufferscale, a new, generic rescaling manager for microservice-based distributed storage systems [18].

To validate our approach in a real-world ecosystem, we showcase the use of Pufferscale as a means to enable storage malleability in the HEPnOS storage system for high energy physics applications.

7.2. Cloud and Edge processing

7.2.1. Benchmarking Edge processing frameworks

Participants: Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

With the spectacular growth of the Internet of Things, edge processing emerged as a relevant means to offload data processing and analytics from centralized Clouds to the devices that serve as data sources (often provided with some processing capabilities). While a large plethora of frameworks for edge processing were recently proposed, the distributed systems community has no clear means today to discriminate between them. Some preliminary surveys exist, focusing on a feature-based comparison.

We claim that a step further is needed, to enable a performance-based comparison. To this purpose, the definition of a benchmark is a necessity. We make a step towards the definition of a methodology for benchmarking Edge processing frameworks [20].

7.2.2. Analytical models for performance evaluation of stream processing

Participants: José Aguilar Canepa, Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

One of the challenges of enabling the Edge computing paradigm is to identify the situations and scenarios in which Edge processing is suitable to be applied. To this end, applications can be modeled as a graph consisting of tasks as nodes and data dependencies between them as edges. The problem comes down to deploying the application graph onto the network graph, that is, operators need to be put on machines, and finding the optimal cut in the graph between the Edge and Cloud resources (i.e., nodes in the network graph).

We have designed an algorithm that finds the optimal execution plan, with a rich cost model that lets users to optimize whichever goal they might be interested in, such as monetary costs, energetic consumption or network traffic, to name a few.

In order to validate the cost model and the effectiveness of the algorithm, a series of experiments were designed using two real-life stream processing applications: a closed-circuit television surveillance system, and an earthquake early warning system.

Two network infrastructures were designed to run the applications. The first one is a state-of-art infrastructure where all processing is done on the Cloud to serve as benchmark. The second one is an infrastructure produced by the algorithm. Both scenarios were executed on the Grid'5000. Several experiments are currently underway. The trade-offs of executing Cloud/Edge workloads with this model were published in [19].

7.2.3. Modeling smart cities applications

Participants: Edgar Romo Montiel, Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

Smart City applications have particular characteristics in terms of data processing and storage, which need to be taken into account by the underlying serving layers. The objective of this new activity is to devise clear models of the data handled by such applications. The data characteristics and the processing requirements does not have to match one-to-one. In some cases, some particular types of data might need one or more types of processing, depending on the use case. For example, small and fast data coming from sensors do not always have to be processed in real-time, but they could also be processed in a batch manner at a later stage.

This activity is the namely the topic of the SmartFastData associated team with the Instituto Politécnico Nacional of Mexico.

In a first phase, we focused on modeling the stream rates of data from sets of sensors in Smart Cities, specifically, from vehicles inside a closed coverage area. Those vehicles are connected in a V2I VANET, and they interact to applications in the Cloud such as traffic reports, navigation apps, multimedia downloading etc. This led to the design of a mathematical model to predict the time that a mobile sensor resides within a geographical designated area.

The proposed model uses Coxian distributions to estimate the time a vehicle requests Cloud services, so that the core challenge is to adjust their parameters. It was achieved by validating the model against real-life data traces from the City of Luxembourg, through extensive experiments on the Grid'5000.

Next, these models were used to estimate the resources needed in the Cloud (or at the Edge) in order to process the whole stream of data. We designed an auto-Scaling module able to adapt the resources with respect to the load. Using the Grid'5000, we evaluated the various possibility to place the prediction module: *(i)* at the Edge, close to data with less accuracy but faster results; or *(ii)* in the Cloud, with higher accuracy due to the global data, but higher latency as well.

7.3. AI across the digital continuum

7.3.1. Machine Learning in the context of Edge stream processing.

Participants: Pedro de Souza Bento Da Silva, Alexandru Costan, Gabriel Antoniu.

Our research aims to improve the accuracy of Earthquake Early Warning (EEW) systems by means of machine learning. EEW systems are designed to detect and characterize medium and large earthquakes before their damaging effects reach a certain location.

Traditional EEW methods based on seismometers fail to accurately identify large earthquakes due to their sensitivity to the ground motion velocity. The recently introduced high-precision GPS stations, on the other hand, are ineffective to identify medium earthquakes due to its propensity to produce noisy data. In addition, GPS stations and seismometers may be deployed in large numbers across different locations and may produce a significant volume of data consequently, affecting the response time and the robustness of EEW systems.

In practice, EEW can be seen as a typical classification problem in the machine learning field: multi-sensor data are given in input, and earthquake severity is the classification result. We introduce the Distributed Multi-Sensor Earthquake Early Warning (DMSEEW) system, a novel machine learning-based approach that combines data from both types of sensors (GPS stations and seismometers) to detect medium and large earthquakes.

DMSEEW is based on a new stacking ensemble method which has been evaluated on a real-world dataset validated with geoscientists. The system builds on a geographically distributed infrastructure (deployable on clouds and edge systems), ensuring an efficient computation in terms of response time and robustness to partial infrastructure failures. Our experiments show that DMSEEW is more accurate than the traditional seismometer-only approach and the combined-sensors (GPS and seismometers) approach that adopts the rule of relative strength.

These results have been accepted for publication at AAAI, a "A*" conference in the area of Artificial Intelligence [21].

7.3.2. ZettaFlow: Unified Fast Data Storage and Analytics Platform for IoT

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

The ZettaFlow platform (system of systems) provides a high-performance multi-model analytics-oriented storage and processing system, while supporting publish-subscribe streams and streaming, key-value and in-memory columnar APIs [16].

The ZettaFlow project is funded by EIT Digital from October 2019 to December 2020. It includes three partners: Inria for the platform development, TU Berlin for edge to cloud IoT optimizations with microservices, and Systematic Paris Region for the go-to-market strategy.

Our goal is to create a startup that will commercialize the ZettaFlow platform: a dynamic, unified and auto-balanced real-time storage and analytics industrial IoT platform. ZettaFlow will provide real-time visibility into machines, assets and factory operations and will automate data driven decisions for high-performance industrial processes.

ZettaFlow will bring a threefold impact to the IoT market.

1. Enable novel real-time edge applications that truly automate manufacturing, transportation and utilities processes.
2. Reduce deployment efforts and time-to-decision of IoT edge-cloud applications by 75% through automation, unified dynamic data management and streaming analytics.
3. Reduce human costs for monitoring and engineering (through edge intelligence) and IoT hardware costs by 50% through unified data collection/storage/analytics.

POLARIS Project-Team

7. New Results

7.1. Design of Experiments

Performance engineering of scientific HPC applications requires to measure repeatedly the performance of applications or of computation kernels, which consume a large amount of time and resources. It is essential to design experiments so as to reduce this cost as much as possible. Our contribution along this axis is twofold: (1) the investigation sound exploration techniques and (2) the control of experiments to ensure the measurements are as representative as possible of real workload.

Writing, porting, and optimizing scientific applications makes autotuning techniques fundamental to lower the cost of leveraging the improvements on execution time and power consumption provided by the latest software and hardware platforms. Despite the need for economy, most autotuning techniques still require large budgets of costly experimental measurements to provide good results, while rarely providing exploitable knowledge after optimization. In [16], we investigate the use of *Design of Experiments* to propose a user-transparent autotuning technique that operates under tight budget constraints by significantly reducing the measurements needed to find good optimizations. Our approach enables users to make informed decisions on which optimizations to pursue and when to stop. We present an experimental evaluation of our approach and show it is capable of leveraging user decisions to find the best global configuration of a GPU Laplacian kernel using half of the measurement budget used by other common autotuning techniques. We show that our approach is also capable of finding speedups of up to $50\times$, compared to gcc's -O3, for some kernels from the SPAPT benchmark suite, using up to $10\times$ fewer measurements than random sampling. Although the results are very encouraging, our approach relies on assumptions on the geometry of the search space that are difficult to test in very large dimension. We are thus currently pursuing this line of research using non parametric approaches based on gaussian process regression, space filling designs and iteratively selecting configurations that yield the best expected improvement.

Our second contribution is related to the control of measurements. In [40], we relate a surprising observation on the performance of the highly optimized and regular DGEMM function on modern processors. The DGEMM function is a widely used implementation of the matrix product. While the asymptotic complexity of the algorithm only depends on the sizes of the matrices, we show that the performance is significantly impacted by the matrices content. Although it would be expected that special values like 1 or 0 may yield to specific behavior, we show that arbitrary constant values are no different and that random values incur a significant performance drop. Our experiments show that this may be due to bit flips in the CPU causing an energy consumption overhead. Such phenomenon reminds the importance of thoroughly randomizing every single parameter of experiments to avoid bias toward specific behavior.

7.2. Predictive Simulation of HPC Applications

Finely tuning MPI applications (number of processes, granularity, collective operation algorithms, topology and process placement) is critical to obtain good performance on supercomputers. With a rising cost of modern supercomputers, running parallel applications at scale solely to optimize their performance is extremely expensive. Using SimGrid, we work toward providing a methodology allowing to provide inexpensive but faithful predictions of expected performance.

The methodology we propose relies on SimGrid/SMPI and captures the complexity of adaptive applications by emulating the MPI code while skipping insignificant parts. In [18] we demonstrate its capability with High Performance Linpack (HPL), the benchmark used to rank supercomputers in the TOP500 and which requires a careful tuning. We explain (1) how we both extended the SimGrid's SMPI simulator and slightly modified the open-source version of HPL to allow a fast emulation on a single commodity server at the scale of a supercomputer and (2) how to model the different components (network, BLAS, ...) of the system. We show that a careful modeling of both spatial and temporal node variability allows us to obtain predictions within a few percents of real experiments. The modeling of BLAS operations is particularly important and we have thus started investigating in the context of simulating a sparse direct solver how to automatically performance models for commonly used BLAS kernels [33]. A key difficulty remains the acquisition of faithful performance measurements as modern processors are often quite unstable. This effort is therefore particularly related to the aforementioned "Design of Experiments" line of research.

7.3. Simulation of Smart Grids

In [35], we present ASGriDS, an asynchronous Smart Grid simulation framework. ASGriDS is multi-domain, it simultaneously models the power network along with its physical loads/generators, controllers, and communication infrastructure. ASGriDS provides a unified workflow in a pythonic environment, to describe, run and control complex SmartGrid deployment scenarios. ASGriDS is an event-driven simulator that can run in either real-time or accelerated real-time. As it is modular and its components interact asynchronously, it can run either locally on a distributed infrastructure, also in hardware-in-the- loop setups, and on top of emulated/physical communication links. In this paper, we present the design of our simulator and we demonstrate its use with a generation control problem on a low voltage network. We use ASGriDS to deploy a real-time controller based on optimal power flow, on top of TCP and UDP based communication network, under various packet loss conditions.

7.4. Batch Scheduling

Despite the impressive growth and size of super-computers, the computational power they provide still cannot match the demand. Efficient and fair resource allocation is a critical task. Super-computers use Resource and Job Management Systems to schedule applications, which is generally done by relying on generic index policies such as First Come First Served and Shortest Processing time First in combination with Backfilling strategies. Unfortunately, such generic policies often fail to exploit specific characteristics of real workloads.

In [36], we focus on improving the performance of online schedulers by studying mixed policies, which are created by combining multiple job characteristics in a weighted linear expression, as opposed to classical pure policies which use only a single characteristic. This larger class of scheduling policies aims at providing more flexibility and adaptability. We use space coverage and black-box optimization techniques to explore this new space of mixed policies and we study how can they adapt to the changes in the workload. We perform an extensive experimental campaign through which we show that (1) the best pure policy is far from optimal and that (2) using a carefully tuned mixed policy would allow to significantly improve the performance of the system. (3) We also provide empirical evidence that there is no one size fits all policy, by showing that the rapid workload evolution seems to prevent classical online learning algorithms from being effective.

A careful investigation of why such mixed strategy fail to globally exploit weekly workload features reveal that some users sometimes provide widely inaccurate information, which dramatically fools the batch scheduling heuristic. Indeed, users typically provide a loose upper bound estimate for job execution times that are hardly useful. Previous studies attempted to improve these estimates using regression techniques. Although these attempts provide reasonable predictions, they require a long period of training data. Furthermore, aiming for perfect prediction may be of limited use for scheduling purposes. In [50], we propose a simpler approach by classifying jobs as small or large and prioritizing the execution of small jobs over large ones. Indeed, small jobs are the most impacted by queuing delays but they typically represent a light load and incur a small burden on the other jobs. The classifier operates online and learns by using data collected over the previous weeks, facilitating its deployment and enabling fast adaptations to changes in workload characteristics. We evaluate our approach using four scheduling policies on six HPC platform workload traces. We show that: (i) incorporating such classification significantly reduces the average bounded slowdown of jobs in all scenarios, and (ii) the obtained improvements are comparable, in most scenarios, to the ideal hypothetical situation where the scheduler would know the exact running time of jobs in advance.

7.5. Load Balancing

In distributed systems, load balancing is a powerful concept to improve the distribution of jobs across multiple computing resources and to control performance metrics such as delays and throughputs while avoiding the overload of any single resource. This section describes three contributions:

- In multi-server distributed queueing systems, the access of stochastically arriving jobs to resources is often regulated by a dispatcher, also known as load balancer. A fundamental problem consists in designing a load balancing algorithm that minimizes the delays experienced by jobs. During the last two decades, the power-of-d-choice algorithm, based on the idea of dispatching each job to the least loaded server out of d servers randomly sampled at the arrival of the job itself, has emerged as a breakthrough in the foundations of this area due to its versatility and appealing asymptotic properties. In [8], we consider the power-of-d-choice algorithm with the addition of a local memory that keeps track of the latest observations collected over time on the sampled servers. Then, each job is sent to a server with the lowest observation. We show that this algorithm is asymptotically optimal in the sense that the load balancer can always assign each job to an idle server in the large-system limit. This holds true if and only if the system load λ is less than $1 - 1/d$. If this condition is not satisfied, we show that queue lengths are tightly bounded by $\lceil \frac{-\log(1-\lambda)}{\log(\lambda d+1)} \rceil$. This is in contrast with the classic version of the power-of-d-choice algorithm, where at the fluid scale a strictly positive proportion of servers containing i jobs exists for all $i \geq 0$, in equilibrium. Our results quantify and highlight the importance of using memory as a means to enhance performance in randomized load balancing.
- When dispatching jobs to parallel servers, or queues, the highly scalable round-robin (RR) scheme reduces the variance of interarrival times at all queues to a great extent but has no impact on the variances of service processes. Contrariwise, size-interval task assignment (SITA) routing has little impact on the variances of interarrival times but makes the service processes as deterministic as possible. In [6], we unify both 'static' approaches to design a scalable load balancing framework able to control the variances of the arrival and service processes jointly. It turns out that the resulting combination significantly improves performance and is able to drive the mean job delay to zero in the large-system limit; it is known that this property is not achieved when both approaches are considered separately. Within realistic parameters, we show that the optimal number of size intervals that partition the support of the job size distribution is small with respect to the system size. This enhances the applicability of the proposed load balancing scheme at a large scale. In fact, we find that adding a little bit of information about job sizes to a dispatcher operating under RR improves performance a lot. Under the optimal scaling of size intervals and assuming highly variable job sizes, numerical simulations indicate that the proposed algorithm is competitive with the (less scalable) join-the-shortest-workload algorithm even when the system size grows large.
- Size-based routing provides robust strategies to improve the performance of computer and communication systems with highly variable workloads because it is able to isolate small jobs from large ones in a static manner. The basic idea is that each server is assigned all jobs whose sizes belong to a distinct and continuous interval. In the literature, dispatching rules of this type are referred to as SITA (Size Interval Task Assignment) policies. Though their evident benefits, the problem of finding a SITA policy that minimizes the overall mean (steady-state) waiting time is known to be intractable. In particular it is not clear when it is preferable to balance or unbalance server loads and, in the latter case, how. In [7], we provide an answer to these questions in the celebrated limiting regime where the system capacity grows linearly with the system demand to infinity. Within this framework, we prove that the minimum mean waiting time achievable by a SITA policy necessarily converges to the mean waiting time achieved by SITA-E, the SITA policy that equalizes server loads, provided that servers are homogeneous. However, within the set of SITA policies we also show that SITA-E can perform arbitrarily bad if servers are heterogeneous. In this case we prove that there exist exactly C! asymptotically optimal policies, where C denotes the number of server types, and all of them are linked to the solution of a single strictly convex optimization problem. It turns out that the mean

waiting time achieved by any of such asymptotically optimal policies does not depend on how job-size intervals are mapped to servers. Our theoretical results are validated by numerical simulations with respect to realistic parameters and suggest that the above insights are also accurate in small systems composed of a few servers, i.e., ten.

7.6. FoG Computing

To this day, the Internet of Things (IoT) continues its explosive growth. Nevertheless, with the exceptional evolution of traffic demand, existing infrastructures are struggling to resist. In this context, Fog computing is shaping the future of IoT applications. It offers nearby computational, networking and storage resources to respond to the stringent requirements of these applications. However, despite its several advantages, Fog computing raises new challenges which slow its adoption down. Hence, there is a lack of practical solutions to enable the exploitation of this novel concept.

In [19], we propose FITOR, an orchestration system for IoT applications in the Fog environment. This solution builds a realistic Fog environment while offering efficient orchestration mechanisms. In order to optimize the provisioning of Fog-Enabled IoT applications, FITOR relies on O-FSP, an optimized fog service provisioning strategy which aims to minimize the provisioning cost of IoT applications, while meeting their requirements. Based on extensive experiments, the results obtained show that O-FSP optimizes the placement of IoT applications and outperforms the related strategies in terms of i) provisioning cost ii) resource usage and iii) acceptance rate. In [46], we propose a novel strategy, which we call GO-FSP and which optimizes the placement of IoT application components while coping with their strict performance requirements. To do so, we first propose an Integer Linear Programming (ILP) formulation for the IoT application provisioning problem. The latter targets to minimize the deployment cost while ensuring a load balancing between heterogeneous devices. Then, a GRASP-based approach is proposed to achieve the aforementioned objectives. Finally, we make use of the FITOR orchestration system to evaluate the performance of our solution under real conditions. Obtained results show that our scheme outperforms the related strategies. We are currently comparing such strategy with other strategies based on online learning mechanisms under various information scenarios (delayed and noisy feedback, inaccurate application load information, etc.).

Last, fog computing also extends the capacities of the cloud to the edge of the network, near the physical world, so that Internet of Things (IoT) applications can benefit from properties such as short delays, real-time and privacy. Unfortunately, devices in the Fog-IoT environment are usually unstable and prone to failures. In this context, the consequences of failures may impact the physical world and can, therefore, be critical. In [28], we present a framework for end-to-end resilience of Fog-IoT applications. The framework was implemented and experimented on a smart home testbed.

7.7. Research Management: Research Reproducibility and Credit

We are actively promoting better research practices, in particular in term of research reproducibility and contribution recognition. Our contribution this year is threefold

First, we have participated to the writing of a book introducing reproducible research [39]. For a researcher, there is nothing more frustrating than the failure to reproduce major results obtained a few months back. The causes of such disappointments can be multiple and insidious. This phenomenon plays an important role in the so-called "research reproducibility crisis". This book takes a current perspective onto a number of potentially dangerous situations and practices, to examplify and highlight the symptoms of non-reproducibility in research. Each time, it provides efficient solutions ranging from good-practices that are easily and immediately implementable to more technical tools, all of which are free and have been put to the test by the authors themselves. Students and engineers and researchers should find efficient and accessible ways leading them to improve their reproducible research practices.

Second, to allow students and engineers and researchers to receive proper training in reproducible research, we have run the second session of the Mooc "Reproducible research: Methodological principles for a transparent science" on the FUN platform from April, 1 to June, 13 2019. This MOOC allows scientists to learn modern and reliable tools such as Markdown for taking structured notes, Desktop search applications, GitLab for version control and collaborative working, and Computational notebooks (Jupyter, RStudio, and Org-Mode) for efficiently combining the computation, presentation, and analysis of data. More than 2,100 persons registered to this session and we are currently working on a third session which is expected to start in the beginning of the year 2020.

Third, software is a fundamental pillar of modern scientific research, not only in computer science, but actually across all fields and disciplines. However, there is a lack of adequate means to cite and reference software, for many reasons. An obvious first reason is software authorship, which can range from a single developer to a whole team, and can even vary in time. The panorama is even more complex than that, because many roles can be involved in software development: software architect, coder, debugger, tester, team manager, and so on. Arguably, the researchers who have invented the key algorithms underlying the software can also claim a part of the authorship. And there are many other reasons that make this issue complex. We provide in [5] a contribution to the ongoing efforts to develop proper guidelines and recommendations for software citation, building upon the internal experience of Inria, the French research institute for digital sciences. As a central contribution, we make three key recommendations. (1) We propose a richer taxonomy for software contributions with a qualitative scale. (2) We claim that it is essential to put the human at the heart of the evaluation. And (3) we propose to distinguish citation from reference which is particularly important in the context of reproducible research.

7.8. Mean Field Games and Control

In [10], we consider mean field games with discrete state spaces (called discrete mean field games in the following) and we analyze these games in continuous and discrete time, over finite as well as infinite time horizons. We prove the existence of a mean field equilibrium assuming continuity of the cost and of the drift. These conditions are more general than the existing papers studying finite state space mean field games. Besides, we also study the convergence of the equilibria of N -player games to mean field equilibria in our four settings. On the one hand, we define a class of strategies in which any sequence of equilibria of the finite games converges weakly to a mean field equilibrium when the number of players goes to infinity. On the other hand, we exhibit equilibria outside this class that do not converge to mean field equilibria and for which the value of the game does not converge. In discrete time this non-convergence phenomenon implies that the Folk theorem does not scale to the mean field limit.

In [20], we consider a class of nonlinear systems of differential equations with uncertainties, i.e., with lack of knowledge in some of the parameters that is represented by a time-varying unknown bounded functions. An under-approximation of such systems consists of a subset of its reachable set, for any value of the unknown parameters. By relying on optimal control theory through Pontryagin's principle, we provide an algorithm for the under-approximation of a linear combination of the state variables in terms of a fully automated tool-chain named UTOPIC. This allows to establish tight under-approximations of common benchmarks models with dimensions as large as sixty-five.

7.9. Energy and Network Optimization

This section describes four contributions on energy and network optimization.

- One of the key challenges in Internet of Things (IoT) networks is to connect many different types of autonomous devices while reducing their individual power consumption. This problem is exacerbated by two main factors: first, the fact that these devices operate in and give rise to a highly dynamic and unpredictable environment where existing solutions (e.g., water-filling algorithms) are no longer relevant; and second, the lack of sufficient information at the device end. To address these issues, we propose a regret-based formulation that accounts for arbitrary network dynamics: this allows us to derive an online power control scheme that is provably capable of adapting to such changes, while relying solely on strictly causal feedback. In so doing, we identify an important tradeoff between the amount of feedback available at the transmitter side and the resulting system performance: if the device has access to unbiased gradient observations, the algorithm's regret after T stages is $O(T^{-1/2})$ (up to logarithmic factors); on the other hand, if the device only has access to scalar, utility-based information, this decay rate drops to $O(T^{-1/4})$. The above is validated by an extensive suite of numerical simulations in realistic channel conditions, which clearly exhibit the gains of the proposed online approach over traditional water-filling methods. This contribution appeared in [11].

- Many businesses possess a small infrastructure that they can use for their computing tasks, but also often buy extra computing resources from clouds. Cloud vendors such as Amazon EC2 offer two types of purchase options: on-demand and spot instances. As tenants have limited budgets to satisfy their computing needs, it is crucial for them to determine how to purchase different options and utilize them (in addition to possible self-owned instances) in a cost-effective manner while respecting their response-time targets. In this paper, we propose a framework to design policies to allocate self-owned, on-demand and spot instances to arriving jobs. In particular, we propose a near-optimal policy to determine the number of self-owned instances and an optimal policy to determine the number of on-demand instances to buy and the number of spot instances to bid for at each time unit. Our policies rely on a small number of parameters and we use an online learning technique to infer their optimal values. Through numerical simulations, we show the effectiveness of our proposed policies, in particular that they achieve a cost reduction of up to 64.51% when spot and on-demand instances are considered and of up to 43.74% when self-owned instances are considered, compared to previously proposed or intuitive policies. This contribution appeared in [13].
- In [22], we consider the classical problem of minimizing offline the total energy consumption required to execute a set of n real-time jobs on a single processor with varying speed. Each real-time job is defined by its release time, size, and deadline (all integers). The goal is to find a sequence of processor speeds, chosen among a finite set of available speeds, such that no job misses its deadline and the energy consumption is minimal. Such a sequence is called an optimal speed schedule. We propose a linear time algorithm that checks the schedulability of the given set of n jobs and computes an optimal speed schedule. The time complexity of our algorithm is in $O(n)$, to be compared with $O(n \log(n))$ for the best known solutions. Besides the complexity gain, the main interest of our algorithm is that it is based on a completely different idea: instead of computing the critical intervals, it sweeps the set of jobs and uses a dynamic programming approach to compute an optimal speed schedule. Our linear time algorithm is still valid (with some changes) with an arbitrary power function (not necessarily convex) and arbitrary switching times
- Network utility maximization (NUM) is an iconic problem in network traffic management which is at the core of many current and emerging network design paradigms - and, in particular, software-defined networks (SDNs). Thus, given the exponential growth of modern-day networks (in both size and complexity), it is crucial to develop scalable algorithmic tools that are capable of providing efficient solutions in time which is dimension-free, i.e., independent-or nearly-independent-on the size of the system. To do so, we leverage a suite of modified gradient methods known as "mirror descent" and we derive a scalable and efficient algorithm for the NUM problem based on gradient exponentiation. We show that the convergence speed of the proposed algorithm only carries a logarithmic dependence on the size of the network, so it can be implemented reliably and efficiently in massively large networks where traditional gradient methods are prohibitively slow. These theoretical results are sub-sequently validated by extensive numerical simulations showing an improvement of several order of magnitudes over standard gradient methods in large-scale networks. This contribution appeared in [31].
- In the DNS resolution process, packet losses and ensuing retransmission timeouts induce marked latencies: the current UDP-based resolution process takes up to 5 seconds to detect a loss event. In [24], [24], we find that persistent DNS connections based on TCP or TLS can provide an elegant solution to this problem. With controlled experiments on a testbed, we show that persistent DNS

connections significantly reduces worst-case latency. We then leverage a large-scale platform to study the performance impact of TCP/TLS on recursive resolvers. We find that off-the-shelf software and reasonably powerful hardware can effectively provide recursive DNS service over TCP and TLS, with a manageable performance hit compared to UDP.

7.10. Privacy, Fairness, and Transparency in Online Social Medias

This section describes four contributions on privacy, fairness and transparency in online social medias

- The Facebook advertising platform has been subject to a number of controversies in the past years regarding privacy violations, lack of transparency, as well as its capacity to be used by dishonest actors for discrimination or propaganda. In this study, we aim to provide a better understanding of the Facebook advertising ecosystem, focusing on how it is being used by advertisers. We first analyze the set of advertisers and then investigate how those advertisers are targeting users and customizing ads via the platform. Our analysis is based on the data we collected from over 600 real-world users via a browser extension that collects the ads our users receive when they browse their Facebook timeline, as well as the explanations for why users received these ads. Our results reveal that users are targeted by a wide range of advertisers (e.g., from popular to niche advertisers); that a non-negligible fraction of advertisers are part of potentially sensitive categories such as news and politics, health or religion; that a significant number of advertisers employ targeting strategies that could be either invasive or opaque; and that many advertisers use a variety of targeting parameters and ad texts. Overall, our work emphasizes the need for better mechanisms to audit ads and advertisers in social media and provides an overview of the platform usage that can help move towards such mechanisms.

  This contribution appeared in [14].

- To help their users to discover important items at a particular time, major websites like Twitter, Yelp, TripAdvisor or NYTimes provide Top-K recommendations (e.g., 10 Trending Topics, Top 5 Hotels in Paris or 10 Most Viewed News Stories), which rely on crowd-sourced popularity signals to select the items. However, diferent sections of a crowd may have diferent preferences, and there is a large silent majority who do not explicitly express their opinion. Also, the crowd often consists of actors like bots, spammers, or people running orchestrated campaigns. Recommendation algorithms today largely do not consider such nuances, hence are vulnerable to strategic manipulation by small but hyper-active user groups. To fairly aggregate the preferences of all users while recommending top-K items, we borrow ideas from prior research on social choice theory, and identify a voting mechanism called Single Trans-ferable Vote (STV) as having many of the fairness properties we desire in top-K item (s)elections. We develop an innovative mechanism to attribute preferences of silent majority which also make STV completely operational. We show the generalizability of our approach by implementing it on two diferent real-world datasets. Through extensive experimentation and comparison with state-of-the-art techniques, we show that our proposed approach provides maximum user satisfaction, and cuts down drastically on items disliked by most but hyper-actively promoted by a few users.

  This contribution appeared in [17].

- The rise of algorithmic decision making led to active researches on how to define and guarantee fairness, mostly focusing on one-shot decision making. In several important applications such as hiring, however, decisions are made in multiple stage with additional information at each stage. In such cases, fairness issues remain poorly understood. In this paper we study fairness in k-stage selection problems where additional features are observed at every stage. We first introduce two fairness notions, local (per stage) and global (final stage) fairness, that extend the classical fairness notions to the k-stage setting. We propose a simple model based on a probabilistic formulation and show that the locally and globally fair selections that maximize precision can be computed via a linear program. We then define the price of local fairness to measure the loss of precision induced by local constraints; and investigate theoretically and empirically this quantity. In particular, our experiments show that the price of local fairness is generally smaller when the sensitive attribute is observed at the first stage; but globally fair selections are more locally fair when the sensitive attribute is observed at the second stage—hence in both cases it is often possible to have a selection that has a small price of local fairness and is close to locally fair.

This contribution appeared in [21].

- Most social platforms offer mechanisms allowing users to delete their posts, and a significant fraction of users exercise this right to be forgotten. However, ironically, users' attempt to reduce attention to sensitive posts via deletion, in practice, attracts unwanted attention from stalkers specifically to those (deleted) posts. Thus, deletions may leave users more vulnerable to attacks on their privacy in general. Users hoping to make their posts forgotten face a "damned if I do, damned if I don't" dilemma. Many are shifting towards ephemeral social platform like Snapchat, which will deprive us of important user-data archival. In the form of intermittent withdrawals, we present, Lethe, a novel solution to this problem of (really) forgetting the forgotten. If the next-generation social platforms are willing to give up the uninterrupted availability of non-deleted posts by a very small fraction, Lethe provides privacy to the deleted posts over long durations. In presence of Lethe, an adversarial observer becomes unsure if some posts are permanently deleted or just temporarily withdrawn by Lethe; at the same time, the adversarial observer is overwhelmed by a large number of falsely flagged un-deleted posts. To demonstrate the feasibility and performance of Lethe, we analyze large-scale real data about users' deletion over Twitter and thoroughly investigate how to choose time duration distributions for alternating between temporary withdrawals and resurrections of non-deleted posts. We find a favorable trade-off between privacy, availability and adversarial overhead in different settings for users exercising their right to delete. We show that, even against an ultimate adversary with an uninterrupted access to the entire platform, Lethe offers deletion privacy for up to 3 months from the time of deletion, while maintaining content availability as high as 95% and keeping the adversarial precision to 20%.

  This contribution appeared in [27],

7.11. Optimization Methods

This section describes six contributions on optimization.

- In [9], we propose an interior-point method for linearly constrained – and possibly nonconvex – optimization problems. The proposed method – which we call the Hessian barrier algorithm (HBA) – combines a forward Euler discretization of Hessian Riemannian gradient flows with an Armijo backtracking step-size policy. In this way, HBA can be seen as an alternative to mirror descent (MD), and contains as special cases the affine scaling algorithm, regularized Newton processes, and several other iterative solution methods. Our main result is that, modulo a non-degeneracy condition, the algorithm converges to the problem's critical set; hence, in the convex case, the algorithm converges globally to the problem's minimum set. In the case of linearly constrained quadratic programs (not necessarily convex), we also show that the method's convergence rate is $O(1/k^{\rho})$ for some $\rho \in (0, 1]$ that depends only on the choice of kernel function (i.e., not on the problem's primitives). These theoretical results are validated by numerical experiments in standard non-convex test functions and large-scale traffic assignment problems.
- In[15], Lipschitz continuity is a central requirement for achieving the optimal $O(1/T)$ rate of convergence in monotone, deterministic variational inequalities (a setting that includes convex minimization, convex-concave optimization, nonatomic games, and many other problems). However, in many cases of practical interest, the operator defining the variational inequality may exhibit singularities at the boundary of the feasible region, precluding in this way the use of fast gradient methods that attain this optimal rate (such as Nemirovski's mirror-prox algorithm and its variants). To address this issue, we propose a novel regularity condition which we call Bregman continuity, and which relates the variation of the operator to that of a suitably chosen Bregman function. Leveraging this condition, we derive an adaptive mirror-prox algorithm which attains the optimal $O(1/T)$ rate of convergence in problems with possibly singular operators, without any prior knowledge of the

degree of smoothness (the Bregman analogue of the Lipschitz constant). We also show that, under Bregman continuity, the mirror-prox algorithm achieves a $O(1/\sqrt{T})$ convergence rate in stochastic variational inequalities.

- In [23] Variational inequalities have recently attracted considerable interest in machine learning as a flexible paradigm for models that go beyond ordinary loss function minimization (such as generative adversarial networks and related deep learning systems). In this setting, the optimal O(1/t) convergence rate for solving smooth monotone variational inequalities is achieved by the Extra-Gradient (EG) algorithm and its variants. Aiming to alleviate the cost of an extra gradient step per iteration (which can become quite substantial in deep learning applications), several algorithms have been proposed as surrogates to Extra-Gradient with a *single* oracle call per iteration. In this paper, we develop a synthetic view of such algorithms, and we complement the existing literature by showing that they retain a O(1/t) ergodic convergence rate in smooth, deterministic problems. Subsequently, beyond the monotone deterministic case, we also show that the last iterate of single-call, *stochastic* extra-gradient methods still enjoys a O(1/t) local convergence rate to solutions of non-monotone variational inequalities that satisfy a second-order sufficient condition.
- In [25], we study a class of online convex optimization problems with long-term budget constraints that arise naturally as reliability guarantees or total consumption constraints. In this general setting, prior work by Mannor et al. (2009) has shown that achieving no regret is impossible if the functions defining the agent's budget are chosen by an adversary. To overcome this obstacle, we refine the agent's regret metric by introducing the notion of a "K-benchmark", i.e., a comparator which meets the problem's allotted budget over any window of length K. The impossibility analysis of Mannor et al. (2009) is recovered when $K = T$; however, for $K = o(T)$, we show that it is possible to minimize regret while still meeting the problem's long-term budget constraints. We achieve this via an online learning policy based on Cautious Online Lagrangiant Descent (COLD) for which we derive explicit bounds, in terms of both the incurred regret and the residual budget violations.
- In [26], owing to their connection with generative adversarial networks (GANs), saddle-point problems have recently attracted considerable interest in machine learning and beyond. By necessity, most theoretical guarantees revolve around convex-concave (or even linear) problems; however, making theoretical inroads towards efficient GAN training depends crucially on moving beyond this classic framework. To make piecemeal progress along these lines, we analyze the behavior of mirror descent (MD) in a class of non-monotone problems whose solutions coincide with those of a naturally associated variational inequality - a property which we call coherence. We first show that ordinary, "vanilla" MD converges under a strict version of this condition, but not otherwise; in particular, it may fail to converge even in bilinear models with a unique solution. We then show that this deficiency is mitigated by optimism: by taking an "extra-gradient" step, optimistic mirror descent (OMD) converges in all coherent problems. Our analysis generalizes and extends the results of Daskalakis et al. (2018) for optimistic gradient descent (OGD) in bilinear problems, and makes concrete headway for establishing convergence beyond convex-concave games. We also provide stochastic analogues of these results, and we validate our analysis by numerical experiments in a wide array of GAN models (including Gaussian mixture models, as well as the CelebA and CIFAR-10 datasets).
- In [30], we develop a new stochastic algorithm with variance reduction for solving pseudo-monotone stochastic variational inequalities. Our method builds on Tseng's forward-backward-forward algorithm, which is known in the deterministic literature to be a valuable alternative to Korpelevich's extragradient method when solving variational inequalities over a convex and closed set governed with pseudo-monotone and Lipschitz continuous operators. The main computational advantage of Tseng's algorithm is that it relies only on a single projection step, and two independent queries of a stochastic oracle. Our algorithm incorporates a variance reduction mechanism, and leads to a.s. convergence to solutions of a merely pseudo-monotone stochastic variational inequality problem. To the best of our knowledge, this is the first stochastic algorithm achieving this by using only a single projection at each iteration.

7.12. Learning

This section describes three contributions on machine learning.

- In [12], we examine the convergence of no-regret learning in games with continuous action sets. For concreteness, we focus on learning via "dual averaging", a widely used class of no-regret learning schemes where players take small steps along their individual payoff gradients and then "mirror" the output back to their action sets. In terms of feedback, we assume that players can only estimate their payoff gradients up to a zero-mean error with bounded variance. To study the convergence of the induced sequence of play, we introduce the notion of variational stability, and we show that stable equilibria are locally attracting with high probability whereas globally stable equilibria are globally attracting with probability 1. We also discuss some applications to mixed-strategy learning in finite games, and we provide explicit estimates of the method's convergence speed.
- Resource allocation games such as the famous Colonel Blotto (CB) and Hide-and-Seek (HS) games are often used to model a large variety of practical problems, but only in their one-shot versions. Indeed, due to their extremely large strategy space, it remains an open question how one can efficiently learn in these games. In this work, we show that the online CB and HS games can be cast as path planning problems with side-observations (SOPPP): at each stage, a learner chooses a path on a directed acyclic graph and suffers the sum of losses that are adversarially assigned to the corresponding edges; and she then receives semi-bandit feedback with side-observations (i.e., she observes the losses on the chosen edges plus some others). We propose a novel algorithm, EXP3-OE, the first-of-its-kind with guaranteed efficient running time for SOPPP without requiring any auxiliary oracle. We provide an expected-regret bound of EXP3-OE in SOPPP matching the order of the best benchmark in the literature. Moreover, we introduce additional assumptions on the observability model under which we can further improve the regret bounds of EXP3-OE. We illustrate the benefit of using EXP3-OE in SOPPP by applying it to the online CB and HS games.

  This contribution appeared in [29], [49]. In an earlier article [38], we also looked at the sequential Colonel Blotto game under bandit feedback and we proposed a blackbox optimization based method to optimize the exploration distribution of the classical ComBand algorithm.
- In [32], we study nonzero-sum hypothesis testing games that arise in the context of adversarial classification, in both the Bayesian as well as the Neyman-Pearson frameworks. We first show that these games admit mixed strategy Nash equilibria, and then we examine some interesting concentration phenomena of these equilibria. Our main results are on the exponential rates of convergence of classification errors at equilibrium, which are analogous to the well-known Chernoff-Stein lemma and Chernoff information that describe the error exponents in the classical binary hypothesis testing problem, but with parameters derived from the adversarial model. The results are validated through numerical experiments.

ROMA Project-Team

7. New Results

7.1. Creation of the start-up “Mumps Technologies SAS”

In January 2019, Jean-Yves L’Excellent left the ROMA team to co-found with Patrick Amestoy and Chiara Puglisi the company “Mumps Technologies” around the free software library MUMPS (Cecill-C licence). MUMPS solves large systems of sparse linear equations on high-performance computers in a robust and effective way. Mumps Technologies carries on collaborations and R&D activities to keep the MUMPS software library state-of-the-art and freely available, while offering to its clients a set of services.

7.2. Scheduling independent stochastic tasks under deadline and budget constraints

This work discusses scheduling strategies for the problem of maximizing the expected number of tasks that can be executed on a cloud platform within a given budget and under a deadline constraint. The execution times of tasks follow IID probability laws. The main questions are how many processors to enroll and whether and when to interrupt tasks that have been executing for some time. We provide complexity results and an asymptotically optimal strategy for the problem instance with discrete probability distributions and without deadline. We extend the latter strategy for the general case with continuous distributions and a deadline and we design an efficient heuristic which is shown to outperform standard approaches when running simulations for a variety of useful distribution laws.

The findings were published in a journal [8].

7.3. Online scheduling of task graphs on heterogeneous platforms

Modern computing platforms commonly include accelerators. We target the problem of scheduling applications modeled as task graphs on hybrid platforms made of two types of resources, such as CPUs and GPUs. We consider that task graphs are uncovered dynamically, and that the scheduler has information only on the available tasks, i.e., tasks whose predecessors have all been completed. Each task can be processed by either a CPU or a GPU, and the corresponding processing times are known. Our study extends a previous $4\sqrt{m/k}$-competitive online algorithm by Amaris et al. [46], where m is the number of CPUs and k the number of GPUs ($m \geq k$). We prove that no online algorithm can have a competitive ratio smaller than $\sqrt{m/k}$. We also study how adding flexibility on task processing, such as task migration or spoliation, or increasing the knowledge of the scheduler by providing it with information on the task graph, influences the lower bound. We provide a $(2\sqrt{m/k}+1)$-competitive algorithm as well as a tunable combination of a system-oriented heuristic and a competitive algorithm; this combination performs well in practice and has a competitive ratio in $\Theta(\sqrt{m/k})$. We also adapt all our results to the case of multiple types of processors. Finally, simulations on different sets of task graphs illustrate how the instance properties impact the performance of the studied algorithms and show that our proposed tunable algorithm performs the best among the online algorithms in almost all cases and has even performance close to an offline algorithm.

The findings were published in a journal [9].

7.4. A generic approach to scheduling and checkpointing workflows

This work deals with scheduling and checkpointing strategies to execute scientific workflows on failure-prone large-scale platforms. To the best of our knowledge, this work is the first to target fail-stop errors for arbitrary workflows. Most previous work addresses soft errors, which corrupt the task being executed by a processor but do not cause the entire memory of that processor to be lost, contrarily to fail-stop errors. We revisit classical mapping heuristics such as HEFT and MINMIN and complement them with several checkpointing strategies. The objective is to derive an efficient trade-off between checkpointing every task (CKPTALL), which is an overkill when failures are rare events, and checkpointing no task (CKPTNONE), which induces dramatic re-execution overhead even when only a few failures strike during execution. Contrarily to previous work, our approach applies to arbitrary workflows, not just special classes of dependence graphs such as M-SPGS (Minimal Series-Parallel Graphs). Extensive experiments report significant gain over both CKPTALL and CKPTNONE, for a wide variety of workflows.

The findings were published in a journal [10].

7.5. Limiting the memory footprint when dynamically scheduling DAGs on shared-memory platforms

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

The findings were published in a journal [12].

7.6. Scheduling independent stochastic tasks on heterogeneous cloud platforms

This work introduces scheduling strategies to maximize the expected number of independent tasks that can be executed on a cloud platform within a given budget and under a deadline constraint. The cloud platform is composed of several types of virtual machines (VMs), where each type has a unit execution cost that depends upon its characteristics. The amount of budget spent during the execution of a task on a given VM is the product of its execution length by the unit execution cost of that VM. The execution lengths of tasks follow a variety of standard probability distributions (exponential, uniform, half-normal, etc.), which is known beforehand and whose mean and standard deviation both depend upon the VM type. Finally, there is a global available budget and a deadline constraint, and the goal is to successfully execute as many tasks as possible before the deadline is reached or the budget is exhausted (whichever comes first). On each VM, the scheduler can decide at any instant to interrupt the execution of a (long) running task and to launch a new one, but the budget already spent for the interrupted task is lost. The main questions are which VMs to enroll, and whether and when to interrupt tasks that have been executing for some time. We assess the complexity of the problem by showing its NP-completeness and providing a 2-approximation for the asymptotic case where budget and deadline both tend to infinity. Then we introduce several heuristics and compare their performance by running an extensive set of simulations.

This work has been presented at the Cluster 2019 conference [17].

7.7. Improved energy-aware strategies for periodic real-time tasks under reliability constraints

This work revisited the real-time scheduling problem recently introduced by Haque, Aydin and Zhu [62]. In this challenging problem, task redundancy ensures a given level of reliability while incurring a significant energy cost. By carefully setting processing frequencies, allocating tasks to processors and ordering task executions, we improve on the previous state-of-the-art approach with an average gain in energy of 20%. Furthermore, we establish the first complexity results for specific instances of the problem.

This work has been accepted at the RTSS 2019 conference [18].

7.8. Multilevel algorithms for acyclic partitioning of directed acyclic graphs

We investigate the problem of partitioning the vertices of a directed acyclic graph into a given number of parts. The objective function is to minimize the number or the total weight of the edges having end points in different parts, which is also known as the edge cut. The standard load balancing constraint of having an equitable partition of the vertices among the parts should be met. Furthermore, the partition is required to be acyclic; i.e., the interpart edges between the vertices from different parts should preserve an acyclic dependency structure among the parts. In this work, we adopt the multilevel approach with coarsening, initial partitioning, and refinement phases for acyclic partitioning of directed acyclic graphs. We focus on two-way partitioning (sometimes called bisection), as this scheme can be used in a recursive way for multiway partitioning. To ensure the acyclicity of the partition at all times, we propose novel and efficient coarsening and refinement heuristics. The quality of the computed acyclic partitions is assessed by computing the edge cut. We also propose effective ways to use the standard undirected graph partitioning methods in our multilevel scheme. We perform a large set of experiments on a dataset consisting of (i) graphs coming from an application and (ii) some others corresponding to matrices from a public collection. We report significant improvements compared to the current state of the art.

This work is published in a journal [11].

7.9. A multi-dimensional Morton-ordered block storage for mode-oblivious tensor computations

Computation on tensors, treated as multidimensional arrays, revolve around generalized basic linear algebra subroutines (BLAS). We propose a novel data structure in which tensors are blocked and blocks are stored in an order determined by Morton order. This is not only proposed for efficiency reasons, but also to induce efficient performance regardless of which mode a generalized BLAS call is invoked for; we coin the term mode-oblivious to describe data structures and algorithms that induce such behavior. Experiments on one of the most bandwidth-bound generalized BLAS kernel, the tensor–vector multiplication, not only demonstrate superior performance over two state-of-the-art variants by up to 18%, but additionally show that the proposed data structure induces a 71% less sample standard deviation for tensor–vector multiplication across d modes, where d varies from 2 to 10. Finally, we show our data structure naturally expands to other tensor kernels and demonstrate up to 38% higher performance for the higher-order power method.

This work is published in a journal [13]

7.10. Effective heuristics for matchings in hypergraphs

The problem of finding a maximum cardinality matching in a d-partite, d-uniform hypergraph is an important problem in combinatorial optimization and has been theoretically analyzed. We first generalize some graph matching heuristics for this problem. We then propose a novel heuristic based on tensor scaling to extend the matching via judicious hyperedge selections. Experiments on random, synthetic and real-life hypergraphs show that this new heuristic is highly practical and superior to the others on finding a matching with large cardinality.

This work is published in the proceedings of SEA^2, where it has received the best paper award [16].

7.11. Karp-Sipser based kernels for bipartite graph matching

We consider Karp–Sipser, a well known matching heuristic in the context of data reduction for the maximum cardinality matching problem. We describe an efficient implementation as well as modifications to reduce its time complexity in worst case instances, both in theory and in practical cases. We compare experimentally against its widely used simpler variant and show cases for which the full algorithm yields better performance .

This work appears in the proceedings of ALENEX2020 [20]

7.12. Efficient and effective sparse tensor reordering

This paper formalizes the problem of reordering a sparse tensor to improve the spatial and temporal locality of operations with it, and proposes two reordering algorithms for this problem, which we call BFS-MCS and Lexi-Order. The BFS-MCS method is a Breadth First Search (BFS)-like heuristic approach based on the maximum cardinality search family; Lexi-Order is an extension of doubly lexical ordering of matrices to tensors. We show the effects of these schemes within the context of a widely used tensor computation, the Candecomp/Parafac decomposition (CPD), when storing the tensor in three previously proposed sparse tensor formats: coordinate (COO), compressed sparse fiber (CSF), and hierarchical coordinate (HiCOO). A new partition-based superblock scheduling is also proposed for HiCOO format to improve load balance. On modern multicore CPUs, we show Lexi-Order obtains up to 4.14× speedup on sequential HiCOO-Mttkrp and 11.88× speedup on its parallel counterpart. The performance of COO-and CSF-based Mttkrps also improves. Our two reordering methods are more effective than state-of-the-art approaches.

This work appears in the proceedings of ICS2019 [21].

7.13. High performance tensor–vector multiplication on shared-memory systems

Tensor–vector multiplication is one of the core components in tensor computations. We have recently investigated high performance, single core implementation of this bandwidth-bound operation. Here, we investigate its efficient, shared-memory implementations. Upon carefully analyzing the design space, we implement a number of alternatives using OpenMP and compare them experimentally. Experimental results on up to 8 socket systems show near peak performance for the proposed algorithms.

This work appears in the proceedings of PPAM2019 and is supported with a technical report [22], [36].

7.14. Matrix symmetrization and sparse direct solvers

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. 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 will appear in the proceedings of CSC2020 [23].

7.15. A scalable clustering-based task scheduler for homogeneous processors using DAG partitioning

When scheduling a directed acyclic graph (DAG) of tasks on computational platforms, a good trade-off between load balance and data locality is necessary. List-based scheduling techniques are commonly used greedy approaches for this problem. The downside of list-scheduling heuristics is that they are incapable of making short-term sacrifices for the global efficiency of the schedule. In this work, we describe new list-based scheduling heuristics based on clustering for homogeneous platforms, under the realistic duplex single-port communication model. Our approach uses an acyclic partitioner for DAGs for clustering. The clustering enhances the data locality of the scheduler with a global view of the graph. Furthermore, since the partition is acyclic, we can schedule each part completely once its input tasks are ready to be executed. We present an extensive experimental evaluation showing the trade-offs between the granularity of clustering and the parallelism, and how this affects the scheduling. Furthermore, we compare our heuristics to the best state-of-the-art list-scheduling and clustering heuristics, and obtain more than three times better makespan in cases with many communications.

This work appears in the proceedings of IPDPS 2019 [25].

7.16. Improving Locality-Aware Scheduling with Acyclic Directed Graph Partitioning

We investigate efficient execution of computations, modeled as Directed Acyclic Graphs (DAGs), on a single processor with a two-level memory hierarchy, where there is a limited fast memory and a larger slower memory. Our goal is to minimize execution time by minimizing redundant data movement between fast and slow memory. We utilize a DAG partitioner that finds localized, acyclic parts of the whole computation that can fit into fast memory, and minimizes the edge cut among the parts. We propose a new scheduler that executes each part one-by-one, obeying the dependency among parts, aiming at reducing redundant data movement needed by cut-edges. Extensive experimental evaluation shows that the proposed DAG-based scheduler significantly reduces redundant data movement.

This work will appear in the proceedings of PPAM 2019 [24].

7.17. Replication Is More Efficient Than You Think

We revisit replication coupled with checkpointing for fail-stop errors. Replication enables the application to survive many fail-stop errors, thereby allowing for longer checkpointing periods. Previously published works use replication with the no-restart strategy, which never restart failed processors until the application crashes. We introduce the restart strategy where failed processors are restarted after each checkpoint. which may introduce additional overhead during checkpoints but prevents the application configuration from degrading throughout successive checkpointing periods. We show how to compute the optimal checkpointing period for this strategy, which is much larger than the one with no-restart, thereby decreasing I/O pressure. We show through simulations that using the restart strategy significantly decreases the overhead induced by replication, in terms of both total execution time and energy consumption.

This work appears in the proceedings of SC 2019 [15], [28].

7.18. Generic matrix multiplication for multi-GPU accelerated distributed-memory platforms over PaRSEC

We introduce a generic and flexible matrix-matrix multiplication algorithm $C = A \times B$ for state-of-the-art computing platforms. Typically, these platforms are distributed-memory machines whose nodes are equipped with several accelerators. To the best of our knowledge, SLATE is the only library that provides a publicly available implementation on such platforms, and it is currently limited to problem instances where the C

matrix can entirely fit in the memory of the GPU accelerators. Our algorithm relies on the classical tile-based outer-product algorithm, but enhances it with several control dependencies to increase data re-use and to optimize communication flow from/to the accelerators within each node. The algorithm is written with the PaRSEC runtime system, which allows for a fast and generic implementation, while achieving close-to-peak performance.

This work appears in the proceedings of Scala 2019 [19].

7.19. Reservation strategies for stochastic jobs

We are interested in scheduling stochastic jobs on a reservation-based platform. Specifically, we consider jobs whose execution time follows a known probability distribution. The platform is reservation-based, meaning that the user has to request fixed-length time slots. The cost then depends on both (i) the request duration (pay for what you ask); and (ii) the actual execution time of the job (pay for what you use).

A reservation strategy determines a sequence of increasing-length reservations, which are paid for until one of them allows the job to successfully complete. The goal is to minimize the total expected cost of the strategy. We provide some properties of the optimal solution, which we characterize up to the length of the first reservation. We then design several heuristics based on various approaches, including a brute-force search of the first reservation length while relying on the characterization of the optimal strategy, as well as the discretization of the target continuous probability distribution together with an optimal dynamic programming algorithm for the discrete distribution.

We evaluate these heuristics using two different platform models and cost functions: The first one targets a cloud-oriented platform (e.g., Amazon AWS) using jobs that follow a large number of usual probability distributions (e.g., Uniform, Exponential, LogNormal, Weibull, Beta), and the second one is based on interpolating traces from a real neuroscience application executed on an HPC platform. An extensive set of simulation results show the effectiveness of the proposed reservation-based approaches for scheduling stochastic jobs.

This work appears in the proceedings of IPDPS 2019 [14].

STORM Project-Team

7. New Results

7.1. Multi-Valued Expression Analysis for Collective Checking

Determining if a parallel program behaves as expected on any execution is challenging due to non-deterministic executions. Static analysis helps to detect all execution paths that can be executed concurrently by identifying multi-valued expressions, i.e. expressions evaluated differently among processes. This can be used to find collective errors in parallel programs. The PARallel COntrol flow Anomaly CHecker (PARCOACH) framework has been extended with a multi-valued expressions detection to find such errors [9]. The new analysis corrects the previous one and analyzes parallel applications using MPI, OpenMP, UPC and CUDA.

7.2. Hiding the latency of MPI communication

As developers spend significant effort performing manual latency optimizations, the goal of Van Man Nguyen Ph.D Thesis is to automatically provide maximal communication overlap for MPI communication (collective, point-to-point and RMA Put/Get operations). A method that moves operations and their completion (e.g. Isend/Wait) as far apart as possible in the program while preserving memory consistency is under development.

7.3. Performance monitoring and Steering Framework

Two frameworks were developed within the context of the project H2020 EXA2PRO to offer performance monitoring and steering APIs into the StarPU runtime system, to be targeted by external tools.

The performance monitoring framework enables StarPU to export performance counters in a generic, safe, extensible way, to give external tools access to internal metrics and statistics, such as the peak number of tasks in the dependence waiting queue, the cumulated execution time by worker thread, and the number of ready tasks of a given kind waiting for an execution slot.

The performance steering framework defines a set of runtime-actionable knobs that can be used to steer the execution of an application on top of StarPU from an external tool, with similar properties of genericity, safety and extensibility as the performance monitoring framework.

7.4. Heterogeneous task scheduling

Taking advantage of heterogeneous systems requires to carefully choose which tasks should be accelerated. Simple heuristics allow to get fairly good performance, but do not have approximation ratio that would provide performance guarantees. The ROMA Inria team designed advanced heuristics which do have approximation ratios. We have implemented one within StarPU and indeed improved the performance over existing heuristics. The implementation required to revamp part of the StarPU toolkit dedicated to writing scheduling heuristics.

7.5. Task scheduling with memory constraints

When dealing with larger and larger datasets processed by task-based applications, the amount of system memory may become too small to fit the working set, depending on the task scheduling order. The ROMA Inria team proposed a heuristic to introduce additional dependencies to the task graph enough so that any scheduling order will meet the memory constraint, while avoiding to extend the critical path length. On the other hand, banker algorithms allow to achieve this online, within the task scheduler, but do not have an overall view of the task graph, and may thus severely increase the critical path. We have thus started to design a collaboration between visionary heuristics which take a global but coarse view of the task graph, and online heuristics which have a local but precise view of the task graph.

7.6. Leveraging compiler analysis for task scheduling

Polyhedral analysis of task graph submission loops allow to get at compilation time a representation of the task graph, and perform insightful analyses thanks to the obtained overview of the whole task graph. Task scheduling heuristics, on the other hand, usually keep only a limited view over the task graph, to avoid prohibitive algorithmic costs. We have started to collaborate with the CASH Inria team to transfer some of the insights of the compiler to the compiler. We have notably made the compiler automatically compute a cut of the task graph below which the availability parallelism is lower than the capacities of the target hardware. The scheduler can then at that point switch between a heuristic which privileges task acceleration, and a heuristic which privileges the critical path. Only preliminary results have been obtained so far.

7.7. Failure Tolerance for StarPU

Since supercomputers keep growing in terms of core numbers, the reliability decreases the same way. The project H2020 EXA2PRO aims to propose solutions for the failure tolerance problem, including StarPU. While exploring decades of research about the resilience techniques, we have identified properties in our runtime's paradigm that can be exploited in order to propose a solution with lower overhead than the generic existing ones. An implementation of a solution is currently being developed for evaluation, with an interface that can be easily plugged into StarPU.

7.8. Static and Dynamic Adaptation of Task parallelism

This work is the result of Pierre Huchant PhD thesis, and the objectives are to adapt statically and dynamically OpenCL tasks. The adaptation consists in splitting tasks automatically into multiple sub-tasks, taking into account the heterogeneity of the architecture (sub-tasks are specifically created for one processing unit), the load imbalance within a parallel OpenCL, between the different iterations in space, and if the task graph is repeatedly executed, between the iterations in time, and it takes into account the time of the communications generated by splitting the tasks [2]. The method is able to cope with sequential task graphs (tasks are parallel themselves but scheduled sequentially) and deals with tasks manipulating complex data structures as shown on an N-body particle simulation mini-app.

7.9. AFF3CT

The AFF3CT library, developed jointly between IMS and the STORM team, which aims to model error correcting codes for numerical communications has been further improved in different ways. Additional new algorithms have been designed and evaluated within the AFF3CT library [6], [7], and a new approach for generating and exploring automatically high performance error correction codes from matrix description is an on-going work. Besides, an on-going work is on the automatic parallelization of the tasks describing the simulation of a whole chain of signal transmission. In order to be able to make accessible the simulation results obtained with AFF3CT and to be able to replay easily the same simulations, a web interface allowing users to browse through the results, and the simulation setup for a large range of inputs has been designed. This graphical interface is currently in use at IMS by other researchers. Finally, a call for the creation of a consortium on AFF3CT is available on the web page of AFF3CT https://aff3ct.github.io/.

7.10. Matlab API for AFF3CT

As part of the AFF3CT development action, an API compatible with the Matlab mathematical software was designed on top of the AFF3CT fast forward error correction toolbox to allow the library to be used in a high-level manner, directly from the Matlab environment. Due to the relatively large number of classes exported by AFF3CT, an automatized process was designed to let AFF3CT classes be wrapped adequately for Matlab's MEX interface to external libraries.

7.11. InKS framework

The InKS framework was developed by Ksander Ejjaaouani as part of his Ph.D Thesis co-supervised by the university of Strasbourg, the Maison de la Simulation and the STORM team. It enables separating algorithms of time loop based scientific simulations into platform independent algorithms and platform specific optimisation files, thus enforcing separation of concerns between the algorithmic design process on one side, and the optimization process of specifying platform-dependent aspects such as operation ordering and memory placement on the other side.

7.12. HPC Big Data Convergence

A Java interface for StarPU has been implemented and allows to execute Map Reduce applications on top of StarPU. We have made some preliminary experiments on Cornac, a big data application for visualising huge graphs.

7.13. Hierarchical Tasks

We are continuing our work, on the partitioning of the data and the prioritization of task graphs to optimize the use of resources of a machine. Hierarchical tasks allow a better control over the submission of an application's task graph by allowing to dynamically adapt the granularity of the calculations. In the ANR project Solharis, hierarchical tasks are proposed as a solution for a better management of dynamic task graphs. We have continued to explore new solutions for maximizing the performance of hierarchical tasks.

7.14. ADT Gordon

In collaboration with the HIEPACS and TADAAM Inria teams, we are strengthening the relations between the Chameleon linear algebra library from HIEPACS, our StarPU runtime scheduler, and the NewMadeleine high-performance communication library from TADAAM. More precisely, we have improved the interoperation between StarPU and NewMadeleine, to more carefully decide when NewMadeleine should proceed with communications. We have then introduced the notion of dynamic collective operations, which opportunistically introduce communication trees to balance the communication load.

7.15. StarPU in Julia

Julia is a modern language for parallelism and simulation that aims to ease the effort for developing high performance codes. In this context, we carry on the development of a StarPU binding inside Julia. It is possible to launch StarPU tasks inside Julia, either given as libraries, or described in Julia directly. The tasks described in Julia are compiled into either source OpenMP code or CUDA code. We improved further the support of StarPU in Julia, but this is still a work in progress.

7.16. Simulation of OpenMP task based programs

A simulator for OpenMP task based programs is being designed as part of Inria's IPL HAC-Specis project, and the Ph.D Thesis of Idriss Daoudi. The goal is to extend the SimGrid HPC simulation framework with the ability to simulate OpenMP applications.

7.17. OpenMP enabled version of Chameleon

An OpenMP enabled version of the Chameleon linear algebra library was designed within the context of European Project PRACE-5IP. This enables the Chameleon library to be available on any platform for which an OpenMP compiler is installed, without any requirement for third party task-based runtime systems. A preliminary support of the OpenMP port for heterogeneous, accelerated platform was also designed as part of this work.

TADAAM Project-Team

7. New Results

7.1. Management of heterogeneous and non-volatile memories in HPC

The emergence of non-volatile memory that may be used either as fast storage or slow high-capacity memory brings many opportunities for application developers.

We studied the impact of those new technologies on the allocation of resources in HPC platforms. We showed that co-scheduling HPC applications will possibly different needs in term of storage and memories brings constraints of the way non-volatile memory should be exposed by the hardware and operating system to bring both flexibility and performance. [21]

We also worked with Lawrence Livermore National Lab to propose an API to help application choose between the different kinds of available memory (high-bandwidth (HBM), normal (DDR), slow (non-volatile)). We exposed several useful criteria for selecting target memories as well as ways to rank them. [22]

7.2. Modeling and Visualizing Many-core HPC Platforms

As the number of cores keeps increasing inside processors, new kinds of hierarchy are added to organize and interconnect them. We worked with Intel to leverage new groups of cores such as *Dies* in newest Xeon Advanced Performance models. We also designed ways to clarify the modeling and visualisation of those many cores by factorizing identical parts of the platforms.

7.3. Co-scheduling HPC workloads on cache-partitioned CMP platforms

Co-scheduling techniques are used to improve the throughput of applications on chip multiprocessors (CMP), but sharing resources often generates critical interferences.

In collaboration with ENS Lyon and Georgia Tech, we looked at the interferences in the last level of cache (LLC) and use the *Cache Allocation Technology* (CAT) recently provided by Intel to partition the LLC and give each co-scheduled application their own cache area.

We considered m iterative HPC applications running concurrently and answer the following questions: (i) how to precisely model the behavior of these applications on the cache partitioned platform? and (ii) how many cores and cache fractions should be assigned to each application to maximize the platform efficiency? Here, platform efficiency is defined as maximizing the performance either globally, or as guaranteeing a fixed ratio of iterations per second for each application. Through extensive experiments using CAT, we demonstrated the impact of cache partitioning when multiple HPC application are co-scheduled onto CMP platforms. [2]

7.4. Modeling High-throughput Applications for in situ Analytics

In this work [3], we proposed to model HPC applications in the framework of in situ analytics. Typically, an HPC application is composed of a simulation tasks (data and compute intensive), and a set of analysis tasks that post-process the data. Currently, the performance of the I/O system in HPC platform prohibits the storage of all simulation data to process analysis post-mortem. Hence, in situ framework proposes to treat the data "on the fly", directly where it is produced. Hence, it leverages the amount of data to store as we only keep the result of analytics phase. However, simulation and analysis have to be scheduled in parallel and compete for shared resources. It generates resource conflicts and can lead to severe performance degradation for the simulation.

Hence, we proposed to model both platform (number of nodes and cores, memory, etc) and application (profile of each tasks) in order to optimize the execution of such applications. We propose a resource partitioning model that affects computational resources to the different tasks, as so as a scheduling of those tasks in order to maximize resource usage and minimize total application makespan. Tasks are assumed to be fully parallel to solve the partitioning problem.

We evaluated different scheduling heuristics combined to the resource partitioning model and show important features that influence in situ analytics performance.

This work is done in collaboration with Bruno RAFFIN from Inria team DATAMOVE of Inria Grenoble.

7.5. Modeling Non-Uniform Memory Access and Heterogeneous Memories on Large Compute Nodes with the Cache-Aware Roofline Model

The trend of increasing the number of cores on-chip is enlarging the gap between compute power and memory performance. This issue leads to design systems with heterogeneous memories, creating new challenges for data locality. Before the release of those memory architectures, the Cache-Aware Roofline Model [43] (CARM) offered an insightful model and methodology to improve application performance with knowledge of the cache memory subsystem.

With the help of the HWLOC library, we are able to leverage the machine topology to extend the CARM for modeling NUMA and heterogeneous memory systems, by evaluating the memory bandwidths between all combinations of cores and NUMA nodes. The new Locality Aware Roofline Model [6] (LARM) scopes most contemporary types of large compute nodes and characterizes three bottlenecks typical of those systems, namely contention, congestion and remote access. We also designed a hybrid memory bandwidth model to better estimate the roof when heterogeneous memories are involved or when read and write bandwidths differ.

We also developed an hybrid bandwidth model that combines the performance of different memories and their respective read/write bandwidth with the application memory access pattern to predict the performance of these accesses on heterogeneous memory platforms.

This work has been achieved in collaboration with the authors of the CARM from University of Lisbon.

7.6. Statistical Learning for Task and Data Placement in NUMA Architecture

Achieving high performance for multi-threaded application requires both a careful placement of threads on computing units and a thorough allocation of data in memory. Finding such a placement is a hard problem to solve, because performance depends on complex interactions in several layers of the memory hierarchy.

We proposed a black-box approach to decide if an application execution time can be impacted by the placement of its threads and data, and in such a case, to choose the best placement strategy to adopt [18]. We show that it is possible to reach near-optimal placement policy selection by looking at hardware performance counters, and at counters obtained from application instrumentation. Furthermore, solutions work across several recent processor architectures (from Haswell to Skylake), across several applications, and decisions can be taken with a single run of low overhead profiling.

This work has been achieved in collaboration with Thomas ROPARS from University of Grenoble.

7.7. On-the-fly scheduling vs. reservation-based scheduling for unpredictable workflows

Scientific insights in the coming decade will clearly depend on the effective processing of large datasets generated by dynamic heterogeneous applications typical of workflows in large data centers or of emerging fields like neuroscience. In this work [8], we show how these big data workflows have a unique set of characteristics that pose challenges for leveraging HPC methodologies, particularly in scheduling. Our findings indicate that execution times for these workflows are highly unpredictable and are not correlated with the size of the dataset involved or the precise functions used in the analysis. We characterize this inherent variability and sketch the need for new scheduling approaches by quantifying significant gaps in achievable performance. Through simulations, we show how on-the-fly scheduling approaches can deliver benefits in both system-level and user-level performance measures. On average, we find improvements of up to 35% in system utilization and up to 45% in average stretch of the applications, illustrating the potential of increasing performance through new scheduling approaches.

7.8. Scheduling strategies for stochastic jobs

Following the observations of made in 7.7 , we studied stochastic jobs (coming from neuroscience applications) which we want to schedule on a reservation-based platform (e.g. cloud, HPC).

The execution time of jobs is modeled using a (known) probability distribution. The platform to run the job is reservation-based, meaning that the user has to request fixed-length time slots for its job to be executed. The aim of this project is to study efficient strategies of reservation for an user given the cost associated to the machine. These reservations are all paid until a job is finally executed.

As a first step we derived efficient strategies without any additional asumptions [15]. This allowed us to set up properly the problem. These strategies were general enough that they could take as input any probability distributions, and performed better than any more natural strategies. Then we extended our strategies by including checkpoint/restart to well-chosen reservations in order to avoid wasting the benefits of work during underestimated reservations [35]. We were able to develop a fully polynomial-time approximation for continuous distribution of job execution time whose performance we then experimentally studied.

The final works of this project focused on the case without checkpointing: we studied experimentally how the strategies developed in [15] would perform in a parallel setup and showed that they improve both system utilization and job response time. Finally we started to study the robustness of such solutions when the job distributions were not perfectly known [19] and observed that the performance were still correct even with a very low quantity of information.

7.9. Online Prediction of Network Utilization

Stealing network bandwidth helps a variety of HPC runtimes and services to run additional operations in the background without negatively affecting the applications. A key ingredient to make this possible is an accurate prediction of the future network utilization, enabling the runtime to plan the background op- erations in advance, such as to avoid competing with the application for network bandwidth. In this work [23], we have proposed a portable deep learning predictor that only uses the information available through MPI introspection to construct a recurrent sequence-to-sequence neural network capable of forecasting network utilization. We leverage the fact that most HPC applications exhibit periodic behaviors to enable predictions far into the future (at least the length of a period). Our online approach does not have an initial training phase, it continuously improves itself during application execution without incurring significant computational overhead. Experimental results show better accuracy and lower computational overhead compared with the state-of-the-art on two representative applications.

7.10. An Introspection Monitoring Library

In this work [36] we have described how to improve communication time of MPI parallel applications with the use of a library that enables to monitor MPI applications and allows for introspection (the program itself can query the state of the monitoring system). Based on previous work, this library is able to see how collective communications are decomposed into point-to-point messages. It also features monitoring sessions that allow suspending and restarting the monitoring, limiting it to specific portions of the code. Experiments show that the monitoring overhead is very small and that the proposed features allow for dynamic and efficient rank reordering enabling up to 2-time reduction of communication parts of some program.

7.11. Tag matching in constant time

Tag matching is the operation, inside an MPI library, of pairing a packet arriving from the network, with its corresponding receive request posted by the user. This operation is not straightforward given that matching criterions are the communicator, the source of the message, a user-supplied tag, and since there are wildcards for tag and source. State of the art algorithms are linear with the number of pending packets and requests, or don’t support wildcards.

We proposed [17] an algorithm that is able perform the matching operation in constant time, in all cases, even with wildcard requests. We implemented the algorithm in our NEWMADELEINE communication library, and demonstrated it actually improves performance of Cholesky factorization with CHAMELEON running on top of STARPU.

7.12. Dynamic broadcasts in StarPU/NewMadeleine

We worked on the improvement of broadcast performance in STARPU runtime with NEWMADELEINE. Although STARPU supports MPI, its distributed and asynchronous model to schedule tasks makes it impossible to use MPI optimized routines, such as `MPI_Bcast`. Indeed these functions need that all nodes participating in the collective are synchronized and know each others, which makes it unusable in practice for STARPU.

We proposed [42], a dynamic broadcast algorithm that runs without synchronization among participants, and where only the root node needs to know the others. Recipient don't even have to know whether the message will arrive as a plain send/receive or through a dynamic broadcast, which allows for a seamless integration in STARPU. We implemented the algorithm in our NEWMADELEINE communication library, leveraging its event-based paradigm and background progression of communications. Preliminary experiments using Cholesky factorization from the CHAMELEON library show a sensible performance improvement.

7.13. Task based asynchronous MPI collectives optimisation

Asynchronous collectives are more complex than plain non-blocking point-to-point communications. They need specific mechanisms for progression. Task based progression is a good way to improve the performance of applications with overlap.

We worked on a benchmarking tool [41] measuring specific collective overlapping, taking into account time shift between different nodes. Using this tool, we were able to experiment with different task execution policies in the NEWMADELEINE communication library.

We propose a progression policy consisting of a dedicated a core for progression tasks; modern processors have more and more cores, so it is profitable on that kind of processors. The only function of this core is to progress communications, so we use a particularly aggressive algorithm for this progression.

7.14. Dynamic placement of progress thread for overlapping MPI non-blocking collectives on manycore processor

To amortize the cost of MPI collective operations, non-blocking collectives have been proposed so as to allow communications to be overlapped with computation. Unfortunately, collective communications are more CPU-hungry than point-to-point communications and running them in a communication thread on a single dedicated CPU core makes them slow. On the other hand, running collective communications on the application cores leads to no overlap. To address these issues, we proposed [5] an algorithm for tree-based collective operations that splits the tree between communication cores and application cores. To get the best of both worlds, the algorithm runs the short but heavy part of the tree on application cores, and the long but narrow part of the tree on one or several communication cores, so as to get a trade-off between overlap and absolute performance. We provided a model to study and predict its behavior and to tune its parameters. We implemented it in the MPC framework, which is a thread-based MPI implementation. We have run benchmarks on manycore processors such as the KNL and Skylake and got good results both in terms of performance and overlap.

7.15. Dynamic placement of Hybrid MPI +X coupled applications

We continued our collaboration with CERFACS in order to propose the HIPPO software that addresses the issue of dynamic placement of computing kernels that feature each their own placement/mapping/binding policy of MPI processes and OpenMP threads. In such a case, enforcing a global placement policy for the whole application composed of several such kernels may be detrimental to the overall performance. HIPPO (based on our HSPLIT library and the HWLOC software) is able to make the selection of the relevent resource on which some master MPI processes are going to execute and spawn OpenMP parallel sections while the remaining MPI processes are put in a “quiescence” state. HIPPO is currently at the prototype stage and the interface and the set of provided functionnalities need some refinement, however, preliminary results are very encouraging, especially on climate modelling applications from Météo France.

7.16. Scheduling on Two Unbounded Resources with Communication Costs

Heterogeneous computing systems are popular and powerful platforms, containing several heterogeneous computing elements (e.g. CPU+GPU). In [13], we consider a platform with two types of machines , each containing an unbounded number of elements. We want to execute an application represented as a Directed Acyclic Graph (DAG) on this platform. Each task of the application has two possible execution times, depending on the type of machine it is executed on. In addition we consider a cost to transfer data from one platform to the other between successive tasks. We aim at minimizing the execution time of the DAG (also called makespan). We show that the problem is NP-complete for graphs of depth at least three but polynomial for graphs of depth at most two. In addition, we provide polynomial-time algorithms for some usual classes of graphs (trees, series-parallel graphs).

7.17. H-Revolve: A Framework for Adjoint Computation on Synchrone Hierarchical Platforms

In this work [38], we study the problem of checkpointing strategies for adjoint computation on synchrone hierarchical platforms. Specifically we consider computational platforms with several levels of storage with different writing and reading costs. When reversing a large adjoint chain, choosing which data to checkpoint and where is a critical decision for the overall performance of the computation. We introduce H-Revolve, an optimal algorithm for this problem. We make it available in a public Python library along with the implementation of several state-of-the-art algorithms for the variant of the problem with two levels of storage. We provide a detailed description of how one can use this library in an adjoint computation software in the field of automatic differentiation or backpropagation. Finally, we evaluate the performance of H-Revolve and other checkpointing heuristics though an extensive campaign of simulation.

7.18. Sizing and Partitioning Strategies for Burst-Buffers to Reduce IO Contention

Burst-Buffers are high throughput and small size storage which are being used as an intermediate storage between the PFS (Parallel File System) and the computational nodes of modern HPC systems. They can allow to hinder to contention to the PFS, a shared resource whose read and write performance increase slower than processing power in HPC systems. A second usage is to accelerate data transfers and to hide the latency to the PFS. In this work [14], we concentrate on the first usage. We propose a model for Burst-Buffers and application transfers. We consider the problem of dimensioning and sharing the Burst-Buffers between several applications. This dimensioning can be done either dynamically or statically. The dynamic allocation considers that any application can use any available portion of the Burst-Buffers. The static allocation considers that when a new application enters the system, it is assigned some portion of the Burst-Buffers, which cannot be used by the other applications until that application leaves the system and its data is purged from it. We show that the general sharing problem to guarantee fair performance for all applications is an NP-Complete problem. We propose a polynomial time algorithms for the special case of finding the optimal buffer size such that no application is slowed down due to PFS contention, both in the static and dynamic cases. Finally, we provide evaluations of our algorithms in realistic settings. We use those to discuss how to minimize the overhead of the static allocation of buffers compared to the dynamic allocation.

7.19. Optimal Memory-aware Backpropagation of Deep Join Networks

Deep Learning training memory needs can prevent the user to consider large models and large batch sizes. In our work [4] (extended version [34]), we propose to use techniques from memory-aware scheduling and Automatic Differentiation (AD) to execute a backpropagation graph with a bounded memory requirement at the cost of extra recomputations. The case of a single homogeneous chain, i.e. the case of a network whose all stages are identical and form a chain, is well understood and optimal solutions have been proposed in the AD literature. The networks encountered in practice in the context of Deep Learning are much more diverse, both in terms of shape and heterogeneity. In this work, we define the class of backpropagation graphs, and extend those on which one can compute in polynomial time a solution that minimizes the total number of recomputations. In particular we consider join graphs which correspond to models such as Siamese or Cross Modal Networks.

7.20. Optimal checkpointing for heterogeneous chains: how to train deep neural networks with limited memory

This work [33] introduces a new activation checkpointing method which allows to significantly decrease memory usage when training Deep Neural Networks with the back-propagation algorithm. Similarly to checkpoint-ing techniques coming from the literature on Automatic Differentiation, it consists in dynamically selecting the forward activations that are saved during the training phase, and then automatically recomputing missing activations from those previously recorded. We propose an original computation model that combines two types of activation savings: either only storing the layer inputs, or recording the complete history of operations that produced the outputs (this uses more memory, but requires fewer recomputations in the backward phase), and we provide an algorithm to compute the optimal computation sequence for this model. This paper also describes a PyTorch implementation that processes the entire chain, dealing with any sequential DNN whose internal layers may be arbitrarily complex and automatically executing it according to the optimal checkpointing strategy computed given a memory limit. Through extensive experiments, we show that our implementation consistently outperforms existing checkpoint-ing approaches for a large class of networks, image sizes and batch sizes.

7.21. I/O scheduling strategy for HPC applications

With the ever-growing need of data in HPC applications, the congestion at the I/O level becomes critical in supercomputers. Architectural enhancement such as burst buffers and pre-fetching are added to machines, but are not sufficient to prevent congestion. Recent online I/O scheduling strategies have been put in place, but they add an additional congestion point and overheads in the computation of applications.

In this project, we studied application pattern (such as periodicity), in order to develop efficient scheduling strategies [7], [32] for their I/O transfers.

7.22. A New Framework for Evaluating Straggler Detection Mechanisms in MapReduce

In this work [10] we present a new framework for evaluating straggler detection mechanisms in MapReduce. We then show how to use it efficiently.

7.23. Clarification of the MPI semantics

In the framework of the MPI Forum, we have been involved in several active working groups, in particular the “Terms and Conventions” Working Group. The work carried out in this group has lead to a timely study and proposed clarifications, revisions, and enhancements to the Message Passing Interface’s (MPI’s) Semantic Terms and Conventions. To enhance MPI, a clearer understanding of the meaning of the key terminology has proven essential, and, surprisingly, important concepts remain underspecified, ambiguous and, in some cases, inconsistent and/or conflicting despite 26 years of standardization. This work [16] addresses these concerns comprehensively and usefully informs MPI developers, implementors, those teaching and learning MPI, and power users alike about key aspects of existing conventions, syntax, and semantics. This work will also be a useful driver for great clarity in current and future standardization and implementation efforts for MPI.

7.24. Adaptive Request Scheduling for the I/O Forwarding Layer using Reinforcement Learning

I/O optimization techniques such as request scheduling can improve performance mainly for the access patterns they target, or they depend on the precise tune of parameters. In this work [40], we propose an approach to adapt the I/O forwarding layer of HPC systems to the application access patterns by tuning a request scheduler. Our case study is the TWINS scheduling algorithm, where performance improvements depend on the time window parameter, which depends on the current workload. Our approach uses a reinforcement learning technique — contextual bandits — to make the system capable of learning the best parameter value to each access pattern during its execution, without a previous training phase. We evaluate our proposal and demonstrate it can achieve a precision of 88% on the parameter selection in the first hundreds of observations of an access pattern. After having observed an access pattern for a few minutes (not necessarily contiguously), we demonstrate that the system will be able to optimize its performance for the rest of the life of the system (years).

DIVERSE Project-Team

6. New Results

6.1. Results on Variability modeling and management

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

6.1.1. Variability and testing.

The performance of software systems (such as speed, memory usage, correct identification rate) tends to be an evermore important concern, often nowadays on par with functional correctness for critical systems. Systematically testing these performance concerns is however extremely difficult, in particular because there exists no theory underpinning the evaluation of a performance test suite, i.e., to tell the software developer whether such a test suite is "good enough" or even whether a test suite is better than another one. This work [37] proposes to apply **Multimorphic testing** and empirically assess the effectiveness of performance test suites of software systems coming from various domains. By analogy with mutation testing, our core idea is to leverage the typical configurability of these systems, and to check whether it makes any difference in the outcome of the tests: i.e., are some tests able to "kill" underperforming system configurations? More precisely, we propose a framework for defining and evaluating the coverage of a test suite with respect to a quantitative property of interest. Such properties can be the execution time, the memory usage or the success rate in tasks performed by a software system. This framework can be used to assess whether a new test case is worth adding to a test suite or to select an optimal test suite with respect to a property of interest. We evaluate several aspects of our proposal through 3 empirical studies carried out in different fields: object tracking in videos, object recognition in images, and code generators.

6.1.2. Variability, sampling, and SAT.

Uniform or near-uniform generation of solutions for large satisfiability formulas is a problem of theoretical and practical interest for the testing community. Recent works proposed two algorithms (namely UniGen and QuickSampler) for reaching a good compromise between execution time and uniformity guarantees, with empirical evidence on SAT benchmarks. In the context of highly-configurable software systems (e.g., Linux), it is unclear whether UniGen and QuickSampler can scale and sample uniform software configurations. We perform a thorough experiment on 128 real-world feature models. We find that UniGen is unable to produce SAT solutions out of such feature models. Furthermore, we show that QuickSampler does not generate uniform samples and that some features are either never part of the sample or too frequently present. Finally, using a case study, we characterize the impacts of these results on the ability to find bugs in a configurable system. Overall, our results suggest that we are not there: more research is needed to explore the cost-effectiveness of uniform sampling when testing large configurable systems. More details [51]. In general, we are investigating sampling algorithms for cost-effectively exploring configuration spaces (see also [63], [67].

6.1.3. Variability and 3D printing.

Configurators rely on logical constraints over parameters to aid users and determine the validity of a configuration. However, for some domains, capturing such configuration knowledge is hard, if not infeasible. This is the case in the 3D printing industry, where parametric 3D object models contain the list of parameters and their value domains, but no explicit constraints. This calls for a complementary approach that learns what configurations are valid based on previous experiences. In this work [41], we report on preliminary experiments showing the capability of state-of-the-art classification algorithms to assist the configuration process. While machine learning holds its promises when it comes to evaluation scores, an in-depth analysis reveals the opportunity to combine the classifiers with constraint solvers.

6.1.4. Variability and video processing.

In an industrial project [24], we addressed the challenge of developing a software-based video generator such that consumers and providers of video processing algorithms can benchmark them on a wide range of video variants. We have designed and developed a variability modeling language, called VM, resulting from the close collaboration with industrial partners during two years. We expose the specific requirements and advanced variability constructs we developed and used to characterize and derive variations of video sequences. The results of our experiments and industrial experience show that our solution is effective to model complex variability information and supports the synthesis of hundreds of realistic video variants. From the software language perspective, we learned that basic variability mechanisms are useful but not enough; attributes and multi-features are of prior importance; meta-information and specific constructs are relevant for scalable and purposeful reasoning over variability models. From the video domain and software perspective, we report on the practical benefits of a variability approach. With more automation and control, practitioners can now envision benchmarking video algorithms over large, diverse, controlled, yet realistic datasets (videos that mimic real recorded videos) – something impossible at the beginning of the project.

6.1.5. Variability and adversarial machine learning

Software product line engineers put a lot of effort to ensure that, through the setting of a large number of possible configuration options, products are acceptable and well-tailored to customers' needs. Unfortunately, options and their mutual interactions create a huge configuration space which is intractable to exhaustively explore. Instead of testing all products, machine learning is increasingly employed to approximate the set of acceptable products out of a small training sample of configurations. Machine learning (ML) techniques can refine a software product line through learned constraints and a priori prevent non-acceptable products to be derived. In this work [53], we use adversarial ML techniques to generate adver- sarial configurations fooling ML classifiers and pinpoint incorrect classifications of products (videos) derived from an industrial video generator. Our attacks yield (up to) a 100% misclassification rate and a drop in accuracy of 5%. We discuss the implications these results have on SPL quality assurance.

6.1.6. Variability, Linux and machine learning

Given a configuration, can humans know in advance the build status, the size, the compilation time, or the boot time of a Linux kernel? Owing to the huge complexity of Linux (there are more than 15000 options with hard constraints and subtle interactions), machines should rather assist contributors and integrators in mastering the configuration space of the kernel. We have developed TuxML https://github.com/TuxML/ an open-source tool based on Docker/Python to massively gather data about thousands of kernel configurations. 200K+ configurations have been automatically built and we show how machine learning can exploit this information to predict properties of unseen Linux configurations, with different use cases (identification of influential/buggy options, finding of small kernels, etc.) The vision is that a continuous understanding of the configuration space is undoubtedly beneficial for the Linux community, yet several technical challenges remain in terms of infrastructure and automation.

Two preprints are available [62] and [49].

A talk has been given at Embedded Linux Conference Europe 2019 (co-located with Open Source Summit 2019) in Lyon about "Learning the Linux Kernel Configuration Space: Results and Challenges" [54].

6.1.7. Variability and machine learning

We gave a tutorial [49] at SPLC 2019 and introduce how machine learning can be used to support activities related to the engineering of configurable systems and software product lines. To the best of our knowledge, this is the first practical tutorial in this trending field. The tutorial is based on a systematic literature review [67] and includes practical tasks (specialization, performance prediction) on real-world systems (VaryLaTeX, x264).

6.2. Results on Software Language Engineering

6.2.1. Software Language Extension Problem

The problem of software language extension and composition drives much of the research in *Software Language Engineering* (SLE). Although various solutions have already been proposed, there is still little understanding of the specific ins and outs of this problem, which hinders the comparison and evaluation of existing solutions. In [34], we introduce the Language Extension Problem as a way to better qualify the scope of the challenges related to language extension and composition. The formulation of the problem is similar to the seminal Expression Problem introduced by Wadler in the late nineties, and lift it from the extensibility of single constructs to the extensibility of groups of constructs, i.e., software languages. We provide a comprehensive definition of the actual constraints when considering language extension, and believe the Language Extension Problem will drive future research in SLE, the same way the original Expression Problem helped to understand the strengths and weaknesses of programming languages and drove much research in programming languages.

6.2.2. A unifying framework for homogeneous model composition

The growing use of models for separating concerns in complex systems has lead to a proliferation of model composition operators. These composition operators have traditionally been defined from scratch following various approaches differing in formality, level of detail, chosen paradigm, and styles. Due to the lack of proper foundations for defining model composition (concepts, abstractions, or frameworks), it is difficult to compare or reuse composition operators. In [33], we stipulate the existence of a unifying framework that reduces all structural composition operators to structural merging, and all composition operators acting on discrete behaviors to event scheduling. We provide convincing evidence of this hypothesis by discussing how structural and behavioral homogeneous model composition operators (i.e., weavers) can be mapped onto this framework. Based on this discussion, we propose a conceptual model of the framework, and identify a set of research challenges, which, if addressed, lead to the realization of this framework to support rigorous and efficient engineering of model composition operators for homogeneous and eventually heterogeneous modeling languages.

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

Executable Domain-Specific Modeling Languages (xDSMLs) enable the application of early dynamic verification and validation (V&V) techniques for behavioral models. At the core of such techniques, execution traces are used to represent the evolution of models during their execution. In order to construct execution traces for any xDSML, generic trace metamodels can be used. Yet, regarding trace manipulations, generic trace metamodels lack efficiency in time because of their sequential structure, efficiency in memory because they capture superfluous data, and usability because of their conceptual gap with the considered xDSML. Our contribution in [26] is a novel generative approach that defines a multidimensional and domain-specific trace metamodel enabling the construction and manipulation of execution traces for models conforming to a given xDSML. Efficiency in time is improved by providing a variety of navigation paths within traces, while usability and memory are improved by narrowing the scope of trace metamodels to fit the considered xDSML. We evaluated our approach by generating a trace metamodel for fUML and using it for semantic differencing, which is an important V&V technique in the realm of model evolution. Results show a significant performance improvement and simplification of the semantic differencing rules as compared to the usage of a generic trace metamodel.

6.2.4. From DSL specification to interactive computer programming environment

The adoption of Domain-Specific Languages (DSLs) relies on the capacity of language workbenches to automate the development of advanced and customized environments. While DSLs are usually well tailored for the main scenarios, the cost of developing mature tools prevents the ability to develop additional capabilities for alternative scenarios targeting specific tasks (e.g., API testing) or stakeholders (e.g., education). In [47], we propose an approach to automatically generate interactive computer programming environments from existing specifications of textual interpreted DSLs. The approach provides abstractions to complement the DSL specification, and combines static analysis and language transformations to automate the transformation of the language syntax, the execution state and the execution semantics. We evaluate the approach over a representative set of DSLs, and demonstrate the ability to automatically transform a textual syntax to load partial programs limited to a single statement, and to derive a Read-Eval-Print-Loop (REPL) from the specification of a language interpreter.

6.2.5. Live-UMLRT: A Tool for Live Modeling of UML-RT Models

In the context of Model-driven Development (MDD) models can be executed by interpretation or by the translation of models into existing programming languages, often by code generation. In [42] we present Live-UMLRT, a tool that supports live modeling of UML-RT models when they are executed by code generation. Live-UMLRT is entirely independent of any live programming support offered by the target language. This independence is achieved with the help of a model transformation which equips the model with support for, e.g., debugging and state transfer both of which are required for live modeling. A subsequent code generation then produces a self-reflective program that allows changes to the model elements at runtime (through synchronization of design and runtime models). We have evaluated Live-UMLRT on several use cases. The evaluation shows that (1) code generation, transformation, and state transfer can be carried out with reasonable performance, and (2) our approach can apply model changes to the running execution faster than the standard approach that depends on the live programming support of the target language. A demonstration video: https://youtu.be/6GrR-Y9je7Y.

6.2.6. Applying model-driven engineering to high-performance computing: Experience report, lessons learned, and remaining challenges

In [35], we present a framework for generating optimizing compilers for performance-oriented embedded DSLs (EDSLs). This framework provides facilities to automatically generate the boilerplate code required for building DSL compilers on top of the existing extensible optimizing compilers. We evaluate the practicality of our framework by demonstrating a real-world use-case successfully built with it.

6.2.7. Software languages in the wild (Wikipedia)

Wikipedia is a rich source of information across many knowledge domains. Yet, recovering articles relevant to a specific domain is a difficult problem since such articles may be rare and tend to cover multiple topics. Furthermore, Wikipedia's categories provide an ambiguous classification of articles as they relate to all topics and thus are of limited use. In [46], we develop a new methodology to isolate Wikipedia's articles that describe a specific topic within the scope of relevant categories; the methodology uses super- vised machine learning to retrieve a decision tree classifier based on articles' features (URL patterns, summary text, infoboxes, links from list articles). In a case study, we retrieve 3000+ articles that describe software (computer) languages. Available fragments of ground truths serve as an essential part of the training set to detect relevant articles. The results of the classification are thoroughly evaluated through a survey, in which 31 domain experts participated.

6.3. Results on Heterogeneous and dynamic software architectures

We have selected three main contributions for DIVERSE's research axis #4: one is in the field of runtime management of resources for dynamically adaptive system, one in the field of temporal context model for dynamically adaptive system and a last one to improve the exploration of hidden real-time structures of programming behavior at run time.

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

In Kevoree, one of the goal is to work on the shipping phases in which we aim at making deployment, and the reconfiguration simple and accessible to a whole development team. This year, we mainly explore two main axes.

In the first one, we try to improve the proposed models that could be used at run time to improve resource usage in two domains: cloud computing [30], [57] and energy [58].

6.3.2. Investigating Machine Learning Algorithms for Modeling SSD I/O Performance for Container-based Virtualization

One of the cornerstones of the cloud provider business is to reduce hardware resources cost by maximizing their utilization. This is done through smartly sharing processor, memory, network and storage, while fully satisfying SLOs negotiated with customers. For the storage part, while SSDs are increasingly deployed in data centers mainly for their performance and energy efficiency, their internal mechanisms may cause a dramatic SLO violation. In effect, we measured that I/O interference may induce a 10x performance drop. We are building a framework based on autonomic computing which aims to achieve intelligent container placement on storage systems by preventing bad I/O interference scenarios. One prerequisite to such a framework is to design SSD performance models that take into account interactions between running processes/containers, the operating system and the SSD. These interactions are complex. In this work [30], we investigate the use of machine learning for building such models in a container based Cloud environment. We have investigated five popular machine learning algorithms along with six different I/O intensive applications and benchmarks. We analyzed the prediction accuracy, the learning curve, the feature importance and the training time of the tested algorithms on four different SSD models. Beyond describing modeling component of our framework, this paper aims to provide insights for cloud providers to implement SLO compliant container placement algorithms on SSDs. Our machine learning-based framework succeeded in modeling I/O interference with a median Normalized Root-Mean-Square Error (NRMSE) of 2.5%.

6.3.3. Cuckoo: Opportunistic MapReduce on Ephemeral and Heterogeneous Cloud Resources

Cloud infrastructures are generally over-provisioned for handling load peaks and node failures. However, the drawback of this approach is that a large portion of data center resources remains unused. In this work [57], we propose a framework that leverages unused resources of data centers, which are ephemeral by nature, to run MapReduce jobs. Our approach allows: i) to run efficiently Hadoop jobs on top of heterogeneous Cloud resources, thanks to our data placement strategy, ii) to predict accurately the volatility of ephemeral resources, thanks to the quantile regression method, and iii) for avoiding the interference between MapReduce jobs and co-resident workloads, thanks to our reactive QoS controller. We have extended Hadoop implementation with our framework and evaluated it with three different data center workloads. The experimental results show that our approach divides Hadoop job execution time by up to 7 when compared to the standard Hadoop implementation. In [44], we presented a demo that leverages unused but volatile Cloud resources to run big data jobs. It is based on a learning algorithm that accurately predicts future availability of resources to automatically scale the ran jobs. We also designed a mechanism that avoids interference between the Big data jobs and co-resident workloads. Our solution is based on Open-Source components such as kubernetes and Apache Spark.

6.3.4. Leveraging cloud unused resources for Big data application while achieving SLA

Companies are more and more inclined to use collaborative cloud resources when their maximum internal capacities are reached in order to minimize their TCO. The downside of using such a collaborative cloud, made of private clouds' unused resources, is that malicious resource providers may sabotage the correct execution of third-party-owned applications due to its uncontrolled nature. In this work [43], we propose an approach that allows sabotage detection in a trustless environment. To do so, we designed a mechanism that (1) builds an application fingerprint considering a large set of resources usage (such as CPU, I/O, memory) in a trusted environment using random forest algorithm, and (2) an online remote fingerprint recognizer that monitors application execution and that makes it possible to detect unexpected application behavior. Our approach has been tested by building the fingerprint of 5 applications on trusted machines. When running these applications on untrusted machines (with either homogeneous, heterogeneous or unspecified hardware from the one that was used to build the model), the fingerprint recognizer was able to ascertain whether the execution of the application is correct or not with a median accuracy of about 98% for heterogeneous hardware and about 40% for the unspecified one.

6.3.5. Benefits of Energy Management Systems on local energy efficiency, an agricultural case study

Energy efficiency is a concern impacting both ecology and economy. Most approaches aiming at reducing the energy impact of a site focus on only one specific aspect of the ecosystem: appliances, local generation or energy storage. A trade-off analysis of the many factors to consider is challenging and must be supported by tools. This work proposes a Model-Driven Engineering approach mixing all these concerns into one comprehensive model [58]. This model can then be used to size either local production means, either energy storage capacity and also help to analyze differences between technologies. It also enables process optimization by modeling activity variability: it takes the weather into account to give regular feedback to the end user. This approach is illustrated by simulation using real consumption and local production data from a representative agricultural site. We show its use by: sizing solar panels, by choosing between battery technologies and specification and by evaluating different demand response scenarios while examining the economic sustainability of these choices.

6.4. Results on Diverse Implementations for Resilience

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

6.4.1. Privacy and Security

6.4.1.1. A Collaborative Strategy for Mitigating Tracking through Browser Fingerprinting

Browser fingerprinting is a technique that collects information about the browser configuration and the environment in which it is running. This information is so diverse that it can partially or totally identify users online. Over time, several countermeasures have emerged to mitigate tracking through browser fingerprinting. However, these measures do not offer full coverage in terms of privacy protection, as some of them may introduce inconsistencies or unusual behaviors, making these users stand out from the rest. In this work [45], we address these limitations by proposing a novel approach that minimizes both the identifiability of users and the required changes to browser configuration. To this end, we exploit clustering algorithms to identify the devices that are prone to share the same or similar fingerprints and to provide them with a new non-unique fingerprint. We then use this fingerprint to automatically assemble and run web browsers through virtualization within a docker container. Thus all the devices in the same cluster will end up running a web browser with an indistinguishable and consistent fingerprint.

6.4.2. Software Testing

6.4.2.1. A Snowballing Literature Study on Test Amplification

The adoption of agile development approaches has put an increased emphasis on developer testing, resulting in software projects with strong test suites. These suites include a large number of test cases, in which developers embed knowledge about meaningful input data and expected properties in the form of oracles. This work [29] surveys various works that aim at exploiting this knowledge in order to enhance these manually written tests with respect to an engineering goal (e.g., improve coverage of changes or increase the accuracy of fault localization). While these works rely on various techniques and address various goals, we believe they form an emerging and coherent field of research, which we call ‘test amplification’. We devised a first set of papers from DBLP, looking for all papers containing ‘test’ and ‘amplification’ in their title. We reviewed the 70 papers in this set and selected the 4 papers that fit our definition of test amplification. We use these 4 papers as the seed for our snowballing study, and systematically followed the citation graph. This study is the first that draws a comprehensive picture of the different engineering goals proposed in the literature for test amplification. In particular, we note that the goal of test amplification goes far beyond maximizing coverage only. We believe that this survey will help researchers and practitioners entering this new field to understand more quickly and more deeply the intuitions, concepts and techniques used for test amplification.

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

In the literature, there is a rather clear segregation between manually written tests by developers and automatically generated ones. In this work, we explore a third solution: to automatically improve existing test cases written by developers. We present the concept, design, and implementation of a system called DSpot, that takes developer-written test cases as input (Junit tests in Java) and synthesizes improved versions of them as output. Those test improvements are given back to developers as patches or pull requests, that can be directly integrated in the main branch of the test code base. In this work [28], we have evaluated DSpot in a deep, systematic manner over 40 real-world unit test classes from 10 notable and open-source software projects. We have amplified all test methods from those 40 unit test classes. In 26/40 cases, DSpot is able to automatically improve the test under study, by triggering new behaviors and adding new valuable assertions. Next, for ten projects under consideration, we have proposed a test improvement automatically synthesized by DSpot to the lead developers. In total, 13/19 proposed test improvements were accepted by the developers and merged into the main code base. This shows that DSpot is capable of automatically improving unit-tests in real-world, large-scale Java software.

6.4.2.3. Leveraging metamorphic testing to automatically detect inconsistencies in code generator families

Generative software development has paved the way for the creation of multiple code generators that serve as a basis for automatically generating code to different software and hardware platforms. In this context, the software quality becomes highly correlated to the quality of code generators used during software development. Eventual failures may result in a loss of confidence for the developers, who will unlikely continue to use these generators. It is then crucial to verify the correct behaviour of code generators in order to preserve software quality and reliability. In this work [25], we leverage the metamorphic testing approach to automatically detect inconsistencies in code generators via so-called “metamorphic relations”. We define the metamorphic relation (i.e., test oracle) as a comparison between the variations of performance and resource usage of test suites running on different versions of generated code. We rely on statistical methods to find the threshold value from which an unexpected variation is detected. We evaluate our approach by testing a family of code generators with respect to resource usage and performance metrics for five different target software platforms. The experimental results show that our approach is able to detect, among 95 executed test suites, 11 performance and 15 memory usage inconsistencies.

6.4.3. Software Co-evolution

6.4.3.1. An Empirical Study on the Impact of Inconsistency Feedback during Model and Code Co-changing

Model and code co-changing is about the coordinated modification of models and code during evolution. Intermittent inconsistencies are a common occurrence during co-changing. A partial co-change is the period in which the developer changed, say, the model but has not yet propagated the change to the code. Inconsistency feedback can be provided to developers for helping them to complete partial co-changes. However, there is no evidence whether such inconsistency feedback is useful to developers. To investigate this problem, we conducted a controlled experiment with 36 subjects who were required to complete ten partially completed change tasks between models and code of two non-trivial systems [31]. The tasks were of different levels of complexity depending on how many model diagrams they affected. All subjects had to work on all change tasks but sometimes with and sometimes without inconsistency feedback. We then measured differences between task effort and correctness. We found that when subjects were given inconsistency feedback during tasks, they were 268% more likely to complete the co-change correctly compared to when they were not given inconsistency feedback. We also found that when subjects were not given inconsistency feedback, they nearly always failed in completing co-change tasks with high complexity where the partially completed changes were spread across different diagrams in the model. These findings suggest that inconsistency feedback (i.e. detection and repair) should form an integral part of co-changing, regardless of whether the code or the model changes first. Furthermore, these findings suggest that merely having access to changes (as with the given partially completed changes) is insufficient for effective co-changing.

6.4.3.2. Detecting and Exploring Side Effects when Repairing Model Inconsistencies

When software models change, developers often fail in keeping them consistent. Automated support in repairing inconsistencies is widely addressed. Yet, merely enumerating repairs for developers is not enough. A repair can as a side effect cause new unexpected inconsistencies (negative) or even fix other inconsistencies as well (positive). To make matters worse, repairing negative side effects can in turn cause further side effects. Current approaches do not detect and track such side effects in depth, which can increase developers' effort and time spent in repairing inconsistencies. This work [66] presents an automated approach for detecting and tracking the consequences of repairs, i.e. side effects. It recursively explores in depth positive and negative side effects and identifies paths and cycles of repairs. This work further ranks repairs based on side effect knowledge so that developers may quickly find the relevant ones. Our approach and its tool implementation have been empirically assessed on 14 case studies from industry, academia, and GitHub. Results show that both positive and negative side effects occur frequently. A comparison with three versioned models showed the usefulness of our ranking strategy based on side effects. It showed that our approach's top prioritized repairs are those that developers would indeed choose. A controlled experiment with 24 participants further highlights the significant influence of side effects and of our ranking of repairs on developers. Developers who received side effect knowledge chose far more repairs with positive side effects and far less with negative side effects, while being 12.3% faster, in contrast to developers who did not receive side effect knowledge.

6.4.3.3. Supporting A Flexible Grouping Mechanism for Collaborating Engineering Teams

Most engineering tools do not provide much support for collaborating teams and today's engineering knowledge repositories lack flexibility and are limited. Engineering teams have different needs and their team members have different preferences on how and when to collaborate. These needs may depend on the individual work style, the role an engineer has, and the tasks they have to perform within the collaborating group. However, individual collaboration is insufficient and engineers need to collaborate in groups. This work [65] presents a collaboration framework for collaborating groups capable of providing synchronous and asynchronous mode of collaboration. Additionally, our approach enables engineers to mix these collaboration modes to meet the preferences of individual group members. We evaluate the scalability of this framework using four real life large collaboration projects. These projects were found from GitHub and they were under active development by the time of evaluation. We have tested our approach creating groups of different sizes for each project. The results showed that our approach scales to support every case for the groups created. Additionally, we scouted the literature and discovered studies that support the usefulness of different groups with collaboration styles.

6.4.4. Software diversification

6.4.4.1. The Maven Dependency Graph: a Temporal Graph-based Representation of Maven Central

The Maven Central Repository provides an extraordinary source of data to understand complex architecture and evolution phenomena among Java applications. As of September 6, 2018, this repository includes 2.8M artifacts (compiled piece of code implemented in a JVM-based language), each of which is characterized with metadata such as exact version, date of upload and list of dependencies towards other artifacts. Today, one who wants to analyze the complete ecosystem of Maven artifacts and their dependencies faces two key challenges: (i) this is a huge data set; and (ii) dependency relationships among artifacts are not modeled explicitly and cannot be queried. In this work [55], we present the Maven Dependency Graph. This open source data set provides two contributions: a snapshot of the whole Maven Central taken on September 6, 2018, stored in a graph database in which we explicitly model all dependencies; an open source infrastructure to query this huge dataset.

6.4.4.2. The Emergence of Software Diversity in Maven Central

Maven artifacts are immutable: an artifact that is uploaded on Maven Central cannot be removed nor modified. The only way for developers to upgrade their library is to release a new version. Consequently, Maven Central accumulates all the versions of all the libraries that are published there, and applications that declare a dependency towards a library can pick any version. In this work [59], we hypothesize that the immutability of Maven artifacts and the ability to choose any version naturally support the emergence of software diversity within Maven Central. We analyze 1,487,956 artifacts that represent all the versions of 73,653 libraries. We observe that more than 30% of libraries have multiple versions that are actively used by latest artifacts. In the case of popular libraries, more than 50% of their versions are used. We also observe that more than 17% of libraries have several versions that are significantly more used than the other versions. Our results indicate that the immutability of artifacts in Maven Central does support a sustained level of diversity among versions of libraries in the repository.

EASE Project-Team

6. New Results

6.1. Smart City and ITS

Participants: Indra Ngurah, Christophe Couturier, Rodrigo Silva, Frédéric Weis, Jean-Marie Bonnin [contact].

In the last years, we contributed to the specification of the hybrid (ITS-G5 + Cellular) communication architecture of the French field operation test project SCOOP@F. The proposed solution relies on the MobileIP family of standards and the ISO/ETSI ITS Station architecture we contributed to standardize at IETF and ISO. On this topic our contribution mainly focussed on bringing concepts from the state of the art to real equipments. For the last year of the SCOOP@F part 2 project, we took part to the performance evaluation process by providing a test and validation platform for IP mobility protocols (MobileIP, NEMO) and IPsec cyphering. This platform allows us to identify the performance limits of current implementation of mobility and security protocols. Moreover it spotted implementation incompatibilities between the open source implementations of theses protocols (namely UMIP and StrongSwan) and helped the industrial partners of the project to identify associated risks.

InDiD is the logical follow up of SCOOP@F part 2. This 3.5 years long European project (mid 2019-2023) aims at testing ITS applications on a large scale national deployment of connected vehicles and infrastructure. This version of the project specifically complex use cases (so called day 1.5) and urban application. For the beginning of this project, we proposed several innovative use cases. Our "Backward cartography update" scenario has been selected as a priority candidate for implementation. In line with the collaborative approaches of EASE, we propose to use vehicles' observations to inform other vehicles and/or a cartography server about differences between the digital map and the reality.

We also want to explore the benefits of new capabilities of upcoming communication technologies to enrich the interactions between vehicle and smartphones or wearable devices. We defined an architecture for both localisation and communication with vulnerable users (workers in road and construction works). Short range communications between dangers (maneuvering construction vehicles) and workers rely on the advertisement feature of Bluetooth Low Energy (BLE). This connectionless communication mode enables for easy direct communication between any node in the neighborhood. It is inspired from the ITS-Station communication standard and we aim to integrate our work into future versions of the standards. Another contribution in this project aims at enhancing the localisation precision in harsh conditions. Recent version of radio communication standards (eg. Bluetooth 5.1 or 802.11ax) now integrate intrinsic real time localisation primitives giving information such as Angle of Arrival (AoA), Angle of Departure (AoD) or distance evaluation based on Time of Flight (ToF) measurements. We started to study how to merge this information with other localisation evidence sources and how to structure a collaborative framework to share it with other objects in the environment. This early works opens the doors to many other works in the future.

The development of innovative applications for smart cities has also been made possible by the rise of Internet of Things and especially the deployment of numerous low energy devices. The collection of the huge amount of data produced by all these piece of hardware become a challenge for the communication networks. In smart cities, the mobility of vehicles can be used to collect data produced by connected objects and to deliver them to several applications which are delay tolerant. The Vehicular Delay Tolerant Networks (VDTN) can be utilized for such services. We designed DC4LED (Data Collection for Low Energy Devices): a hierarchical VDTN routing which takes advantage of the specific mobility patterns of the various type of vehicles. It provides a low-cost delivery service for applications that need to gather data generated from the field. The idea is to propose a simple routing scheme where cars, taxis, and buses route data hierarchically in a store-carry-forward mechanism to any of the available Internet Point-of-Presences in the city. We compare using simulation tools the performance of DC4LED routing with two legacy VDTN routing schemes which represent the extreme ends of VDTN routing spectrum: First-contact and Epidemic routing. It show that DC4LED has much lower network overhead in comparison with the two legacy routing schemes, which is advantageous for its implementation scalability. The DC4LED also maintains comparable data delivery probability and latency to Epidemic routing.

The situational viewing and surveillance in cities is one such category of applications which can benefit from various networking solutions available to transport images or data from installed sensor cameras. We explore how our DC4LED mechanism can be used to for a city-wide image and data collection service. We study the networking performance in terms of increasing image sizes that can be transported with respect to varying vehicular density in city. We focus mainly on two technologies for sensors to vehicles communications: ZigBee and ITS-G5. We show that, surprisingly such very simple mechanism could meet the requirements of multiple services.

6.2. Autonomic Maintenance of Optical Networks

Participant: Jean-Marie Bonnin [contact].

The application of classification techniques based on machine learning approaches to analyze the behavior of network users has interested many researchers in the last years. In a recent work, we have proposed an architecture for optimizing the upstream bandwidth allocation in Passive Optical Network (PON) based on the traffic pattern of each user. Clustering analysis was used in association with an assignment index calculation in order to specify for PON users their upstream data transmission tendency. A dynamic adjustment of Service Level Agreement (SLA) parameters is then performed to maximize the overall customers' satisfaction with the network. In this work, we extend the proposed architecture by adding a prediction module as a complementary to the first classification phase. Grey Model GM(1,1) is used in this context to learn more about the traffic trend of users and improve their assignment. An experimental study is conducted to show the impact of the forecaster and how it can overcome the limits of the initial model.

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

6.3. Location assessment from local observations

Participants: Yoann Maurel, Paul Couderc [contact].

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

Unfortunately, location reporting methods implemented in existing services are weakly protected: it is often possible to lie in simple cases or to emit signals that deceive the more cautious systems. For example, we have experimented simple and successful replay attacks against Google Location using this approach, as shown on Figure 3 .

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

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

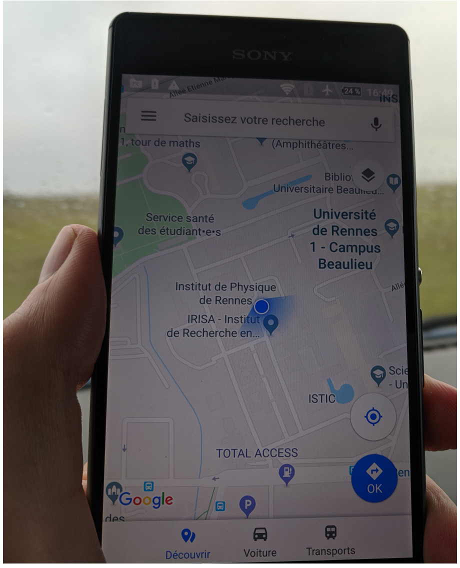

Figure 3. Google map deceived by faked Wi-Fi beacons replayed by an ESP-32

This work was published at CCNC'2019 [1]. We are currently working in broadening this approach, in particular using other attributes of Wi-Fi traffic beside beacon timestamps, and combining the timestamp solution with other type of challenges to propose a diversity of challenges for location validation servers. We are also working on the attack side, which presents interesting perspectives regarding the actual strength of existing services and the potential protection improvements than our approach can provide.

6.4. A methodological framework to promote the use of renewable energy

Participants: Alexandre Rio, Yoann Maurel [contact].

This work is in line with projects aimed at optimizing the use of renewable energies. It is carried out in collaboration with OKWind. This compagny designs and supplies its customers with renewable energy generators such as vertical axis wind turbines and solar trackers. OKWind promotes a micro-grid infrastructure development.

Our application domains are those of agriculture and industry in which it is possible to identify and influence consuming processes. We mainly consider local generation for self-consumption purposes (microgrid) as it limits infrastructure costs, minimizes line losses, reduces the need of the Grid and hopefully reduces the electricity bill.

Renewable energies currently benefit from numerous subsidies to promote their use so as to reduce greenhouse gas emissions. Nevertheless, it seems worth considering the cost-effectiveness of these solutions without these incentives, as they are highly dependent on political will and can be questioned. The reduction in manufacturing costs, particularly in solar energy, suggests that these solutions can eventually compete with traditional sources if they are properly used.

Competitive low-carbon energy is hampered by the stochastic nature of these sources. During peak periods, the electricity produced is competitive, but too often, the scheduled consumption is not aligned with production. In practice, process planning was and is still driven by the electricity price from the grid. On average, the profitability of the installations is therefore not certain. In this context, using battery to shift the load looks appealing but is, as of today, far from being economically viable if not done properly.

Consequently, the achievement of a profitable self-production site is, in practice, a question of trade-off that involves several factors: the scaling of energy sources, the sizing of batteries used, the desired autonomy level, the ecological concerns, and the organization of demand. This trade-off analysis is very challenging: to be carried out effectively and comprehensively, it must be supported by tools that help the stakeholders. While much work has been done in the literature on the impacts of different factors, there are few approaches that offer a comprehensive model.

Our objective is to provide a methodological framework to embrace the diversity of knowledge, of production and consumption tools, of farm activities and of prediction algorithms. This should enable an expert to conduct a trade-off analysis and decide on the best option for each individual site under consideration.

In the first two years of this PhD thesis, we argued that model-driven engineering is suited for the development of such a model and we presented some preliminary implementation. In 2019, we were able to test our approaches in the field and continue to expand the model to account for a wider range of resources. This was published in [5].

6.5. Introducing Data Quality to the Internet of Things

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

The Internet of Things (IoT) connects various distributed heterogeneous devices. Such Things sense and actuate their physical environment. The IoT pervades more and more into industrial environments forming the so-called Industrial IoT (IIoT). Especially in industrial environments such as smart factories, the quality of data that IoT devices provide is highly relevant. However, current frameworks for managing the IoT and exchanging data do not provide data quality (DQ) metrics. Pervasive applications deployed in the factory need to know how data are "good" for use. However, the DQ requirements differ from a process to another. Actually, specifying/expressing DQ requirements is a subjective task, depending on the specific needs of each targeted application. As an example this could mean how accurate a location of an object that is provided by an IoT system differs from the actual physical position of the object. A Data Quality of 100% could mean that the value represents the actual position. A Data Quality of 0% could mean that the object is not at the reported position. In this example, the value 0% or 100% can be given by a specific software module that is able to filter raw data sent to the IoT system and to deliver the appropriate metric for Dev apps. Building ad hoc solutions for DQ management is perfectly acceptable. But the challenge of writing and deploying applications for the Internet of Things remains often understated. We believe that new approaches are needed, for thinking DQ management in the context of extremely dynamic systems that is the characteristic of the IoT.

In 2019, we introduced DQ to the IoT by (1) representing data quality parameters as metadata to each stored and exchanged IoT data item and (2) providing a toolbox that helps developers to assess the data quality of their processed data using the previously introduced data quality metadata. We followed an inductive approach. Therefore, we set up a pilot to gain first-hand experience with DQ, and to test our developed tools. Our pilot focuses on multi-source data inconsistency. Our setting consists of multiple industrial robots that cowork within a factory. The robots on the line follow a fixed path while the other two robots can freely move. For our implementation we use a data-centric IoT middleware, the Virtual State Layer (VSL). It provides many desired properties such as security and dynamic coupling of services at runtime. Most important it has a strong semantic model for representing data that allows adding new metadata for data quality easily. In our pilot the decrease of the DQ is caused by a low periodicity of location reports. We implemented a DQ service that infers the DQ being located in the service chain. The coordination service queries our DQ enriching service. The DQ enrichment service models the behavior of a robot and infers the resulting DQ depending on the time between the location report and the coordination service's query. Our goal was not only to report the DQ to the consuming service but also to offer tools (microservices) to mitigate from bad DQ. To enable a mitigation from the decreasing DQ, we started the sensors at a random time. This results in the same precision decrease periodicity but in shifted reporting times. The shift enables increasing the DQ by using sensor fusion and data filtering.

FOCUS Project-Team

7. New Results

7.1. Service-Oriented and Cloud Computing

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

7.1.1. Service-Oriented Computing and Internet of Things

Session types, i.e. types for structuring service communication, are recently being integrated into mainstream programming languages. In practice, a very important notion for dealing with such types is that of subtyping, since it allows for typing larger classes of system, where a program has not precisely the expected behavior but a similar one. We recently showed that, when asynchronous communication is considered, unfortunately, such a subtyping relation is undecidable. In [27] we present an algorithm (the first one that does not restrict type syntax or limit communication) and a tool for checking asynchronous subtyping which is sound, but not complete: in some cases it terminates without returning a final verdict. In [29] we discuss the relationship between session types and service behavioural contracts and we show the existence of a fully abstract interpretation of session types into a fragment of contracts, mapping subtyping into binary compliance-preserving contract refinement. This also yields an original undecidability result for asynchronous contract refinement.

In [43] we elaborate on our previous work on choreographies, which specify in a single artefact the expected behaviour of all the participants in a service oriented system. In particular, we extend dynamic choreographies, which model system updates at runtime, with the feature of dynamic inclusion of new unforeseen participants. In [30] we propose, in the context of platooning (a freight organization system where a group of vehicles follows a predefined trajectory maintaining a desired spatial pattern), a two layered, composable technical solution for federated platooning: a decentralized overlay network that regulates the interactions among the stakeholders, useful to mitigate issues linked to data safety and trustworthiness; and a dynamic federation platform, needed to monitor and interrupt deviant behaviors of federated members.

Finally, in [44] we focused on the use of our service-oriented language Jolie in an Internet of Things (IoT) setting. Technically, a key feature of Jolie is that it supports uniform linguistic abstractions to exploit heterogeneous communication stacks, i.e. for service oriented computing, protocols such as TCP/IP, Bluetooth, and RMI at transport level, and HTTP and SOAP at application level. We extend Jolie in order to support, uniformly as well, also the two most adopted protocols for IoT communication, i.e. CoAP and MQTT, and we report our experience on a case study on home automation.

7.1.2. Cloud Computing

In [18] we investigate the problem of modeling the optimal and automatic deployment of cloud applications and we experiment such an approach by applying it to the Abstract Behavioural Specification language ABS. In [28] we show that automated deployment, proven undecidable in the general case, is, instead, algorithmically treatable for the specific case of microservices: we implement an automatic optimal deployment tool and compute deployment plans for a realistic microservice architecture. In [35] we propose a core formal programming model (combining features from λ-calculus and π-calculus) for serverless computing, also known as Functions-as-a-Service: a recent paradigm aimed at simplifying the programming of cloud applications. The idea is that developers design applications in terms of functions and the infrastructure deals automatically with cloud deployment in terms of distribution and scaling.

7.2. Models for Reliability

Participants: Ivan Lanese, Doriana Medic.

7.2.1. Reversibility

We have continued the study of reversibility started in the past years. First, we continued to study reversibility in the context of the Erlang programming language. In particular, we devised a technique to record a program execution and replay it [37] inside the causal-consistent reversible debugger for Erlang we developed in the last years. More precisely, we may not replay the exact same execution, but any execution which is causal-consistent to it. We proved that this is enough to replay misbehaviours, hence to look for the bugs causing them. Second, we compared [48] various approaches to causal-consistent reversibility in CCS and π-calculus. In CCS, we showed that the two main approaches for causal-consistent reversibility, namely the ones of RCCS [51] and of CCSk [55] give rise to isomorphic LTSs (up to some structural rules). In π-calculus, we showed that one can define a causal semantics for π-calculus parametric on the data structure used to track extruded names, and that different instances capture causal semantics from the literature. All such semantics can be used to define (different) causal-consistent reversible semantics. As a final contribution, we studied reversibility in the context of Petri nets [41]. There, we do not considered causal-consistent reversibility, but a notion of local reversibility typical of Petri nets. In particular, we say that a transition is reversible if one can add a set of effect-reverses (an effect-reverse, if it can trigger, undoes the effect of the transition) to undo it in each marking reachable by it, without changing the set of reachable markings. We showed that, contrarily to what happens in bounded nets, transition reversibility is not decidable in general unbounded nets. It is however decidable in some significant subclasses of Petri nets, in particular all transitions of cyclic nets (nets where the initial marking is reachable from any state) are reversible. Finally, we show how to restructure nets by adding new places so to make their transitions reversible without altering their behaviour.

7.3. Probabilistic Systems and Resource Control

Participants: Martin Avanzini, Mario Bravetti, Raphaelle Crubillé, Ugo Dal Lago, Francesco Gavazzo, Gabriele Vanoni, Akira Yoshimizu.

7.3.1. Probabilistic Programming and Static Analysis

In FoCUS, we are interested in studying probabilistic higher-order programming languages and, more generally, the fundamental properties of probabilistic computation when placed in an interactive scenario, for instance concurrency. One of the most basic but nevertheless desirable properties of programs is of course termination. Termination can be seen as a minimal guarantee about the time complexity of the underlying program. When probabilistic choice comes into play, termination can be defined by stipulating that a program is terminating if its probability of convergence is 1, this way giving rise to the notion of *almost sure termination*. Alternatively, a probabilistic program is said to be *positively* almost surely terminating if its average runtime is finite. The latter condition easily implies the former. Termination, already undecidable for deterministic (universal) programming languages, remains so in the presence of probabilistic choice, even becoming provably harder.

The FoCUS team has been the first in advocating the use of types to guarantee probabilistic termination, in the form of a monadic sized-type system [17]. Developed in collaboration with Grellois by Dal Lago, this system substantially generalises usual sized-types, and allows this way to capture probabilistic, higher-order programs which terminate almost surely. Complementary, in collaboration with Ghyselen, Avanzini and Dal Lago have recently defined a formal system for reasoning about the *expected runtime* of higher-order probabilistic programs, through a *refinement type system* capable of *modeling probabilistic effects* with exceptional accuracy [26]. To the best of our knowledge, this provides the first formal methodology for *average case complexity analysis* of higher-order programs. Remarkably, the system is also *extensionally complete*.

In 2018, we have started to investigate the foundations for *probabilistic abstract reduction systems* (*probabilistic ARSs*), which constitute a general framework to study fundamental properties of probabilistic computations, such as termination or confluence. In 2019, we have significantly revised this initial development [11]. Particularly, we have refined Lyapunov ranking functions by conceiving them as *probabilistic embeddings*. The ramifications of this work are two-fold. First, we obtain a sound and complete method for reasoning about strong positive almost sure termination. Second, this method has been instantiated in the setting of (first-order) *probabilistic rewrite systems*, giving rise to the notion of *barycentric algebras*, generalising the well-known interpretation method. Barycentric algebras have been integrated in the termination prover *NaTT*[0], confirming the feasibility of the approach.

We have also worked on higher-order model checking as a way to prove termination of probabilsitic variations on higher-order recursion schemes [36], obtaining encouraging results. More specifically, an algorithm for approximating the probability of convergence of any such scheme has been designed and proved sound, although the problem of precisely computing the probability of convergence is shown to be undecidable at order 2 or higher. Finally, we have published a new version of a contribution we wrote in 2017 about how implicit computational complexity could help in proving that certain cryptographic constructions have the desired complexity-theoretic properties [12].

7.3.2. Higher-Order end Effectful Programs: Relational Reasoning

In FoCUS, we are also interested in relational reasoning about programs written in higher-order programming languages. In the recent years, this research has been directed to effectful programs, namely programs whose behaviour is not purely functional. Moreover, there has recently been a shift in our interests, driven by the projects REPAS and DIAPASoN, towards quantitative kinds of relational reasoning, in which programs are not necessarily dubbed equivalent (or not), but rather put at a certain distance.

The first contribution we had in this direction is due to Dal Lago and Gavazzo [31], who generalized the so-called open normal-form bisimilarity technique to higher-order programs exhibiting any kind of monadic effect. The key ingredient here is that of a relator, and allows to lift relations on a set to relations on monadic extensions to the same set. This allows to define open normal-form bisimilarity, and to prove it correct. This, together, with other contributions, have also appeared in Gavazzo's PhD Thesis, which has been successfully defended in April 2019 [10], and which has been awarder the Prize for the Best PhD Thesis in Theoretical Computer Science by the Italian Chapter of the EATCS.

We have also given the notion of differential logical relations [33], a generalization of Plotkin's logical relations in which programs are dubbed being at a certain *distance* rather than being just *equivalent*. Noticeably, this distance is not necessarily numeric, but is itself functional if the compared programs have a non-ground type. This allows to evaluate the distance between programs taking into account the possible actions the environment can make on the compared programs.

7.3.3. Alternative Probabilistic Models

We are also interested in exploring probabilistic models going beyond the usual ones, in which determinisitic programming languages are endowed with discrete probabilistic choice.

We have first of all studied bayesian λ-calculi, namely λ-calculi in which not only an operator for probabilistic choice is available, but also one for *scoring*, which serves as the basis to model conditioning in probabilistic programming. We give a geometry of interaction model for such a typed λ-calculus [34], namely a paradigmatic calculus for higher-order Bayesian programming in the style of PCF. The model is based on the category of measurable spaces and partial measurable functions, and is proved adequate with respect to both a distribution-based and a sampling-based operational semantics.

We have also introduced a probabilistic extension of a framework to specify and analyze software product lines [15]. We define a syntax of the language including probabilistic operators and define operational and denotational semantics for it. We prove that the expected equivalence between these two semantic frameworks holds. Our probabilistic framework is supported by a set of scripts to show the model behavior.

7.4. Verification Techniques

Participants: Ugo Dal Lago, Adrien Durier, Daniel Hirschkoff, Ivan Lanese, Cosimo Laneve, Davide Sangiorgi, Akira Yoshimizu, Gianluigi Zavattaro.

[0] See https://www.trs.css.i.nagoya-u.ac.jp/NaTT/.

Extensional properties are those properties that constrain the behavioural descriptions of a system (i.e., how a system looks like from the outside). Examples of such properties include classical functional correctness, deadlock freedom and resource usage.

In the last year of the Focus project, we have worked on three main topics: (*i*) *name mobility and coinductive techniques*, (*ii*) *deadlock analysis*, and (*iii*) *cost analysis of properties* of languages for actors and for smart contracts.

7.4.1. Name Mobility and Coinductive Techniques

In [19], we propose proof techniques for bisimilarity based on unique solution of equations. The results essentially state that an equation (or a system of equations) whose infinite unfolding never produces a divergence has the unique-solution property. We distinguishing between different forms of divergence; derive an abstract formulation of the theorems, on generic LTSs; adapt the theorems to other equivalences such as trace equivalence, and to preorders such as trace inclusion; we compare the resulting techniques to enhancements of the bisimulation proof method (the 'up-to techniques'). In [20], we study how to adapt such techniques to higher-order languages. In such languages proving behavioural equivalences is known to be hard, because interactions involve complex values, namely terms of the language. The soundness of proof techniques is usually delicate and difficult to establish. The language considered is the Higher-Order π-calculus.

The contribution [42] studies the representation of the call-by-need λ-calculus in the pure message-passing concurrency of the π-calculus, precisely the Local Asynchronous π-calculus, that has sharper semantic properties than the ordinary π-calculus. We exploit such properties to study the validity of of β-reduction (meaning that the source and target terms of a beta-reduction are mapped onto behaviourally equivalent processes). Nearly all results presented fail in the ordinary π-calculus.

In [45], we investigate basic properties of the Erlang concurrency model. This model is based on asynchronous communication through mailboxes accessed via pattern matching. In particular, we consider Core Erlang (which is an intermediate step in Erlang compilation) and we define, on top of its operational semantics, an observational semantics following the approach used to define asynchronous bisimulation for the π-calculus. Our work allows us to shed some light on the management of process identifiers in Erlang, different from the various forms of name mobility already studied in the literature. In fact, we need to modify standard definitions to cope with such specific features of Erlang.

The paper [25] reviews the origins and the history of enhancements of the bisimulation and coinduction proof methods.

7.4.2. Deadlock Analysis

The contributions [22] and [50] address deadlock analysis of `Java`-like programs. The two papers respectively cover two relevant features of these languages: (*i*) multi-threading and reentrant locks and (*ii*) co-ordination primitives (`wait`, `notify` and `notifyAll`). In both cases, we define a behavioral type system that associates abstract models to programs (lams and Petri Nets with inhibitor arcs) and define an algorithm for detecting deadlocks. The two systems are consistent and our technique is intended to be an effective tool for the deadlock analysis of programming languages.

The paper [16] addresses the π-calculus. It defines a type system for guaranteeing that typable processes never produce a run-time error and, even if they may diverge, there is always a chance for them to finish their work, i.e., to reduce to an idle process (a stronger property than deadlock freedom). The type system uses so-called *non-idempotent intersections* and, therefore, applies to a large class of processes. Indeed, despite the fact that the underlying property is $\prod_2^0$-complete, there is a way to show that the system is complete, i.e., that any well-behaved process is typable, although for obvious reasons infinitely many derivations need to be considered.

7.4.3. Static Analysis of Properties of Concurrent Programs

We have analyzed the computational time of actor programs, following a technique similar to [52], and we have begun a new research direction that deals with the analysis of `Solidity` smart contracts.

In [23], we propose a technique for estimating the computational time of programs in an actor model. To this aim, we define a compositional translation function returning cost equations, which are fed to an automatic off-the-shelf solver for obtaining the time bounds. Our approach is based on so-called *synchronization sets* that capture possible difficult synchronization patterns between actors and helps make the analysis efficient and precise. The approach is proven to correctly over-approximate the worst computational time of an actor model of concurrent programs. The technique is complemented by a prototype analyzer that returns upper bound of costs for the actor model.

In [38], we analyze the bahaviour of smart contracts, namely programs stored on some blockchain that control the transfer of assets between parties under certain conditions. In particular, we focus on the interactions of smart contracts and external actors (usually, humans) in order to maximize objective functions. 5 To this aim, we define a core language of programs, which is reminiscent of `Solidity`, with a minimal set of smart contract primitives and we describe the whole system as a parallel composition of smart contracts and users. We therefore express the system behaviour as a first order logic formula in Presburger arithmetics and study the maximum profit for each actor by solving arithmetic constraints.

7.5. Computer Science Education

Participants: Michael Lodi, Simone Martini.

We study why and how to teach computer science principles (nowadays often referred to as “computational thinking”, CT), in the context of K-12 education. We are interested in philosophical, sociological, and historical motivations to teach computer science. Furthermore, we study what concepts and skills related to computer science are not only technical abilities, but have a general value for all students. Finally, we try to find/produce/evaluate suitable materials (tools, languages, lesson plans...) to teach these concepts, taking into account: difficulties in learning CS concepts (particularly programming); stereotypes about computer science (teachers’ and students’ mindset); teacher training (both non-specialist and disciplinary teachers); innovative teaching methodologies (primarily based on constructivist and constructionist learning theories).

7.5.1. Computational Thinking, Unplugged Activities, and Constructionism

We reviewed some relevant literature related to learning CS and, more specifically, programming in a constructivist and constructionist light. We investigated some cognitive aspects, for example, the notional machine and its role in understanding, misunderstanding, and difficulties of learning to program. We reviewed programming languages for learning to program, with particular focus on educational characteristics of block-based languages [24].

We analyzed the widespread but debated pedagogical approach of “unplugged activities”: activities without a computer, like physical games, used to teach CS concepts. We explicitly connect computational thinking to the “CS Unplugged” pedagogical approach, by analyzing a representative sample of CS Unplugged activities in light of CT. We found the activities map well onto commonly accepted CT concepts, although caution must be taken not to regard CS Unplugged as being a complete approach to CT education [14].

Moreover, we found similarities (e.g., kinesthetic activities) and differences (e.g., structured vs. creative activities) between Unplugged and constructivism or constructionism. We argue there is a tension between the constructivist need to link the CS concepts to actual implementations and the challenge of teaching CS principles without computers, to undermine the misconceptions of CS as “the science of computers” [13].

7.5.2. CS in Primary School

We designed, produced and implemented in a primary school some “unplugged + plugged” teaching materials and lesson plans [47]. The unplugged activities are structured as an incremental discovery, scaffolded by the instructors, of the fundamental concepts of structured programming (e.g., sequence, conditionals, loops, variables) but also complexity in terms of computational steps and generalization of algorithms. The plugged activities follow the creative learning approach, using Scratch as the primary tool, both for free creative expression and for learning other disciplines (e.g., drawing regular polygons).

7.5.3. *Growth Mindset and Transfer*

Every person holds an idea (mindset) about intelligence: someone thinks it is a fixed trait, like eye colour (fixed mindset), while others believe it can grow like muscles (growth mindset). The latter is beneficial for students to have better results, particularly in STEM disciplines, and to not being influenced by stereotypes. Computer science is a subject that can be affected by fixed ideas (“geek gene”), and some (small) studies showed it can induce fixed ideas. By contrast, some claims stating that studying CS can foster a GM have emerged. However, educational research shows that the transfer of competences is hard. We measured [40] some indicators (e.g., mindset, computer science mindset) at the beginning and the end of a high school year in different classes, both CS and non-CS oriented. At the end of the year, none of the classes showed a statistically significant change in their mindset. Interestingly, non-CS oriented classes showed a significant decrease in their computer science growth mindset, which is not desirable.

7.6. Constraint Programming

Participants: Maurizio Gabbrielli, Liu Tong.

In Focus, we sometimes make use of constraint solvers (e.g., cloud computing, service-oriented computing). Since a few years we have thus began to develop tools based on constraints and constraint solvers.

In [39] we have used constraints in the setting of Service Function Chaining (SFC) deployment. SFCs represent sequences of Virtual Network Functions that compose a service. They are found within Network Function Virtualization (NFV) and Software Defined Networking (SDN) technologies, that recently acquired a great momentum thanks to their promise of being a flexible and cost-effective solution for replacing hardware-based, vendor-dependent network middleboxes with software appliances running on general purpose hardware in the cloud.

We employ constraint programming to solve the SFC design problem. Indeed we argue that constraint programming can be effectively used to address this kind of problems because it provides expressive and flexible modeling languages which come with powerful solvers, thus providing efficient and scalable performance.

INDES Project-Team

6. New Results

6.1. JavaScript Implementation and Browser Security

We have pursued the development of *Hop* and our study on efficient and secure JavaScript implementations.

6.1.1. JavaScript Property Caches

JavaScript objects are dynamic. At any moment of their lifetime, properties can be added or deleted. In principle a property access requires a lookup in the object itself, and, possibly, in all the objects forming its prototype chain. All fast JavaScript implementations deploy strategies to implement this lookup operation in nearly constant time. They generally rely on two ingredients: *hidden classes* and *property caches*. Hidden classes describe object memory layouts. Property caches use these descriptions to access objects directly, avoiding the normal name lookup operations. Hidden classes and property caches make property accesses comparable in speed to field accesses of traditional languages like C and Java.

Hidden classes and property caches are not new. They were invented for Self, the first dynamically typed prototype-based languages, following Smalltalk's idea that already used caches at that time for optimizing method calls. For the past ten years they have enjoyed a revival of interest after it was shown how effective they are at improving Object-Oriented languages performance in general and specially JavaScript. Today most JavaScript implementations such as V8, JavaScriptCode, and SpiderMonkey use them. Hidden classes and property caches apply in specific situations, which unfortunately means that some accesses are unoptimized or not treated very efficiently.

1. **Property addition problem**: hidden classes support the accesses of existing properties but they do not handle efficiently property addition commonly found in object constructors.
2. **Prototype properties problem**: hidden classes and property caches optimize accesses of properties directly stored in the object. They do not optimize accesses of properties stored in one of the objects composing the prototype chain.
3. **Polymorphic properties problem**, as property caches require strict hidden class equivalence for optimizing accesses, polymorphic data structures and polymorphic method invocations need special treatment to not be left unoptimized. This has been addressed by the *Polymorphic Inline Cache* technique proposed by Holzle *et al.* in previous studies, which resorts to a dynamic search in the cache history. As a linear or binary search is involved, it is not as efficient as plain property caches.

Problem 1 is critical for all existing JavaScript programs as it impacts the performance of object construction. Problems 2 and 3 will become prominent with the advent of ECMAScript 6 class-like programming style that is backed up by object prototypes. We propose solutions to these problems. At the cost of one extra test inserted at each property access, we optimize prototype property accesses. Resorting to a static analysis, we propose a technique that we call *speculative caches* for optimizing object construction.

Trading memory space for speed, we propose *cache property tables* that enable accessing polymorphic objects in constant time. For the analogy with C++ virtual tables we call these cache tables *vtables*.

We have implemented these techniques in *Hopc*, the *Hop* static JavaScript compiler and we have presented them in a conference publication [17]. We have shown how the complement and enhance property caches used for accessing object properties of JavaScript like languages. We have shown that they take over classical caches when the searched property is either stored in an object of the prototype chain or defined using accessors. They also support efficiently polymorphic and megamorphic property accesses. Finally, they also support efficient object extensions. These techniques do not apply as frequently as simple property caches that cover a vast majority of accesses. However, since they impose no overhead when not used, they can be integrated in any existing system at no run time cost. We have validated the approach with an experimental report based that shown that the presented techniques improve performance in situations where simple cache miss.

6.1.2. Secure JavaScript

Whereas the dynamic nature of JavaScript plays an essential role in the advantages it offers for easy and fast development, a malicious JavaScript program can easily break the integrity and confidentiality of a web or IoT application. JavaScript dynamic semantics and sharing are deeply intricated and attacker code can trivially exploit these.

We have developed a compiler, called SecureJS to offer security guarantees for JavaScript on clients, servers, and IoT devices. Our compiler is applicable to ECMAScript 5th legacy code, which in particular means that we allow for built-in JavaScript functions. Moreover, we go beyond the JavaScript language and handle a common web API, XMLHttpRequest module. The challenge is to cover most of the JavaScript language efficiently while providing strong security guarantees. For the latter, we formally define and prove the compiler's security guarantees by means of a new security property, coined as *dynamic delimited release*, for JavaScript integrity and confidentiality.

Compiled programs can be effortlessly deployed in client, server, and IoT JavaScript environments and do not require an external isolation mechanism to preserve integrity and confidentiality.

We have validated SecureJS experimentally using ECMAScript Test262 test suits. First, we have shown that SecureJS preserves the correct SecureJS semantics. Second, we have shown that it successfully implements the memory isolation needed to enforce the security property.

The current SecureJS implementation as been architectured to support low-power platforms that only supports ECMAScript 5. In the future we plan to accommodate more recent version of JavaScript for the platforms that supports it. This will extend the possibility of communications between trusted and untrusted codes and this will enable more efficient implementation techniques. A paper describing this work is currently under submission.

6.1.3. Empowering Web Applications with Browser Extensions

Browser extensions are third party programs, tightly integrated to browsers, where they execute with elevated privileges in order to provide users with additional functionalities. Unlike web applications, extensions are not subject to the Same Origin Policy (SOP) and therefore can read and write user data on any web application. They also have access to sensitive user information including browsing history, bookmarks, credentials (cookies) and list of installed extensions. They have access to a permanent storage in which they can store data as long as they are installed in the user's browser. They can trigger the download of arbitrary files and save them on the user's device. For security reasons, browser extensions and web applications are executed in separate contexts. Nonetheless, in all major browsers, extensions and web applications can interact by exchanging messages. Through these communication channels, a web application can exploit extension privileged capabilities and thereby access and exfiltrate sensitive user information.

We have analyzed the communication interfaces exposed to web applications by Chrome, Firefox and Opera browser extensions [18]. As a result, we identified many extensions that web applications can exploit to access privileged capabilities. Through extensions' APIS, web applications can bypass SOP and access user data on any other web application, access user credentials (cookies), browsing history, bookmarks, list of installed extensions, extensions storage, and download and save arbitrary files in the user's device. Our results demonstrate that the communications between browser extensions and web applications pose serious security and privacy threats to browsers, web applications and more importantly to users. We discuss countermeasures and proposals, and believe that our study and in particular the tool we used to detect and exploit these threats, can be used as part of extensions review process by browser vendors to help them identify and fix the aforementioned problems in extensions.

6.2. Timing-side channels attacks

We have pursued our studies on foundations of language-based security following two axes on timing-side channels research:

6.2.1. Speculative constant time

The most robust way to deal with timing side-channels in software is via *constant-time* programming—the paradigm used to implement almost all modern cryptography. Constant-time programs can neither branch on secrets nor access memory based on secret data. These restrictions ensure that programs do not leak secret information via timing side channels, at least on hardware *without* microarchitectural features. However, microarchitectural features are a major source of timing side channels as the growing list of attacks (Spectre, Meltdown, etc) is showing. Moreover code deemed to be constant-time in the usual sense may in fact leak information on processors with microarchitectural features. Thus the decade-old constant-time recipes are no longer enough. We lay the foundations for constant-time in the presence of micro-architectural features that have been exploited in recent attacks: out-of-order and speculative execution. We focus on constant-time for two key reasons. First, *impact*: constant-time programming is largely used in narrow, high-assurance code—mostly cryptographic implementations—where developers already go to great lengths to eliminate leaks via side-channels. Second, *foundations:* constant-time programming is already rooted in foundations, with well-defined semantics. These semantics consider very powerful attackers have control over the cache and the scheduler. A nice effect of considering powerful attackers is that the semantics can already overlook many hardware details—e.g., since the cache is adversarially controlled there is no point in modeling it precisely—making constant-time amenable to automated verification and enforcement.

We have first defined a semantics for an abstract, three-stage (fetch, execute, and retire) machine. This machine supports out-of-order and speculative execution by modeling *reorder buffers* and *transient instructions*, respectively. Our semantics assumes that attackers have complete control over microarchitectural features (e.g., the branch target predictor), and uses adversarial execution *directives* to model adversary's control over predictors. We have then defined *speculative constant-time*, the counterpart of *constant-time* for machines with out-of-order and speculative execution. This definition has allowed us to discover microarchitectural side channels in a principled way—all four classes of Spectre attacks as classified by Canella et al., for example, manifest as violation of our constant-time property. Our semantics even revealed a new Spectre variant, that exploits the aliasing predictor. The variant can be disabled by unsetting a flag, by illusttrates the usefulness of our semantics. This study is described in a paper currently submitted.

6.2.2. Remote timing attacks

A common approach to deal with timing attacks is based on preventing secrets from affecting the execution time, thus achieving security with respect to a strong, *local* attacker who can measure the timing of program runs. Another approach is to allow branching on secrets but prohibit any subsequent attacker-visible side effects of the program. It is sometimes used to handle *internal timing* leaks, i.e., when the timing behavior of threads affects the interleaving of attacker-visible events via the scheduler.

While these approaches are compatible with strong attackers, they are highly restrictive for program runs as soon as they branch on a secret. It is commonly accepted that "adhering to constant-time programming is hard" and "doing so requires the use of low-level programming languages or compiler knowledge, and forces developers to deviate from conventional programming practices".

This restrictiveness stems from the fact that there are many ways to set up timing leaks in a program. For example, after branching on a secret the program might take different time in the branches because of: (i) more time-consuming operations in one of the branches, (ii) cache effects, when in one of the branches data or instructions are cached but not in the other branch, (iii) garbage collection (GC) when in one of the branches GC is triggered but not in the other branch, and (iv) just-in-time (JIT) compilation, when in one of the branches a JIT-compiled function is called but not in the other branch. Researchers have been painstakingly addressing these types of leaks, often by creating mechanisms that are specific to some of these types. Because of the intricacies of each type, addressing their combination poses a major challenge, which these approaches have largely yet to address.

This motivates a general mechanism to tackle timing leaks independently of their type. However, rather than combining enforcement for the different types of timing leaks for strong local attackers, is there a setting where the capabilities of attackers are perhaps not as strong, enabling us to design a general and less restrictive mechanism for a variety of timing attacks with respect to a weaker attacker?

We focus on timing leaks under *remote* execution. A key difference is that the remote attacker does not generally have a reference point of when a program run has started or finished, which significantly restricts attacker capabilities.

We illustrate remote timing attacks by two settings: a server-side setting of IoT apps where apps that manipulate private information run on a server and a client-side setting where e-voting code runs in a browser.

IFTTT (If This Then That), Zapier, and Microsoft Flow are popular IoT platforms driven by enduser programming. App makers publish their apps on these platforms. Upon installation apps manipulate sensitive information, connecting cyberphysical "things" (e.g., smart homes, cars, and fitness armbands) to online services (e.g., Google and Dropbox) and social networks (e.g., Facebook and Twitter). An important security goal is to prevent a malicious app from leaking private information of a user to the attacker.

Recent research identifies ways to leak private information by IoT apps and suggests tracking information flows in IoT apps to control these leaks. The suggested mechanisms perform data-flow (*explicit*) and control-flow (*implicit*) tracking. Unfortunately, they do not address timing leaks, implying that a malicious app maker can still exfiltrate private information, even if the app is subject to the security restrictions imposed by the proposed mechanisms.

In addition, Verificatum, an advanced client-side cryptographic library for e-voting motivates the question of remote timing leaks with respect to attackers who can observe the presence of encrypted messages on the network.

This leads us to the following general research questions:

1. What is the right model for remote timing attacks?
2. How do we rule out remote timing leaks without rejecting useful secure programs?
3. How do we generalize enforcement to multiple security levels?
4. How do we harden existing information flow tools to track remote timing leaks?
5. Are there case studies to give evidence for the feasibility of the approach?

To help answering these questions, we propose an extensional knowledge-based security characterization that captures the essence of remote timing attacks. In contrast to the local attacker that counts execution steps/time since the beginning of the execution, our model of the remote attacker is only allowed to observe inputs and outputs on attacker-visible channels, along with their timestamps. At the same time, the attacker is in charge of the potentially malicious code with capabilities to access the clock, in line with assumptions about remote execution on IoT app platforms and e-voting clients.

A timing leak is typically enabled by branching on a secret and taking different time or exhibiting different cache behavior in the branches. However, as discussed earlier, it is desirable to avoid restrictive options like forcing the execution to take constant time, prohibiting attacker-visible output any time after the branching, or prohibiting branching on a secret in the first place.

Our key observation is that for a remote attacker to successfully set up and exploit a timing leak, program behavior must follow the following pattern: (i) branching on a secret takes place in a program run, and either (ii-a) the branching is followed by more than one attacker-visible I/O event, or (ii-b) the branching is followed by one attacker-visible I/O event, and prior to the branching there is either an attacker-visible I/O event or a reading to the clock.

Based on this pattern, we design Clockwork, a monitor that rules out timing leaks. Our mechanism pushes for permissiveness. For example, runs (free of explicit and implicit flows) that do not access the clock and only have one attacker-visible I/O event are accepted.

Runs that do not perform attacker-visible I/O after branching on a secret are accepted as well. As we will see, these kinds of runs are frequently encountered in secure IoT and e-voting apps.

We implement our monitor for JavaScript, leveraging JSFlow, a state-of-the-art information flow tracker for JavaScript. We demonstrate the feasibility of the approach on a case study with IFTTT, showing how to prevent malicious app makers from exfiltrating users' private information via timing, and a case study with Verificatum, showing how to track remote timing attacks with respect to network attackers. Our case studies demonstrate both the security and permissiveness. While apps with timing leaks are rejected, benign apps that use clock and I/O operations in a non-trivial fashion are accepted.

6.3. Security analysis of ElGamal implementations

Throughout the last century, especially with the beginning of public key cryptography due to Diffie-Hellman, many cryptographic schemes have been proposed. Their security depends on mathematically complex problems such as integer factorization and discrete logarithm. In fact, it is thought that a cryptographic scheme is secure if it resists cryptographic attacks over a long period of time. On one hand, since certain schemes may take several years before being widely studied in depth, they become vulnerable as time passes. On the other hand, a cryptographic scheme is a provable one, if it resists cryptographic attacks relying on mathematical hypothesis.

Being easily adaptable to many kinds of cryptographic groups, the ElGamal encryption scheme enjoys homomorphic properties while remaining semantically secure , provided that the Decisional Diffie-Hellman (DDH) assumption holds on the chosen group. While the homomorphic property forbids resistance against chosen ciphertext attacks, it is very convenient for voting systems. The ElGamal encryption scheme is the most extensively used alternative to RSA, and it is the homomorphic encryption scheme almost exclusively used for voting systems. Moreover, ElGamal is the only homomorphic encryption scheme implemented by default in many hardware security modules.

In order to be provable secure, ElGamal encryption needs to be implemented on top of a group verifying the Decisional Diffie-Hellman (DDH) assumption. Since this assumption does not hold for all groups, one may have to wrap an encoding and a decoding phase to ElGamal to be able to have a generic encryption scheme.

We have submitted a paper that studies ElGamal encryption scheme libraries in order to identify which implementations respect the DDH assumption. The paper presents an analysis of 25 libraries that implement ElGamal encryption scheme in the wild. We focus our analysis on understanding whether the DDH assumption is respected in these implementations, ensuring a secure scheme in which no information about the original message could be leaked. The DDH assumption is crucial for the security of ElGamal because it ensures indistinguishability under chosen-plaintext attacks (IND-CPA). Without the DDH assumption, encryption mechanisms may leak one bit of information about the plaintext and endager the security of the electoral system as one bit has the ability to completely invalidate privacy in an election. One way to comply with the DDH assumption is by using groups of prime order. In particular, when adopting safe primes, one can ensure the existence of a *large* prime order subgroup and restrict messages to belong to this subgroup. Mapping plaintexts into subgroups is called message encoding. Such encoding necessitates to be efficient and precisely invertible to allow decoding after the decryption.

Our results show that out of 25 analyzed libraries, 20 are wrongly implemented because they do not respect the conditions to achieve IND-CPA security under the DDH assumption. This means that encryptions using ElGamal from any of these 20 libraries leak one bit of information.

From the 5 libraries which respect the DDH assumption, we also study and compare various encoding and decoding techniques. We identify four different message encoding and decoding techniques and discuss the different designs and conclude which implementation is more efficient for voting systems.

6.4. Measurement and Detection of Web Tracking

6.4.1. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels

The Web has become an essential part of our lives: billions are using Web applications on a daily basis and while doing so, are placing *digital traces* on millions of websites. Such traces allow advertising companies, as well as data brokers to continuously profit from collecting a vast amount of data associated to the users.

Web tracking has been extensively studied over the last decade. To detect tracking, most of the research studies and user tools rely on *consumer protection lists*. EasyList [0] and EasyPrivacy [0] (EL&EP) are the most popular publicly maintained blacklist of know advertising and tracking domains, used by the popular browser extensions AdBlock Plus [0] and uBlockOrigin [0]. Disconnect [0] is another very popular list for detecting domains known for tracking, used in Disconnect browser extension [0] and in integrated tracking protection of Firefox browser. Relying on EL&EP or Disconnect became the *de facto* approach to detect third-party tracking requests in privacy and measurement community. However it is well-known that these lists detect only known tracking and ad-related requests, and a tracker can easily avoid this detection by registering a new domain or changing the parameters of the request.

Our contributions: To evaluate the effectiveness of filter lists, we propose a new, fine-grained behavior-based tracking detection. Our results are based on a stateful dataset of 8K domains with a total of 800K pages generating 4M third-party requests. We make the following contributions:

- *We analyse all the requests and responses that lead to invisible pixels (by "invisible pixels" we mean* 1×1 *pixel images or images without content).* Pixels are routinely used by trackers to send information or third-party cookies back to their servers: the simplest way to do it is to create a URL containing useful information, and to dynamically add an image tag into a webpage. This makes invisible pixels *the perfect suspects for tracking* and propose a new classification of tracking behaviors. Our results show that pixels are still widely deployed: they are present on more than 94% of domains and constitute 35.66% of all third-party images. We found out that pixels are responsible only for 23.34% of tracking requests, and the most popular tracking content are scripts: a mere loading of scripts is responsible for 34.36% of tracking requests.
- *We uncover hidden collaborations between third parties.* We applied our classification on more than 4M third-party requests collected in our crawl. We have detected new categories of tracking and collaborations between domains. We show that domains sync first party cookies through a *first to third party cookie syncing*. This tracking appears on 67.96% of websites.
- *We show that filter lists miss a significant number of cookie-based tracking.* Our evaluation of the effectiveness of EasyList&EasyPrivacy and Disconnect lists shows that they respectively miss 25.22% and 30.34% of the trackers that we detect. Moreover, we find that if we combine all three lists, 379,245 requests originating from 8,744 domains still track users on 68.70% of websites.
- *We show that privacy browser extensions miss a significant number of cookie-based tracking.* By evaluating the popular privacy protection extensions: Adblock, Ghostery, Disconnect, and Privacy Badger, we show that Ghostery is the most efficient among them and that all extensions fail to block at least 24% of tracking requests.

This paper [15] has been accepted for publication at the Privacy Enhancing Technologies Symposium (PETs) 2020.

6.4.2. A survey on Browser Fingerprinting

This year, we have conducted a survey on the research performed in the domain of browser fingerprinting, while providing an accessible entry point to newcomers in the field. We explain how this technique works and where it stems from. We analyze the related work in detail to understand the composition of modern fingerprints and see how this technique is currently used online. We systematize existing defense solutions into different categories and detail the current challenges yet to overcome.

[0] https://easylist.to/
[0] https://easylist.to/easylist/easyprivacy.txt
[0] https://adblockplus.org/
[0] https://github.com/gorhill/uBlock
[0] https://disconnect.me/trackerprotection/blocked
[0] https://disconnect.me/

A *browser fingerprint* is a set of information related to a user's device from the hardware to the operating system to the browser and its configuration. *Browser fingerprinting* refers to the process of collecting information through a web browser to build a fingerprint of a device. Via a script running inside a browser, a server can collect a wide variety of information from public interfaces called Application Programming Interface (API) and HTTP headers. An API is an interface that provides an entry point to specific objects and functions. While some APIs require a permission to be accessed like the microphone or the camera, most of them are freely accessible from any JavaScript script rendering the information collection trivial. Contrarily to other identification techniques like cookies that rely on a unique identifier (ID) directly stored inside the browser, browser fingerprinting is qualified as completely *stateless*. It does not leave any trace as it does not require the storage of information inside the browser.

The goal of this work is twofold: first, to provide an accessible entry point for newcomers by systematizing existing work, and second, to form the foundations for future research in the domain by eliciting the current challenges yet to overcome. We accomplish these goals with the following contributions:

- A thorough survey of the research conducted in the domain of browser fingerprinting with a summary of the framework used to evaluate the uniqueness of browser fingerprints and their adoption on the web.
- An overview of how this technique is currently used in both research and industry.
- A taxonomy that classifies existing defense mechanisms into different categories, providing a high-level view of the benefits and drawbacks of each of these techniques.
- A discussion about the current state of browser fingerprinting and the challenges it is currently facing on the science, technological, business, and legislative aspects.

This work has been submitted for publication at an international journal.

6.5. Security Analysis of GDPR Subject Access Request Procedures

With the GDPR in place since May 2018, the rights of the European users have been strengthened. The GDPR defines users' rights and aims at protecting their personal data. Every European Data Protection Authority (DPA) provides advices, explanations and recommendations on the use of these rights. However, the GDPR does not provide any prescriptive requirements on how to authenticate a data subject request. This lack of concrete description undermines the practical effect of the GDPR: it hampers the way to exercise the subject access right, to check the lawfulness of the processing and to enforce the derived legal rights therefrom (erasure, rectification, restriction, etc).

Every data subject would like to benefit from the rights specified in GDPR, but still wonders: *How do I exercise my access right?How do I prove my identity to the controller?* These questions are critical to build trust between the data subject and the controller. The data subject is concerned with threats like *impersonation* and *abusive identity check*. Impersonation is the case of a malicious party who attempts to abuse the subject access request (SAR) by impersonating a subject to a controller. Abusive identity check occurs when a data controller is too curious and verifies the identity of a subject by asking irrelevant and unnecessary information like an electricity bill or government issued documents.

Symmetrically, every data controller needs to know how to proceed when they receive an access request: *Is the request legitimate? What is necessary to identify the subject's data?* These concerns aggravate when controllers deal with indirectly-linked identifiers, such as IP addresses, or when they have no prior contact with data subjects, as in *Google Spain* [0]. Most of all, data controllers want to avoid data breaches, as it can result in legal proceedings and heavy fines. Such consequence occurs in two cases: *(i)* the data controller releases data to an illegitimate subject, or *(ii)* he releases data of a subject A to a legitimate subject B.

[0] Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, Case C-131/12, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62012CJ0131&from=EN

All these questions concern the authentication procedure between the data subject and the controller. They both share a common interest in holding a strong authentication procedure to prevent impersonation and data breaches. The subject must be careful during the authentication procedure, as for providing too much personal information could compromise her right of privacy. Additionally, the controller needs to ask the appropriate information to identify the subject's data without ambiguity. There is clearly a tension during this authentication act between the controller, who tries to get as much information as possible, and the data subject who wants to provide as little as possible. Plausibly, subject access rights can probably increase the incidence of personal records being accidentally or deliberately opened to unauthorised third parties [22].

This work studies *the tension during the authentication between the data subject and the data controller*. We first evaluate the threats to the SAR authentication procedure and then we analyze the recommendations of 28 DPAs of European Union countries. We observe that four of them can potentially lead to abusive identity check. On the positive side, six of them are recommending to enforce the data minimization principle during authentication. This principle, on one hand, protects the right to privacy of data subjects, and on the other hand prevents data controllers to massively collect personal data that is not needed for authentication, thus preventing abusive identity check.

We have then evaluated the authentication procedure when exercising the access right of the 50 most popular websites and 30 third-party tracking services. Several popular websites require to systematically provide a national identity card or government-issued documents to authenticate the data subject. Among third-party tracking services, 9 of them additionally to cookies demand other personal data from the data subjects, like the identity card or the full name. We explain that such demands are not justified because additional information can not prove the ownership of the cookie.

We then provide guidelines to Data Protection Authorities, website owners and third party services on how to authenticate data subjects safely while protecting their identities, and without requesting additional unnecessary information (complying with the data minimization principle). More precisely, we explain how data controllers and data subjects must interact and how digital identifiers can be redesigned to be compliant with the GDPR.

This work has been published at the Annual Privacy Forum (APF) 2019 [13].

6.6. Measuring Legal Compliance of Cookie Banners

6.6.1. Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners

In this work, we analyze the legal requirements on how cookie banners are supposed to be implemented to be fully compliant with the ePrivacy Directive and the GDPR.

Our contribution resides in the definition of 17 operational and fine-grained requirements on cookie banner design that are legally compliant, and moreover, we define whether and when the verification of compliance of each requirement is technically feasible.

The definition of requirements emerges from a joint interdisciplinary analysis composed of lawyers and computer scientists in the domain of web tracking technologies. As such, while some requirements are provided by explicitly codified legal sources, others result from the domain-expertise of computer scientists. In our work, we match each requirement against existing cookie banners design of websites. For each requirement, we exemplify with compliant and non-compliant cookie banners.

As an outcome of a technical assessment, we verify per requirement if technical (with computer science tools) or manual (with any human operator) verification is needed to assess compliance of consent and we also show which requirements are impossible to verify with certainty in the current architecture of the Web. For example, we explain how the GDPR's requirement for revocable consent could be implemented in practice: when consent is revoked, the publisher should delete the consent cookie and communicate the withdrawal to all third parties who have previously received consent.

With this approach we aim to support practically-minded parties (compliance officers, regulators, privacy NGOs, researchers, and computer scientists) to assess compliance and detect violations in cookie banners' design and implementation, specially under the current revision of the EU ePrivacy framework.

This working paper is submitted for publication.

6.6.2. Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework

As a result of the GDPR and the ePrivacy Directive, (known as "cookie law"), European users encounter cookie banners on almost every website. Many of such banners are implemented by Consent Management Providers (CMPs), who respect the IAB Europe's Transparency and Consent Framework (TCF). Via cookie banners, CMPs collect and disseminate user consent to third parties. In this work, we systematically study IAB Europe's TCF and analyze consent stored behind the user interface of TCF cookie banners. We analyze the GDPR and the ePrivacy Directive to identify legal violations in implementations of cookie banners based on the storage of consent and detect such violations by crawling 22 949 European websites.

With two automatic and semi-automatic crawl campaigns, we detect violations, and we find that: 175 websites register positive consent even if the user has not made their choice; 236 websites nudge the users towards accepting consent by pre-selecting options; and 39 websites store a positive consent even if the user has explicitly opted out. Performing extensive tests on 560 websites, we find at least one violation in 54% of them.

Finally, we provide a browser extension called "Cookie Glasses" to facilitate manual detection of violations for regular users and Data Protection Authorities.

This working paper is submitted for publication at an international conference.

6.7. Session Types

Session types describe communication protocols between two or more parties by specifying the sequence of exchanged messages and their functionality (sender, receiver and type of carried data). They may be viewed as the analogue, for concurrency and distribution, of data types for sequential computation. Originally conceived as a static analysis technique for an enhanced version of the π-calculus, session types have been subsequently embedded into a range of functional, concurrent, and object-oriented programming languages.

While binary sessions can be described by a single session type, multiparty sessions require two kinds of types: a *global type* that describes the whole session protocol, and *local types* that describe the contributions of the various participants to the protocol. The key requirement to achieve safety properties such as the absence of communication errors and deadlock-freedom, is that the local types of the processes implementing the participants be obtained as projections from the same global type (the one describing the session protocol).

We have pursued our work on multiparty session types along four main directions, in collaboration with colleagues from the Universities of Groningen, Luxemburg, Nice Sophia Antipolis, Turin and Eastern Piedmont. One of these directions is described in Section 6.8.3 , the others are described below.

6.7.1. Reversible Sessions with Flexible Choices

Reversibility has been an active trend of research for the last fifteen years. A reversible computation is a computation that may roll back to a past state. Allowing computations to reverse is a means to improve system flexibility and reliability. In the setting of concurrent process calculi, reversible computations have been first studied for Milner's calculus CCS, then for the π-calculus, and only recently for typed session calculi.

Following up on our previous work on concurrent reversible sessions, we studied a simpler but somewhat more realistic calculus for concurrent reversible multiparty sessions, equipped with a flexible choice operator allowing for different sets of participants in each branch of a choice. This operator was inspired by the notion of *connecting communication* introduced by other authors to describe protocols with optional participants. Our calculus supports a compact representation of the *history* of processes and types, which facilitates the definition of rollback. Moreover, it implements a fine-tuned strategy for backward computation, where only some specific participants, the "choice leaders", can trigger a rollback. We present a session type system for this calculus and show that it enforces the expected properties of session fidelity, forward progress and backward progress. This work has been published in the journal [11].

6.7.2. Multiparty Sessions with Internal Delegation

We have investigated a new form of *delegation* for multiparty session calculi. Usually, the delegation mechanism allows a session participant to appoint a participant in another session to act on her behalf. This means that delegation is inherently an inter-session mechanism, which requires session interleaving. Hence delegation falls outside the descriptive power of global types, which specify single multiparty sessions. As a consequence, properties such as deadlock-freedom or lock-freedom are difficult to ensure in the presence of delegation. In our work, we adopt a different view of delegation, by allowing participants to delegate tasks to each other within the same multiparty session. This way, delegation occurs within a single session (whence the name “internal delegation”) and may be captured by its global type. To increase flexibility in the use of delegation, we use again connecting communications, in order to accommodate optional participants in the branches of choices. By this means, we are also able to express conditional delegation. We present a session type system based on global types with internal delegation, and show that it ensures the usual safety properties of multiparty sessions, together with a progress property.

This work has been published in a special issue of TCS dedicated to Maurice Nivat [12].

6.7.3. Event Structure Semantics for Multiparty Sessions

In the work [14] we investigate the relationship between multiparty session calculi and other concurrency models, by focussing on Event Structures as proposed in the late 80’s. We consider a standard multiparty session calculus where sessions are described as networks of sequential processes, and each process implements a participant in the session. We propose an interpretation of such networks as *Flow Event Structures* (FESs) (a subclass of Winskel’s Stable Event Structures), which allows concurrency between session communications to be explicitly represented. We then introduce global types for these networks, and define an interpretation of global types as *Prime Event Structures* (PESs). Since the syntax of global types does not allow all the concurrency among communications to be expressed, the events of the associated PES need to be defined as equivalence classes of communication sequences up to *permutation equivalence*. We show that when a network is typable by a global type, the FES semantics of the former is equivalent, in a precise technical sense, to the PES semantics of the latter.

This work has been published in a volume dedicated to Rocco De Nicola on the occasion of his 65th birthday [14]. An extended version is available as Research Report [21].

6.8. Web Reactive Programming

6.8.1. HipHop.js

This year, we have completed the design of the *HipHop* programming language. We have finalized the syntax of core instructions, stabilized the interfacing with JavaScript, added variables that supplement signals in local computation, and we have completed the synchronous/asynchronous connections. A paper describing this final version of the paper is currently under submission.

We have also improved significantly the *HipHop* implementation for speed and for debugging.

- Leveraging on the *Hop* speed improvement and by adding a new *HipHop* compilation stage we have been able to accelerate by a factor of about 10× the intrinsic execution time of the reactive machine. The optimization removes nets of the virtual electronic circuits that are generated by the *HipHop* compiler by propagating constant and by collapsing identical nodes. This contributions is included in the main development tree (https://github.com/manuel-serrano/hiphop).
- A central difficulty of the synchronous reactive programming is debugging and error messages. The *HipHop* compilation roughly consists in implementing efficiently and compactly a deterministic automata that represents the user source code. If a causality error is detected during that compilation, unless a precise isolation of the user source code fragments that are involved in that error, the error

message reported to the user is so imprecise that fixing the problem is difficult. We have implemented an algorithm based on *strongly connected components* that enables the needed isolation. This experimental feature is currently publicly available via a dedicating development branch under the *HipHop* github repository.

6.8.2. Interactive music composition

The production of a piece of music by school children using the Skini platform as part of SACEM's call for projects "Fabrique à musique" (Music Factory) was initiated in 2019. It ended in 2019 with the realization of a show at the Nice Conservatory in May 2019. The music piece thus created implemented all of Skini's functionalities, from the distributed sequencer that allowed the pupils to design the basic material, to the control of the live orchestration, not by the audience in this case, but by the 24 students who participated in the project. Following the success of this first experiment, the project is being extended for 2020/2021 by Inria, with another class, as part of the "les cordées de la réussite" program.

Beyond the improvement of the system, and in particular of the distributed sequencer, thanks to the significant performance improvements of HipHop.js, it has been possible to enrich the controls on musical orchestrations by driving transformation elements of Skini's basic elements (the patterns) such as transpositions, use of patterns of different durations, music mode conversions, or tempo control. The coupling of these new processes has enriched the range of possibilities for interaction and has opened up new horizons in the field of pattern-based generative music.

In terms of musical creation, we have been able to implement orchestrations using the platform's new technical possibilities in order to reduce musical processes perceived as too automatic. We can note as convincing results: the possibility to break the too big symmetries on the durations of the patterns, and the variations of tempi subjected to various controls. We were able to demonstrate that with the same HipHop.js music orchestration program, we could efficiently generate very different musical pieces.

Skini music composition has been described in a conference paper presented in the NIME 2019 conference [16].

6.8.3. Multiparty Reactive Sessions

Ensuring that communication-centric systems interact according to an intended protocol is a challenging problem, particularly for systems with some reactive or timed components. To rise to this challenge, we have studied the integration of Session-based Concurrency and Synchronous Reactive Programming (SRP).

Synchronous Reactive Programming (SRP) is a well-established programming paradigm whose essential features are logical instants, broadcast events and event-based preemption. This makes it an ideal vehicle for the specification and analysis of timed reactive systems. *Session-based Concurrency* is the model of concurrent computation induced by session types.

In the Research Report [20], we propose a multiparty session calculus enriched with features from SRP. In this calculus, protocol participants may broadcast messages, suspend themselves while waiting for a message, and react to events. Our main contribution is a session type system for this calculus, which enforces session correctness for non-interleaved sessions and additionally ensures *input timeliness*, a time-related property that entails livelock-freedom (while deadlock-freedom holds by construction in our calculus). Our type system departs significantly from existing ones, specifically as it captures the notion of “logical instant” typical of SRP.

RMOD Project-Team

7. New Results

7.1. Dynamic Languages: Virtual Machines

Illicium A modular transpilation toolchain from Pharo to C. The Pharo programming language runs on the OpenSmalltalk-VM. This Virtual Machine (VM) is mainly written in Slang, a subset of the Smalltalk language dedicated to VM development. Slang is transpiled to C using the Slang-to-C transpiler. The generated C is then compiled to produce the VM executable binary code. Slang is a powerful dialect for generating C because it benefits from the tools of the Smalltalk environment, including a simulator that runs and debugs the VM. However, the Slang-to-C transpiler is often too permissive. For example, the Slang-to-C transpiler generates invalid C code from some Smalltalk concepts it does not support. This makes the Slang code hard to debug as the errors are caught very late during the development process, which is worsen by the loss of the mapping between the generated C code and Slang. The Slang-to-C transpiler is also hard to extend or adapt to modify part of the translation process. We present Illicium, a new modular transpilation toolchain based on a subset of Pharo targeting C through AST transformations. This toolchain translates the Pharo AST into a C AST to generate C code. Using ASTs as source and target artifacts enables analysis, modification and validation at different levels during the translation process. The main translator is split into smaller and replaceable translators to increase modularity. Illicium also allows the possibility to introduce new translators and to chain them together, increasing reusability. To evaluate our approach, we show with a use case how to extend the transpilation process with a translation that requires changes not considered in the original C AST. [7]

GildaVM: a Non-Blocking I/O Architecture for the Cog VM. The OpenSmalltalk virtual machine (VM) was historically designed as a single-threaded VM. All VM code including the Smalltalk interpreter, the garbage collector and the just-in-time compiler run in the same single native thread. While this VM provides concurrency through green threads, it cannot take advantage of multi-core processors. This architecture performs really well in practice until the VM accesses external resources such as e.g., FFI callouts, which block the single VM thread and prevent green threads to benefit from the processor. We present GildaVM, a multi-threaded VM architecture where one thread at a time executes the VM while allowing non-blocking I/O in parallel. The ownership of the VM is orchestrated by a Global Interpreter Lock (GIL) as in the standard implementations of Python and Ruby. However, within a single VM thread concurrency is still possible through green threads. We present a prototype implementation of this architecture running on top of the Stack flavour of the OpenSmalltalk VM. We finally evaluate several aspects of this architecture like FFI and thread-switch overhead. While current benchmarks show good results for long FFI calls, short FFI calls require more research to minimize the overhead of thread-switch. [9]

7.2. Dynamic Languages: Language Constructs for Modular Design

Magic Literals in Pharo Literals are constant values (numbers, strings, etc.) used in the source code. Magic literals are the ones used without a clear explanation of their meaning. Presence of such literals harms source code readability, decreases its modularity, and encourages code duplication. Identifying magic literals is not straightforward. A literal can be considered self-explanatory in one context and magic in another. We need a heuristic to help developers spot magic literals. We study and characterize the literals in Pharo. We implemented a heuristic to detect magic literals and integrated it as a code critic rule for System Browser and Critics Browser in Pharo 7. We run our heuristic on 112,500 Pharo methods which reported 23,292 magic literals spread across 8,986 methods. We manually validated our approach on a random subset of 100 methods and found that 62% of the reported literals in those methods are indeed magic. [3]

Towards easy program migration using language virtualization Migrating programs between language versions is a daunting task. A developer writes a program in a particular version of a language and cannot foresee future language changes. In this article, we explore a solution to gradual program migration based on virtualization at the programming language level. Our language virtualization approach adds a backwards-compatibility layer on top of a recent language version, allowing developers to load and run old programs on the more recent infrastructure. Developers are then able to migrate the program to the new language version or are able to run it as it is. Our virtualization technique is based on a dynamic module implementation and code intercession techniques. Migrated and non-migrated parts co-exist in the meantime allowing an incremental migration procedure. We validate it by migrating legacy Pharo programs, MuTalk and Fuel. [10]

7.3. Dynamic Languages: Debugging

Sindarin: A Versatile Scripting API for the Pharo Debugger Debugging is one of the most important and time consuming activities in software maintenance, yet mainstream debuggers are not well-adapted to several debugging scenarios. This has led to the research of new techniques covering specific families of complex bugs. Notably, recent research proposes to empower developers with scripting DSLs, plugin-based and moldable debuggers. However, these solutions are tailored to specific use-cases, or too costly for one-time-use scenarios. We argue that exposing a debugging scripting interface in mainstream debuggers helps in solving many challenging debugging scenarios. For this purpose, we present Sindarin, a scripting API that eases the expression and automation of different strategies developers pursue during their debugging sessions. Sindarin provides a GDB-like API, augmented with AST-bytecode-source code mappings and object-centric capabilities. To demonstrate the versatility of Sindarin, we reproduce several advanced breakpoints and non-trivial debugging mechanisms from the literature. [4]

Challenges in Debugging Bootstraps of Reflective Kernels The current explosion of embedded systems (i.e., IoT, Edge Computing) implies the need for generating tailored and customized software for these systems. Instead of using specific runtimes (e.g., MicroPython, eLua, mRuby), we advocate that bootstrapping specific language kernels is a promising higher-level approach because the process takes advantage of the generated language abstractions, easing the task for a language developer. Nevertheless, bootstrapping language kernels is still challenging because current debugging tools are not suitable for fixing the possible failures that occur during the process. We take the Pharo bootstrap process as an example to analyse the different challenges a language developer faces. We propose a taxonomy of failures appearing during bootstrap and their causes. Based on this analysis, we identify future research directions: (1) prevention measures based on the reification of implicit virtual machine contracts, and (2) hybrid debugging tools that unify the debugging of high-level code from the bootstrapped language with low-level code from the virtual machine. [6]

7.4. Software Reengineering

Decomposing God Classes at Siemens A group of developers at Siemens Digital Industry Division approached our team to help them restructure a large legacy system. Several problems were identified, including the presence of God classes (big classes with thousands of lines of code and hundred of methods). They had tried different approaches considering the dependencies between the classes, but none were satisfactory. Through interaction during the last three years with a lead software architect of the project, we designed a software visualization tool and an accompanying process that allows her to propose a decomposition of a God Class in a matter of one or two hours even without prior knowledge of the class (although actually implementing the decomposition in the source code could take a week of work). We present the process that was formalized to decompose God Classes and the tool that was designed. We give details on the system itself and some of the classes that were decomposed. The presented process and visualisations have been successfully used for the last three years on a real industrial system at Siemens. [1]

Rotten Green Tests Unit tests are a tenant of agile programming methodologies, and are widely used to improve code quality and prevent code regression. A green (passing) test is usually taken as a robust sign that the code under test is valid. However, some green tests contain assertions that are never executed. We call such tests Rotten Green Tests. Rotten Green Tests represent a case worse than a broken test: they report that the code under test is valid, but in fact do not test that validity. We describe an approach to identify rotten green tests by combining simple static and dynamic call-site analyses. Our approach takes into account test helper methods, inherited helpers, and trait compositions, and has been implemented in a tool called DrTest. DrTest reports no false negatives, yet it still reports some false positives due to conditional use or multiple test contexts. Using DrTest we conducted an empirical evaluation of 19,905 real test cases in mature projects of the Pharo ecosystem. The results of the evaluation show that the tool is effective; it detected 294 tests as rotten-green tests that contain assertions that are not executed. Some rotten tests have been "sleeping" in Pharo for at least 5 years. [2]

Migrating GWT to Angular 6 using MDE In the context of a collaboration with Berger-Levrault, a major IT company, we are working on the migration of a GWT application to Angular. We focus on the GUI aspect of this migration which, even if both are web frameworks, is made difficult because they use different programming languages (Java for one, Typescript for the other) and different organization schemas (e.g. different XML files). Moreover, the new application must mimic closely the visual aspect of the old one so that the users of the application are not disturbed. We propose an approach in three steps that uses a meta-model to represent the GUI at a high abstraction level. We evaluated this approach on an application comprising 470 Java (GWT) classes representing 56 screens. We are able to model all the web pages of the application and 93% of the wid-gets they contain, and we successfully migrated (i.e., the result is visually equal to the original) 26 out of 39 pages (66%). We give examples of the migrated pages, both successful and not. [14] [11] [12]

Empirical Study of Programming to an Interface A popular recommendation to programmers in object-oriented software is to *program to an interface, not an implementation* (PTI). Expected benefits include increased simplicity from abstraction, decreased dependency on implementations, and higher flexibility. Yet, interfaces must be immutable, excessive class hierarchies can be a form of complexity, and *speculative generality* is a known code smell. To advance the empirical knowledge of PTI, we conducted an empirical investigation that involves 126 Java projects on GitHub, aiming to measuring the decreased dependency benefits (in terms of cochange). [13]

Exposing Test Analysis Results with DrTests Tests are getting the cornerstone of continuous development process and software evolution. Tests are the new gold. To improve test quality, a plethora of analyses is proposed such as test smells, mutation testing, test coverage. The problem is that each analysis often needs a particular way to expose its results to the developer. There is a need for an architecture supporting test running and analysis in a modular and extensible way. We present an extensible plugin-based architecture to run and report test results. DrTests is a new test browser that implements such plugin-based architecture. DrTests supports the execution of rotten tests, comments to tests, coverage and profiling tests. [5]

7.5. Blockchain Software Engineering

SmartAnvil: Open-Source Tool Suite for Smart Contract Analysis Smart contracts are new computational units with special properties: they act as classes with aspectual concerns; their memory structure is more complex than mere objects; they are obscure in the sense that once deployed it is difficult to access their internal state; they reside in an append-only chain. There is a need to support the building of new generation tools to help developers. Such support should tackle several important aspects: (1) the static structure of the contract, (2) the object nature of published contracts, and (3) the overall data chain composed of blocks and transactions. We present SmartAnvil an open platform to build software analysis tools around smart contracts. We illustrate the general components and we focus on three important aspects: support for static analysis of Solidity smart contracts, deployed smart contract binary analysis through inspection, and blockchain navigation and querying. SmartAnvil is open-source and supports a bridge to the Moose data and software analysis platform. [18]

The Influence Factors on Ethereum Transaction Fees In Ethereum blockchain, the user needs to set a Gas price to get a transaction processed and approved by Miners. To have the transaction executed, the Gas price has to be greater than or equal to the lowest Ethereum transaction fees. We present a set of data sampled every 15 seconds, from December 1st, 2018 to December 15, 2018, coming from different blockchain web APIs. The aim is to investigate whether and to what extent different variables - such as the number of pending transactions, the value of the USD/Ether pair, average electricity prices around the world, and the number of miners - influence the Ethereum transaction fees. This study is relevant from an economic perspective because more and more companies in different economic fields are adopting Ethereum blockchain. From historical data analysis, we found that only some of these variables do have an influence. For example, the number of pending transactions and the number of miners have a major influence on Ethereum transaction fees when compared to the other variables. [8]

AGORA Project-Team

7. New Results

7.1. Wireless network deployment

Participants: Walid Bechkit, Ahmed Boubrima, Oana Iova, Rodrigue D. Komguem, Abdoul-Aziz Mbacke, Jad Oueis, Hervé Rivano, Razvan Stanica, Fabrice Valois

7.1.1. Deployment of wireless sensor networks for air quality mapping

Wireless sensor networks (WSN) are widely used in environmental applications where the aim is to sense a physical phenomenon such as temperature, air pollution, etc. A careful deployment of sensors is necessary in order to get a better knowledge of these physical phenomena while ensuring the minimum deployment cost [18]. In this work, we focus on using WSN for air pollution mapping and tackle the optimization problem of sensor deployment [3]. Unlike most of the existing deployment approaches, which are either generic or assume that sensors have a given detection range, we define an appropriate coverage formulation based on an interpolation formula that is adapted to the characteristics of air pollution sensing. We derive from this formulation two deployment models for air pollution mapping using integer linear programming while ensuring the connectivity of the network and taking into account the sensing error of nodes. We analyze the theoretical complexity of our models and propose heuristic algorithms based on linear programming relaxation and binary search. We perform extensive simulations on a dataset of the Lyon city, France in order to assess the computational complexity of our proposal and evaluate the impact of the deployment requirements on the obtained results.

7.1.2. Characterization of radio links in case of a ground deployment

In this work, we are interested in characterizing the link properties of a wireless sensor network with nodes deployed at ground level [5]. Such a deployment is fairly common in practice, e.g., when monitoring the vehicular traffic on a road segment or the status of infrastructures such as bridges, tunnels or dams. However, the behavior of off-the-shelf wireless sensor nodes in these settings is not yet completely understood. Through a thorough experimentation campaign, we evaluated not only the impact of the ground proximity on the wireless links, but also the impact of some parameters such as the packet payload, the communication channel frequency and the topography of the deployment area. Our results show that a ground-level deployment has a significant negative impact on the link quality, while parameters such as the packet size produce unexpected consequences. This allows us to parameterize classical theoretical models in order to fit a ground-level deployment scenario. Finally, based on the lessons learned in our field tests, we discuss some considerations that must be taken into account during the design of communication protocols and before the sensor deployment in order to improve network performance.

7.1.3. Sensor deployment in linear wireless sensor networks using the concept of virtual node

In a multi-hop wireless sensor network with a convergecast communication model, there is a high traffic accumulation in the neighborhood of the sink. This area constitutes the bottleneck of the network since the sensors deployed within it rapidly exhaust their batteries. In this work, we consider the problem of sensors deployment for lifetime maximization in a linear wireless sensor network [6]. Existing approaches express the deployment recommendations in terms of distance between consecutive sensors. Solutions imposing such constraints on the deployment may be costly and difficult to manage. We propose a new approach where the network is formed of virtual nodes, each associated to a certain geographical area. An analytical model of the network traffic per virtual node is proposed and a greedy algorithm to calculate the number of sensors that should form each virtual node is presented. Performance evaluation shows that the greedy deployment can improve the network lifetime by up to 40%, when compared to the uniform deployment. Moreover, the proposed approach outperforms the related work when complemented by a scheduling algorithm which reduces the messages overhearing. It is also shown that the lifetime of the network can be significantly improved if the battery capacity of each sensor is dimensioned taking into account the traffic it generates or relays.

7.1.4. Core network function placement in self-deployable mobile networks

Emerging mobile network architectures (e.g., aerial networks, disaster relief networks) are disrupting the classical careful planning and deployment of mobile networks by requiring specific self-deployment strategies. Such networks, referred to as self-deployable, are formed by interconnected rapidly deployable base stations that have no dedicated backhaul connection towards a traditional core network. Instead, an entity providing essential core network functionalities is co-located with one of the base stations. In this work, we tackle the problem of placing this core network entity within a self-deployable mobile network, i.e., we determine with which of the base stations it must be co-located [9], [15] [15]. We propose a novel centrality metric, the flow centrality, which measures a node capacity of receiving the total amount of flows in the network. We show that in order to maximize the amount of exchanged traffic between the base stations and the core network entity, under certain capacity and load distribution constraints, the latter should be co-located with the base station having the maximum flow centrality. We first compare our proposed metric to other state of the art centralities. Then, we highlight the significant traffic loss occurring when the core network entity is not placed on the node with the maximum flow centrality, which could reach 55% in some cases.

7.1.5. Cyber physical systems and Internet of things: emerging paradigms on smart cities

A city is smart when investment in traditional and modern infrastructure, human and social capital, fuel well being, high quality of life, and sustainable economic development. The Smart City paradigm is driven by technological evolution in the field of Information and Communication Technologies, and more specifically the paradigms of Internet of Things, Industrial Internet of Things and their confluence with Cyber Physical Systems [12]. Smart Cities present a number of application domains that are related to their critical infrastructures, including energy and transport. These domains present needs similar to the industrial manufacturing environment utilizing smart devices and employing control automation for their applications. They could thus be labeled as *industrial domains* in the wider sense. This work presents three application domains associated with Smart Cities, namely Smart Lighting, Smart Buildings / Energy, and Smart Urban Mobility, identifies their requirements and challenges and reviews existing solutions.

7.2. Wireless data collection

Participants: Oana Iova, Abderrahman Ben Khalifa, Razvan Stanica

7.2.1. Reliable and efficient support for downward traffic in RPL

Modern protocols for wireless sensor networks efficiently support multi-hop upward traffic from many sensors to a collection point, a key functionality enabling monitoring applications. However, the ever-evolving scenarios involving low-power wireless devices increasingly require support also for downward traffic, e.g., enabling a controller to issue actuation commands based on the monitored data. The IETF Routing Protocol for Low-power and Lossy Networks (RPL) is among the few tackling both traffic patterns. Unfortunately, its support for downward traffic is significantly unreliable and inefficient compared to its upward counterpart. We tackle this problem by extending RPL with mechanisms inspired by opposed, yet complementary, principles [7]. At one extreme, we retain the route-based operation of RPL and devise techniques allowed by the standard but commonly neglected by popular implementations. At the other extreme, we rely on flooding as the main networking primitive. Inspired by these principles, we define three base mechanisms, integrate them in a popular RPL implementation, analyze their individual and combined performance, and elicit the resulting tradeoffs in scalability, reliability, and energy consumption. The evaluation relies on simulation, using both real-world topologies from a smart city scenario and synthetic grid ones, as well as on testbed experiments validating our findings from simulation. Results show that the combination of all three mechanisms into a novel protocol, T-RPL *i)* yields high reliability, close to the one of flooding, *ii)* with a low energy consumption, similar to route-based approaches, and *iii)* improves remarkably the scalability of RPL w.r.t. downward traffic.

7.2.2. Performance evaluation of LED-to-camera communications

The use of LED-to-camera communication opens the door to a wide range of use cases and applications, with diverse requirements in terms of quality of service. However, while analytical models and simulation tools exist for all the major radio communication technologies, the only way of currently evaluating the performance of a network mechanism over LED-to-camera is to implement and test it. Our work aims to fill this gap by proposing a Markov-modulated Bernoulli process to model the wireless channel in LED-to-camera communications, which is shown to closely match experimental results [11]. Based on this model, we develop and validate *CamComSim*, the first network simulator for LED-to-camera communications.

7.2.3. Performance evaluation of channel access methods for dedicated IoT networks

Networking technologies dedicated for the Internet of Things are different from the classical mobile networks in terms of architecture and applications. This new type of network is facing several challenges to satisfy specific user requirements. Sharing the communication medium between (hundreds of) thousands of connected nodes and one base station is one of these main requirements, hence the necessity to imagine new solutions, or to adapt existing ones, for medium access control. In this work, we start by comparing two classical medium access control protocols, CSMA/CA and Aloha, in the context of Internet of Things dedicated networks [13]. We continue by evaluating a specific adaptation of Aloha, already used in low-power wide area networks, where no acknowledgement messages are transmitted in the network. Finally, we apply the same concept to CSMA/CA, showing that this can bring a number of benefits. The results we obtain after a thorough simulation study show that the choice of the best protocol depends on many parameters (number of connected objects, traffic arrival rate, allowed retransmission number), as well as on the metric of interest (e.g. packet reception probability or energy consumption).

7.2.4. On the use of wide channels in WiFi networks

An increased density of access points is common today in WiFi deployments, and more and more parameters need to be configured in such networks. In this work, we question current industrial guidelines for both residential and enterprise scenarios [14]. More precisely, we investigate the joint channel, power, and carrier sense threshold allocation problem in IEEE 802.11ac networks, showing that the current practice, which is to use narrower channels at maximum power when the deployment is dense, yields much worse performance than a solution using the widest possible channel with a much lower power.

7.3. Network data exploitation

Participants: Florent Delaine, Panagiota Katsikouli, Hervé Rivano, Razvan Stanica

7.3.1. Calibration algorithms for environmental sensor networks

The recent developments in both nanotechnologies and wireless technologies have enabled the rise of small, low cost and energy efficient environmental sensing devices. Many projects involving dense sensor networks deployments have followed, in particular within the Smart City trend. If such deployments are now within economical and technical reach, their maintenance and reliability remain however a challenge. In particular, reaching, then maintaining, the targeted quality of measurement throughout deployment duration is an important issue. Indeed, factory calibration is too expensive for systematic application to low-cost sensors and as these sensors are usually prone to drifting because of premature aging. In addition, there are concerns about the applicability of factory calibration to field conditions [4]. These challenges have fostered many researches on in situ calibration. In situ means that the sensors are calibrated without removing them from their deployment location, preferably without physical intervention, often leveraging their communication capabilities. It is a critical challenge for the economical sustainability of networks with large scale deployments. In this work, we focus on in situ calibration methods for environmental sensor networks. We propose a taxonomy of the methodologies in the literature. Our classification relies on both the architecture of the network of sensors and the algorithmic principles of the calibration methods. This review allows us to identify and discuss two main challenges: how to improve the performance evaluation of such methods and how to enable a quantified comparison of these strategies?

7.3.2. Characterizing and Removing Oscillations in Mobile Phone Location Data

Human mobility analysis is a multidisciplinary research subject that has attracted a growing interest over the last decade. A substantial amount of such recent studies is driven by the availability of original sources of real-world information about individual movement patterns. An important task in the analysis of mobility data is reliably distinguishing between the stop locations and movement phases that compose the trajectories of the monitored subjects. The problem is especially challenging when mobility is inferred from mobile phone location data: here, oscillations in the association of mobile devices to base stations lead to apparent user mobility even in absence of actual movement [10]. In this work, we leverage a unique dataset of spatiotemporal individual trajectories that allows capturing both the user and network operator perspectives in mobile phone location data, and investigate the oscillation phenomenon. We present probabilistic and machine learning approaches for detecting oscillations in mobile phone location data, and a filtering technique for removing those. Our analyses and comparison with state-of-the-art approaches demonstrate the superiority of our solution, both in terms of removed oscillations and of error with respect to ground-truth trajectories.

COATI Project-Team

7. New Results

7.1. Network Design and Management

Participants: Julien Bensmail, Jean-Claude Bermond, Christelle Caillouet, David Coudert, Frédéric Giroire, Frédéric Havet, Nicolas Nisse, Stéphane Pérennes, Joanna Moulierac, Foivos Fioravantes, Adrien Gausseran, Andrea Tomassilli.

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

We mainly focus on the following topics: Firstly, we study Software Defined Networking (SDN) and Network Function Virtualization (NFV) and how to exploit their potential benefits. We propose algorithms for the Provisioning Service Function Chains (SFC) and algorithms to reconfigure the SFC in order to improve the network operational costs without any interruption (with a *make-before-break approach*) and Virtual Network Functions (VNF) placement algorithms to address the mono- and multi-tenant issues in edge and core networks. We also propose algorithms for distributed Mininet [0] in order to improve the performance, and also bandwidth-optimal failure recovery scheme for robust programmable networks. Secondly, we study optimization problems within optical networks: wavelength reconfiguration for seamless migration and spectrum assignment in elastic optical tree-networks. Thirdly, we study the scheduling of network tasks within a data center while taking into account the communication between the network resources. We also study distributed link scheduling in wireless networks. Finally, we investigate on the placement of drones for maximizing the coverage of a landscape by drones in order to localize targets or collect data from sensors.

7.1.1. Software Defined Networks and Network Function Virtualization

Recent advances in networks such as Software Defined Networking (SDN) and Network Function Virtualization (NFV) are changing the way network operators deploy and manage Internet services. On the one hand, SDN introduces a logically centralized controller with a global view of the network state. On the other hand, NFV enables the complete decoupling of network functions from proprietary appliances and runs them as software applications on general purpose servers. In such a way, network operators can dynamically deploy Virtual Network Functions (VNFs). SDN and NFV, both separately, bring to network operators new opportunities for reducing costs, enhancing network flexibility and scalability, and shortening the time-to-market of new applications and services. Moreover, the centralized routing model of SDN jointly with the possibility of instantiating VNFs on demand may open the way for an even more efficient operation and resource management of networks. For instance, an SDN/NFV-enabled network may simplify the Service Function Chain (SFC) deployment and provisioning by making the process easier and cheaper. We addressed several questions in this context.

In [15], we aim at investigating how to leverage both SDN and NFV in order to exploit their potential benefits. We took steps to address the new opportunities offered in terms of network design, network resilience, and energy savings, and the new problems that arise in this new context, such as the optimal network function placement in the network. We show that a symbiosis between SDN and NFV can improve network performance and significantly reduce the network's Capital Expenditure (CapEx) and Operational Expenditure (OpEx).

[0] Mininet provides a virtual test bed and development environment for software-defined networks (SDN). See http://mininet.org.

In [50], [57], [58], we consider the problem of reconfiguring SFC with the goal of bringing the network from a sub-optimal to an optimal operational state. We propose optimization models based on the *make-before-break* mechanism, in which a new path is set up before the old one is torn down. Our method takes into consideration the chaining requirements of the flows and scales well with the number of nodes in the network. We show that, with our approach, the network operational cost defined in terms of both bandwidth and installed network function costs can be reduced and a higher acceptance rate can be achieved, while not interrupting the flows.

In [59], we consider the placement of functions in 5G networks in which functions must not only be deployed in large central data centers, but also in the edge. We propose an algorithm that solves the Virtual Network Function Chain Placement Problem allowing a fine management of these rare resources in order to respond to the greatest number of requests possible. Because networks can be divided into several entities belonging to different tenants who are reluctant to reveal their internal topologies, we propose a heuristic that allows the NFV orchestrator to place the function chains based only on an abstract view of the infrastructure network. We leverage this approach to address the complexity of the problem in large mono- or multi-tenant networks. We analyze the efficiency of our algorithm and heuristic with respect to a wide range of parameters and topologies.

In [53], [80], [69], we rethink the network dimensioning problem with protection against Shared Risk Link Group (SLRG) failures in the SDN context. We propose a path-based protection scheme with a global rerouting strategy, in which, for each failure situation, there may be a new routing of all the demands. Our optimization task is to minimize the needed amount of bandwidth. After discussing the hardness of the problem, we develop a scalable mathematical model that we handle using the Column Generation technique. Through extensive simulations on real-world IP network topologies and on random generated instances, we show the effectiveness of our method. Finally, our implementation in OpenDaylight demonstrates the feasibility of the approach and its evaluation with Mininet shows that technical implementation choices may have a dramatic impact on the time needed to reestablish the flows after a failure takes place.

Finally, in [49], [78], [79], we consider the problem of performing large scale SDN networks simulations in a distributed environment. Indeed, networks have become complex systems that combine various concepts, techniques, and technologies. Hence, modelling or simulating them is now extremely complicated and researchers massively resort to prototyping techniques. Among other tools, Mininet is the most popular when it comes to evaluate SDN propositions. It allows to emulate SDN networks on a single computer. However, under certain circumstances experiments (e.g., resource intensive ones) may overload the host running Mininet. To tackle this issue, we propose Distrinet [49], [78], [79], a way to distribute Mininet over multiple hosts. Distrinet uses the same API than Mininet, meaning that it is compatible with Mininet programs. Distrinet is generic and can deploy experiments in Linux clusters or in the Amazon EC2 cloud. Thanks to optimization techniques, Distrinet minimizes the number of hosts required to perform an experiment given the capabilities of the hosting infrastructure, meaning that the experiment is run in a single host (as Mininet) if possible. Otherwise, it is automatically deployed on a platform using a minimum amount of resources in a Linux cluster or with a minimum cost in Amazon EC2.

7.1.2. Optimization of optical networks operation

7.1.2.1. Wavelength Defragmentation for Seamless Migration

Dynamic traffic in optical networks leads to spectrum fragmentation, which significantly reduces network performance, i.e., increases blocking rate and reduces spectrum usage. Telecom operators face the operational challenge of operating non-disruptive defragmentation, i.e., within the make-before-break paradigm when dealing with lightpath rerouting in wavelength division multiplexed (WDM) fixed-grid optical networks. In [39], we propose a make-before-break (MBB) Routing and Wavelength Assignment (RWA) defragmentation process, which provides the best possible lightpath network provisioning, i.e., with minimum bandwidth requirement. We tested extensively the models and algorithms we propose on four network topologies with different GoS (Grade of Service) defragmentation triggering events. We observe that, for a given throughput, the spectrum usage of the best make-before-break lightpath rerouting is always less than 2.5% away from that of an optimal lightpath provisioning.

7.1.2.2. On spectrum assignment in elastic optical tree-networks

To face the explosion of the Internet traffic, a new generation of optical networks is being developed; the Elastic Optical Networks (EONs). EONs use the optical spectrum efficiently and flexibly, but that gives rise to more difficulty in the resource allocation problems. In [31], we study the problem of Spectrum Assignment (SA) in Elastic Optical Tree-Networks. Given a set of traffic requests with their routing paths (unique in the case of trees) and their spectrum demand, a spectrum assignment consists in allocating to each request an interval of consecutive slots (spectrum units) such that a slot on a given link can be used by at most one request. The objective of the SA problem is to find an assignment minimizing the total number of spectrum slots to be used. We prove that SA is NP-hard in undirected stars of 3 links and in directed stars of 4 links, and show that it can be approximated within a factor of 4 in general stars. Afterwards, we use the equivalence of SA with a graph coloring problem (interval coloring) to find constant-factor approximation algorithms for SA on binary trees with special demand profiles.

7.1.3. Scheduling

7.1.3.1. When Network Matters: Data Center Scheduling with Network Tasks

We consider in [51] the placement of jobs inside a data center. Traditionally, this is done by a task orchestrator without taking into account network constraints. According to recent studies, network transfers represent up to 50% of the completion time of classical jobs. Thus, network resources must be considered when placing jobs in a data center. In this paper, we propose a new scheduling framework, introducing network tasks that need to be executed on network machines alongside traditional (CPU) tasks. The model takes into account the competition between communications for the network resources, which is not considered in the formerly proposed scheduling models with communication. Network transfers inside a data center can be easily modeled in our framework. As we show, classical algorithms do not efficiently handle a limited amount of network bandwidth. We thus propose new provably efficient algorithms with the goal of minimizing the makespan in this framework. We show their efficiency and the importance of taking into consideration network capacity through extensive simulations on workflows built from Google data center traces.

7.1.3.2. Distributed Link Scheduling in Wireless Networks

In [55], we investigate distributed transmission scheduling in wireless networks. Due to interference constraints, “neighboring links” cannot be simultaneously activated, otherwise transmissions will fail. Here, we consider any binary model of interference. We use the model described by Bui, Sanghavi, and Srikant in [85], [93]. We assume that time is slotted and during each slot there are two phases: one control phase which determines what links will be activated and a data phase in which data are sent. We assume random arrivals on each link during each slot, so that a queue is associated to each link. Since nodes do not have a global knowledge of the network, our aim (like in [85], [93]) is to design for the control phase a distributed algorithm which determines a set of non-interfering links. To be efficient the control phase should be as short as possible; this is done by exchanging control messages during a constant number of mini-slots (constant overhead). In this paper, we design the first fully distributed local algorithm with the following properties: it works for any arbitrary binary interference model; it has a constant overhead (independent of the size of the network and the values of the queues), and it does not require any knowledge of the queue-lengths. We prove that this algorithm gives a maximal set of active links, where in each interference set there is at least one active link. We also establish sufficient conditions for stability under general Markovian assumptions. Finally, the performance of our algorithm (throughput, stability) is investigated and compared via simulations to that of previously proposed schemes.

7.1.3.3. Backbone colouring and algorithms for TDMA scheduling

We investigate graph colouring models for the purpose of optimizing TDMA link scheduling in Wireless Networks. Inspired by the BPRN-colouring model recently introduced by Rocha and Sasaki, we introduce a new colouring model, namely the BMRN-colouring model, which can be used to model link scheduling problems where particular types of collisions must be avoided during the node transmissions.

In [25], we initiate the study of the BMRN-colouring model by providing several bounds on the minimum number of colours needed to BMRN-colour digraphs, as well as several complexity results establishing the hardness of finding optimal colourings. We also give a special focus on these considerations for planar digraph topologies, for which we provide refined results. Some of these results extend to the BPRN-colouring model as well. We notably prove that every planar digraph can be 8-BMRN*-coloured, while there exist planar digraphs for which 8 colours are needed in a BMRN*-colouring [72]. We also proved that the problem of deciding whether a planar digraph can be k-BMRN*-coloured is NP-hard for every $k \in \{3, ..., 6\}$.

7.1.4. Optimizing drone coverage

7.1.4.1. Self-organized UAV-based Supervision and Connectivity

The use of drones has become more widespread in recent years. Many use cases have been developed involving these autonomous vehicles, ranging from simple delivery of packages to complex emergency situations following catastrophic events. The miniaturization and very low cost of these machines make it possible today to create large meshes to ensure network coverage in disaster areas, for instance. However, the problems of scaling up and self-organization are necessary to solve problems in these use cases.

In the position paper [45], we first present different new requirements for the deployment of unmanned aerial vehicles (UAV) networks, involving the use of many drones. Then, we introduce solutions from distributed algorithms and real-time data processing to ensure quasi-optimal solutions to the raised problems.

In [44], [65], we propose VESPA, a distributed algorithm using only one-hop information of the drones, to discover targets with unknown location and auto-organize themselves to ensure connectivity between them and the sink in a multi-hop aerial wireless network. We prove that connectivity, termination and coverage are preserved during all stages of our algorithm, and we evaluate the algorithm performances through simulations. Comparison with a prior work shows the efficiency of VESPA both in terms of discovered targets and number of used drones.

7.1.4.2. Optimal placement of drones for fast sensor energy replenishment using wireless power transfer

Lifetime is the main issue of wireless sensors networks. Since the nodes are often placed in inaccessible places, the replacement of their battery is not an easy task. Moreover, the node maintenance is a costly and time consuming operation when the nodes are high in numbers. Energy harvesting technologies have recently been developed to replenish part or all of the required energy that allows a node to function. In [47], [48], we use dedicated chargers carried by drones that can fly over the network and transmit energy to the nodes using radio-frequency (RF) signals. We formulate and optimally solve the Optimal Drone Placement and Planning Problem (OD3P) by using a given number of flying drones, in order to efficiently recharge wireless sensor nodes. Unlike other works in the literature, we assume that the drones can trade altitude with coverage and recharge power, while each drone can move across different positions in the network to extend coverage. We present a linear program as well as a fast heuristic algorithm to meet the minimum energy demands of the nodes in the shortest possible amount of time. Our simulation results show the effectiveness of our approaches for network scenarios with up to 50 sensors and a 50×50m terrain size.

7.1.4.3. Efficient Data Collection and Tracking with Flying Drones

Data collection is an important mechanism for wireless sensor networks to be viable. In [34], we address the Aerial Data Collection Problem (ADCP) from a set of mobile wireless sensors located on the ground, using a fleet of flying devices. The objective is i) to deploy a set of UAVs in a 3D space to cover and collect data from all the mobile wireless sensors at each time step through a ground-to-air communication, ii) to send these data to a central base station using multi-hop wireless air-to-air communications through the network of UAVs, iii) while minimizing the total deployment cost (communication and deployment) over time. The Aerial Data Collection Problem (ADCP) is a complex time and space coverage, and connectivity problem. We first present a mixed-integer linear program solving ADCP optimally for small instances. Then, we develop a second model solved by column generation for larger instances, with optimal or heuristic pricing programs. Results show that our approach provides very accurate solutions minimizing the data collection cost. Moreover, only a very small number of columns are generated throughout the resolution process, showing the efficiency of our approach.

7.1.5. Other results

7.1.5.1. The Structured Way of Dealing with Heterogeneous Live Streaming Systems

In peer-to-peer networks for video live streaming, peers can share the forwarding load in two types of systems: unstructured and structured. In unstructured overlays, the graph structure is not well-defined, and a peer can obtain the stream from many sources. In structured overlays, the graph is organized as a tree rooted at the server and parent-child relationships are established between peers. Unstructured overlays ensure robustness and a higher degree of resilience compared to the structured ones. Indeed, they better manage the dynamics of peer participation or churn. Nodes can join and leave the system at any moment. However, they are less bandwidth efficient than structured overlays. In [54], we propose new simple distributed repair protocols for video live streaming structured systems. We show, through simulations and with real traces from Twitch, that structured systems can be very efficient and robust to failures, even for high churn and when peers have very heterogeneous upload bandwidth capabilities.

7.1.5.2. Optimal SF Allocation in LoRaWAN Considering Physical Capture and Imperfect Orthogonality

In [46], we propose a theoretical framework for maximizing the long range wide-area networks (LoRaWAN) capacity in terms of the number of end nodes, when they all have the same traffic generation process. The model optimally allocates the spreading factor to the nodes so that attenuation and collisions are optimized. We use an accurate propagation model considering Rayleigh channel, and we take into account physical capture and imperfect spreading factors (SF) orthogonality while guaranteeing a given transmission success probability to each served node in the network. Numerical results show the effectiveness of our SF allocation policy. Our framework also quantifies the maximum capacity of single cell networks and the gain induced by multiplying the gateways on the covered area. We finally evaluate the impact of physical capture and imperfect SF orthogonality on the SF allocation and network performances.

7.2. Graph Algorithms

Participants: Julien Bensmail, Jean-Claude Bermond, David Coudert, Frédéric Giroire, Frédéric Havet, Emanuele Natale, Nicolas Nisse, Stéphane Pérennes, Francois Dross, Fionn Mc Inerney, Thibaud Trolliet.

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

7.2.1. Complexity of graph problems

7.2.1.1. Fully Polynomial FPT Algorithms for Some Classes of Bounded Clique-width Graphs.

Recently, hardness results for problems in P were achieved using reasonable complexity theoretic assumptions such as the Strong Exponential Time Hypothesis. According to these assumptions, many graph theoretic problems do not admit truly subquadratic algorithms. A central technique used to tackle the difficulty of the above mentioned problems is fixed-parameter algorithms with polynomial dependency in the fixed parameter (P-FPT). Applying this technique to clique-width, an important graph parameter, remained to be done. In [35], we study several graph theoretic problems for which hardness results exist such as cycle problems, distance problems and maximum matching. We give hardness results and P-FPT algorithms, using clique-width and some of its upper bounds as parameters. We believe that our most important result is an algorithm in $O(k^4 \cdot n + m)$-time for computing a maximum matching where k is either the modular-width of the graph or the P_4-sparseness. The latter generalizes many algorithms that have been introduced so far for specific subclasses such as cographs. Our algorithms are based on preprocessing methods using modular decomposition and split decomposition. Thus they can also be generalized to some graph classes with unbounded clique-width.

7.2.1.2. *Explicit Linear Kernels for Packing Problems*

During the last years, several algorithmic meta-theorems have appeared (Bodlaender et al. [83], Fomin et al. [88], Kim et al. [90]) guaranteeing the existence of linear kernels on sparse graphs for problems satisfying some generic conditions. The drawback of such general results is that it is usually not clear how to derive from them constructive kernels with reasonably low explicit constants. To fill this gap, we recently presented [89] a framework to obtain explicit linear kernels for some families of problems whose solutions can be certified by a subset of vertices. In [37], we enhance our framework to deal with packing problems, that is, problems whose solutions can be certified by collections of *subgraphs* of the input graph satisfying certain properties. $\mathcal{F}$-Packing is a typical example: for a family $\mathcal{F}$ of connected graphs that we assume to contain at least one planar graph, the task is to decide whether a graph G contains k vertex-disjoint sub-graphs such that each of them contains a graph in $\mathcal{F}$ as a minor. We provide explicit linear kernels on sparse graphs for the following two orthogonal generalizations of $\mathcal{F}$-Packing: for an integer $\ell \geq 1$, one aims at finding either minor-models that are pairwise at distance at least ℓ in G (ℓ-$\mathcal{F}$-Packing), or such that each vertex in G belongs to at most ℓ minors-models ($\mathcal{F}$-Packing with-Membership). Finally, we also provide linear kernels for the versions of these problems where one wants to pack *subgraphs* instead of minors.

7.2.1.3. *Low Time Complexity Algorithms for Path Computation in Cayley Graphs.*

We study the problem of path computation in Cayley Graphs (CG) from an approach of word processing in groups. This approach consists in encoding the topological structure of CG in an automaton called Diff, then techniques of word processing are applied for computing the shortest paths. In [17], we present algorithms for computing the K-shortest paths, the shortest disjoint paths and the shortest path avoiding a set of nodes and edges. For any CG with diameter D, the time complexity of the proposed algorithms is $O(KD|\text{Diff}|)$, where $|\text{Diff}|$ denotes the size of Diff. We show that our proposal outperforms the state of art of topology-agnostic algorithms for disjoint shortest paths and stays competitive with respect to proposals for specific families of CG. Therefore, the proposed algorithms set a base in the design of adaptive and low-complexity routing schemes for networks whose interconnections are defined by CG.

7.2.1.4. *Convex hull in graphs.*

In [40], we prove that, given a closure function the smallest preimage of a closed set can be calculated in polynomial time in the number of closed sets. This implies that there is a polynomial time algorithm to compute the convex hull number of a graph, when all its convex subgraphs are given as input. We then show that deciding if the smallest preimage of a closed set is logarithmic in the size of the ground set is LOGSNP-hard if only the ground set is given. A special instance of this problem is to compute the dimension of a poset given its linear extension graph, that is conjectured to be in P.
The intent to show that the latter problem is LOGSNP-complete leads to several interesting questions and to the definition of the isometric hull, i.e., a smallest isometric subgraph containing a given set of vertices S. While for $|S| = 2$ an isometric hull is just a shortest path, we show that computing the isometric hull of a set of vertices is NP-complete even if $|S| = 3$. Finally, we consider the problem of computing the isometric hull number of a graph and show that computing it is Σ_2^P complete.

7.2.2. *Combinatorial games in graphs*

7.2.2.1. *Graph searching and combinatorial games in graphs.*

The Network Decontamination problem consists of coordinating a team of mobile agents in order to clean a contaminated network. The problem is actually equivalent to tracking and capturing an invisible and arbitrarily fast fugitive. This problem has natural applications in network security in computer science or in robotics for search or pursuit-evasion missions. Many different objectives have been studied: the main one being the minimization of the number of mobile agents necessary to clean a contaminated network.

Many environments (continuous or discrete) have also been considered. In the book chapter [61], we focus on networks modeled by graphs. In this context, the optimization problem that consists of minimizing the number of agents has a deep graph-theoretical interpretation. Network decontamination and, more precisely, *graph searching* models, provide nice algorithmic interpretations of fundamental concepts in the Graph Minors theory by Robertson and Seymour.

For all these reasons, graph searching variants have been widely studied since their introduction by Breish (1967) and mathematical formalizations by Parsons (1978) and Petrov (1982). The book chapter [61] consists of an overview of the algorithmic results on graph decontamination and graph searching. Moreover, [19] is the preface to the special issue of TCS on the 8th Workshop on GRAph Searching, Theory and Applications, Anogia, Crete, Greece, April 10 - April 13, 2017.

In [52], we focus on another game with mobile agents in a graph. Precisely, in the eternal domination game played on graphs, an attacker attacks a vertex at each turn and a team of guards must move a guard to the attacked vertex to defend it. The guards may only move to adjacent vertices on their turn. The goal is to determine the eternal domination number γ^{∞}_{all} of a graph which is the minimum number of guards required to defend against an infinite sequence of attacks. [52] continues the study of the eternal domination game on strong grids $P_n \boxtimes P_m$. Cartesian grids $P_n \square P_m$ have been vastly studied with tight bounds existing for small grids such as $k \times n$ grids for $k \in \{2,3,4,5\}$. It was recently proven that $\gamma^{\infty}_{all}(P_n \square P_m) = \gamma(P_n \square P_m) + O(n+m)$ where $\gamma(P_n \square P_m)$ is the domination number of $P_n \square P_m$ which lower bounds the eternal domination number [91]. We prove that, for all $n, m \in \mathbb{N}^*$ such that $m \geq n$, $\lfloor \frac{nm}{9} \rfloor + \Omega(n+m) = \gamma^{\infty}_{all}(P_n \boxtimes P_m) = \lceil \frac{nm}{9} \rceil + O(m\sqrt{n})$ (note that $\lceil \frac{nm}{9} \rceil$ is the domination number of $P_n \boxtimes P_m$). Our technique may be applied to other “grid-like" graphs.

In [66], we adapt the techniques of [91] to prove that the eternal domination number of strong grids is upper bounded by $\frac{mn}{7} + O(m+n)$. While this does not improve upon a recently announced bound of $\lceil \frac{m}{3} \rceil \lceil \frac{n}{3} \rceil + O(m\sqrt{n})$ [52] in the general case, we show that our bound is an improvement in the case where the smaller of the two dimensions is at most 6179.

7.2.2.2. The Orthogonal Colouring Game

In [18], we introduce the Orthogonal Colouring Game, in which two players alternately colour vertices (from a choice of $m \in N$ colours) of a pair of isomorphic graphs while respecting the properness and the orthogonality of the colouring. Each player aims to maximize her score, which is the number of coloured vertices in the copy of the graph she owns. Our main result is that the second player has a strategy to force a draw in this game for any $m \in N$ for graphs that admit a strictly matched involution. An involution σ of a graph G is strictly matched if its fixed point set induces a clique and any non-fixed point $v \in V(G)$ is connected with its image $\sigma(v)$ by an edge. We give a structural characterization of graphs admitting a strictly matched involution and bounds for the number of such graphs. Examples of such graphs are the graphs associated with Latin squares and sudoku squares.

In [62], we prove that recognising graphs that admit a strictly matched involution is NP-complete.

7.2.2.3. Complexity of Games Compendium

Since games and puzzles have been studied under a computational lens, researchers unearthed a rich landscape of complexity results showing deep connections between games and fundamental problems and models in computer science. Complexity of Games (CoG, https://steven3k.gitlab.io/isnphard-test/) is a compendium of complexity results on games and puzzles. It aims to serve as a reference guide for enthusiasts and researchers on the topic and is a collaborative and open source project that welcomes contributions from the community.

7.2.3. Algorithms for social networks

7.2.3.1. KADABRA, an ADaptive Algorithm for Betweenness via Random Approximation

In [32], we present KADABRA, a new algorithm to approximate betweenness centrality in directed and undirected graphs, which significantly outperforms all previous approaches on real-world complex networks. The efficiency of the new algorithm relies on two new theoretical contributions, of independent interest. The first contribution focuses on sampling shortest paths, a subroutine used by most algorithms that approximate betweenness centrality. We show that, on realistic random graph models, we can perform this task in time $|E|^{\frac{1}{2}+o(1)}$ with high probability, obtaining a significant speedup with respect to the $\Theta(|E|)$ worst-case performance. We experimentally show that this new technique achieves similar speedups on real-world complex networks, as well. The second contribution is a new rigorous application of the adaptive sampling technique. This approach decreases the total number of shortest paths that need to be sampled to compute all betweenness centralities with a given absolute error, and it also handles more general problems, such as computing the k most central nodes. Furthermore, our analysis is general, and it might be extended to other settings.

7.2.3.2. Distributed Community Detection via Metastability of the 2-Choices Dynamics

In [56], we investigate the behavior of a simple majority dynamics on networks of agents whose interaction topology exhibits a community structure. By leveraging recent advancements in the analysis of dynamics, we prove that, when the states of the nodes are randomly initialized, the system rapidly and stably converges to a configuration in which the communities maintain internal consensus on different states. This is the first analytical result on the behavior of dynamics for non-consensus problems on non-complete topologies, based on the first symmetry-breaking analysis in such setting. Our result has several implications in different contexts in which dynamics are adopted for computational and biological modeling purposes. In the context of Label Propagation Algorithms, a class of widely used heuristics for community detection, it represents the first theoretical result on the behavior of a distributed label propagation algorithm with quasi-linear message complexity. In the context of evolutionary biology, dynamics such as the Moran process have been used to model the spread of mutations in genetic populations [Lieberman, Hauert, and Nowak 2005]; our result shows that, when the probability of adoption of a given mutation by a node of the evolutionary graph depends super-linearly on the frequency of the mutation in the neighborhood of the node and the underlying evolutionary graph exhibits a community structure, there is a non-negligible probability for species differentiation to occur.

7.2.3.3. On the Necessary Memory to Compute the Plurality in Multi-Agent Systems

Consensus and Broadcast are two fundamental problems in distributed computing, whose solutions have several applications. Intuitively, Consensus should be no harder than Broadcast, and this can be rigorously established in several models. Can Consensus be easier than Broadcast?

In models that allow noiseless communication, we prove in [60] a reduction of (a suitable variant of) Broadcast to binary Consensus, that preserves the communication model and all complexity parameters such as randomness, number of rounds, communication per round, etc., while there is a loss in the success probability of the protocol. Using this reduction, we get, among other applications, the first logarithmic lower bound on the number of rounds needed to achieve Consensus in the uniform GOSSIP model on the complete graph. The lower bound is tight and, in this model, Consensus and Broadcast are equivalent.

We then turn to distributed models with noisy communication channels that have been studied in the context of some bio-inspired systems. In such models, only one noisy bit is exchanged when a communication channel is established between two nodes, and so one cannot easily simulate a noiseless protocol by using error-correcting codes. An $\Omega(\epsilon^{-2}n)$ lower bound on the number of rounds needed for Broadcast is proved by Boczkowski et al. [82] in one such model (noisy uniform PULL, where ϵ is a parameter that measures the amount of noise). In such model, we prove a new $\Theta(\epsilon^{-2}n\log n)$ bound for Broadcast and a $\Theta(\epsilon^{-2}\log n)$ bound for binary Consensus, thus establishing an exponential gap between the number of rounds necessary for Consensus versus Broadcast.

7.2.3.4. How long does it take for all users in a social network to choose their communities?

In [30], we consider a community formation problem in social networks, where the users are either friends or enemies. The users are partitioned into conflict-free groups (i.e., independent sets in the conflict graph $G^- = (V, E)$ that represents the enmities between users). The dynamics goes on as long as there exists any set of at most k users, k being any fixed parameter, that can change their current groups in the partition simultaneously, in such a way that they all strictly increase their utilities (number of friends i.e., the cardinality of their respective groups minus one). Previously, the best-known upper-bounds on the maximum time of convergence were $O(|V|\alpha(G^-))$ for $k \leq 2$ and $O(|V|^3)$ for $k = 3$, with $\alpha(G^-)$ being the independence number of G^-. Our first contribution consists in reinterpreting the initial problem as the study of a dominance ordering over the vectors of integer partitions. With this approach, we obtain for $k \leq 2$ the tight upper-bound $O(|V| \min\left\{\alpha(G^-), \sqrt{|V|}\right\})$ and, when G^- is the empty graph, the exact value of order $\frac{(2|V|)^{3/2}}{3}$. The time of convergence, for any fixed $k \geq 4$, was conjectured to be polynomial. In this paper we disprove this. Specifically, we prove that for any $k \geq 4$, the maximum time of convergence is in $\Omega(|V|\Theta(\log |V|))$.

7.2.3.5. A Comparative Study of Neural Network Compression

There has recently been an increasing desire to evaluate neural networks locally on computationally-limited devices in order to exploit their recent effectiveness for several applications; such effectiveness has nevertheless come together with a considerable increase in the size of modern neural networks, which constitute a major downside in several of the aforementioned computationally-limited settings. There has thus been a demand of compression techniques for neural networks. Several proposal in this direction have been made, which famously include hashing-based methods and pruning-based ones. However, the evaluation of the efficacy of these techniques has so far been heterogeneous, with no clear evidence in favor of any of them over the others. In [70], we address this latter issue by providing a comparative study. While most previous studies test the capability of a technique in reducing the number of parameters of state-of-the-art networks, we follow [86] in evaluating their performance on basic architectures on the MNIST dataset and variants of it, which allows for a clearer analysis of some aspects of their behavior. To the best of our knowledge, we are the first to directly compare famous approaches such as HashedNet, Optimal Brain Damage (OBD), and magnitude-based pruning with L1 and L2 regularization among them and against equivalent-size feed-forward neural networks with simple (fully-connected) and structural (convolutional) neural networks. Rather surprisingly, our experiments show that (iterative) pruning-based methods are substantially better than the HashedNet architecture, whose compression doesn't appear advantageous to a carefully chosen convolutional network. We also show that, when the compression level is high, the famous OBD pruning heuristics deteriorates to the point of being less efficient than simple magnitude-based techniques.

7.3. Graph and digraph theory

Participants: Julien Bensmail, Frédéric Havet, Nicolas Nisse, Stéphane Pérennes, Francois Dross, Fionn Mc Inerney, Thi Viet Ha Nguyen, Nathann Cohen.

COATI studies theoretical problems in graph theory. If some of them are directly motivated by applications, others are more fundamental.

We are putting an effort on understanding better directed graphs (also called *digraphs*) and partitioning problems, and in particular colouring problems. We also try to better the understand the many relations between orientations 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.

To ease the reading, we split our results in this section into several subsections dedicated to particular topics.

7.3.1. Graph and digraph colourings

7.3.1.1. Distinguishing labellings and the 1-2-3 Conjecture

We are interested in several distinguishing labelling (or edge-weighting) problems, where the general aim, given a graph, is to label the edges in such a way that certain properties are fulfilled. The main problem we have been considering is the **1-2-3 Conjecture**, which claims that every connected graph different from K_2 admits a labelling with $1, 2, 3$ such that no two adjacent vertices are incident to the same sum of weights. Some of our latest results provide evidence towards the 1-2-3 Conjecture. We also investigated questions inspired from the conjecture, such that the role of the weights $1, 2, 3$ in the statement of the conjecture, the deep connection with proper vertex-colourings and other standard notions of graph theory.

7.3.1.2. A 1-2-3-4 result for the 1-2-3 Conjecture in 5-regular graphs

To date, the best-known result towards the 1-2-3 Conjecture is due to Kalkowski, Karoński and Pfender, who proved that it holds when relaxed to 5-edge-weightings. Their proof builds upon a weighting algorithm designed by Kalkowski for a total version of the problem. In [23], we present new mechanisms for using Kalkowski's algorithm in the context of the 1-2-3 Conjecture. As a main result we prove that every 5-regular graph admits a 4-edge-weighting that permits to distinguish its adjacent vertices via their incident sums.

7.3.1.3. On $\{a, b\}$-edge-weightings of bipartite graphs with odd a, b

For any $S \subset \mathbb{Z}$ we say that a graph G has the S-property if there exists an S-edge-weighting $w : E(G) \to S$ such that for any pair of adjacent vertices u, v we have $\sum_{e \in E(v)} w(e) \neq \sum_{e \in E(u)} w(e)$, where $E(v)$ and $E(u)$ are the sets of edges incident to v and u, respectively. In general, deciding if a graph G has the $\{a, b\}$-property is NP-complete for every a, b. This question is open for bipartite graphs however. The only known results of this sort are that bipartite graphs without the $\{1, 2\}$-property can be recognized easily, and similarly for 2-connected bipartite graphs without the $\{0, 1\}$-property. In [28], we focus on $\{a, a + 2\}$-edge-weightings where $a \in \mathbb{Z}$ is odd. We show that a 2-connected bipartite graph has the $\{a, a + 2\}$-property if and only if it is not a so-called odd multi-cactus. In the case of trees, we show that only one case is pathological. That is, we show that all trees have the $\{a, a + 2\}$-property for odd $a \neq -1$, while there is an easy characterization of trees without the $\{-1, 1\}$-property.

7.3.1.4. 1-2-3 Conjecture in Digraphs: More Results and Directions

When arc-weighting a digraph, there are, at each vertex, two sums of incident weights: the in-coming sum σ^- and the out-going sum σ^+. Thus, there are many ways for generalizing the 1-2-3 Conjecture to digraphs. In the recent years, four main variants have been considered, where, for every arc $\overrightarrow{uv}$, it is required that one of $\sigma^-(u), \sigma^+(u)$ is different from one of $\sigma^-(v), \sigma^+(v)$. All of these four variants are well understood, except for the one where, for every arc $\overrightarrow{uv}$, it is required that $\sigma^-(u) \neq \sigma^+(v)$. Regarding this version, Horňak, Przybyło and Woźniak recently proved that almost every digraph can be 4-arc-weighted so that, for every arc $\overrightarrow{uv}$, the sum of weights incoming to u is different from the sum of weights outgoing from v. They conjectured a stronger result, namely that the same statement with 3 instead of 4 should also be true. We verify this conjecture in [73]. This work takes place in a recent "quest" towards a directed version of the 1-2-3 Conjecture, the variant above being one of the last introduced ones. We take the occasion of this work to establish a summary of all results known in this field, covering known upper bounds, complexity aspects, and choosability. On the way we prove additional results which were missing in the whole picture. We also mention the aspects that remain open.

7.3.1.5. Edge Weights and Vertex Colours: Minimizing Sum Count

Put differently, the 1-2-3 Conjecture asks whether, via weights with very low magnitude, we can "encode" a proper vertex-colouring of any graph. Note, however, that we do not care about whether such a result colouring is optimal, i.e., whether its number of colours is close to the chromatic number. In [22], we investigate the minimum number of distinct sums/colours we can produce via a neighbour-sum-distinguishing edge-weighting of a given graph G, and the role of the assigned weights in that context. Clearly, this minimum number is bounded below by the chromatic number $\chi(G)$ of G. When using weights of $\mathbb{Z}$, we show that, in general, we can produce neighbour-sum-distinguishing edge-weightings generating $\chi(G)$ distinct sums, except in the peculiar case where G is a balanced bipartite graph, in which case $\chi(G) + 1$ distinct sums can be generated. These results are best possible. When using k consecutive weights $1, ..., k$, we provide both lower and upper bounds, as a function of the maximum degree Δ, on the maximum least number of sums that can be generated for a graph with maximum degree Δ. For trees, which, in general, admit neighbour-sum-distinguishing 2-edge-weightings, we prove that this maximum, when using weights 1 and 2, is of order $2 \log_2 \Delta$. Finally, we also establish the NP-hardness of several decision problems related to these questions.

7.3.1.6. On Minimizing the Maximum Color for the 1-2-3 Conjecture

In the line of the previous investigation, one way to get some sort of progress is to design proper labellings where the maximum color of a vertex is as small as possible. In [64], we investigate the consequences of labeling graphs as in the 1-2-3 Conjecture when it is further required to make the maximum resulting color as small as possible. We first investigate the hardness of determining the minimum maximum color by a labeling for a given graph, which we show is NP-complete in the class of bipartite graphs but polynomial-time solvable in the class of graphs with bounded treewidth. We then provide bounds on the minimum maximum color that can be generated both in the general context, and for particular classes of graphs. Finally, we study how using larger labels permits to reduce the maximum color.

7.3.1.7. Decomposing degenerate graphs into locally irregular subgraphs

A (undirected) graph is locally irregular if no two of its adjacent vertices have the same degree. A decomposition of a graph G into k locally irregular subgraphs is a partition $E_1, ..., E_k$ of $E(G)$ into k parts each of which induces a locally irregular subgraph. Not all graphs decompose into locally irregular subgraphs; however, it was conjectured that, whenever a graph does, it should admit such a decomposition into at most three locally irregular subgraphs. This conjecture was verified for a few graph classes in recent years. It was introduced because it was noticed that, in some contexts, there are connections between locally irregular decompositions and the 1-2-3 Conjecture. In [63], we consider the decomposability of degenerate graphs with low degeneracy. Our main result is that decomposable k-degenerate graphs decompose into at most $3k+1$ locally irregular subgraphs, which improves on previous results whenever $k \leq 9$. We improve this result further for some specific classes of degenerate graphs, such as bipartite cacti, k-trees, and planar graphs. Although our results provide only little progress towards the leading conjecture above, the main contribution of this work is rather the decomposition schemes and methods we introduce to prove these results.

7.3.1.8. A general decomposition theory for the 1-2-3 Conjecture and locally irregular decompositions

In [21], we propose an approach encapsulating locally irregular decompositions and proper labelings. As a consequence, we get another interpretation of several existing results related to the 1-2-3 Conjecture. We also come up with new related conjectures, to which we give some support.

7.3.1.9. Decomposability of graphs into subgraphs fulfilling the 1-2-3 Conjecture

In particular, one of the side problems we run into is decomposing graphs into subgraphs verifying the 1-2-3 Conjecture. In [29], we prove that every d-regular graph, $d \geq 2$, can be decomposed into at most 2 subgraphs (without isolated edges) fulfilling the 1-2-3 Conjecture if $d \notin \{10, 11, 12, 13, 15, 17\}$, and into at most 3 such subgraphs in the remaining cases. Additionally, we prove that in general every graph without isolated edges can be decomposed into at most 24 subgraphs fulfilling the 1–2–3 Conjecture, improving the previously best upper bound of 40. Both results are partly based on applications of the Lovász Local Lemma.

7.3.1.10. On the 2-edge-coloured chromatic number of grids

The oriented (2-edge-coloured, respectively) chromatic number $\chi_o(G)$ ($\chi_2(G)$, respectively) of an undirected graph G is defined as the maximum oriented (2-edge-coloured, respectively) chromatic number of an orientation (signature, respectively) of G. Although the difference between $\chi_o(G)$ and $\chi_2(G)$ can be arbitrarily large, there are, however, contexts in which these two parameters are quite comparable. In [24], we compare the behaviour of these two parameters in the context of (square) grids. While a series of works has been dedicated to the oriented chromatic number of grids, we are not aware of any work dedicated to their 2-edge-coloured chromatic number. We investigate this throughout this paper. We show that the maximum 2-edge-coloured chromatic number of a grid lies between 8 and 11. We also focus on 2-row grids and 3-row grids, and exhibit bounds on their 2-edge-coloured chromatic number, some of which are tight. Although our results indicate that the oriented chromatic number and the 2-edge-coloured chromatic number of grids are close in general, they also show that these parameters may differ, even for easy instances.

7.3.1.11. From light edges to strong edge-colouring of 1-planar graphs

A strong edge-colouring of an undirected graph G is an edge-colouring where every two edges at distance at most 2 receive distinct colours. The strong chromatic index of G is the least number of colours in a strong edge-colouring of G. A conjecture of Erdős and Nešetřil, stated back in the 80's, asserts that every graph with maximum degree Δ should have strong chromatic index at most roughly $1.25\Delta^2$. Several works in the last decades have confirmed this conjecture for various graph classes. In particular, lots of attention have been dedicated to planar graphs, for which the strong chromatic index decreases to roughly 4Δ, and even to smaller values under additional structural requirements. In [26], we initiate the study of the strong chromatic index of 1-planar graphs, which are those graphs that can be drawn on the plane in such a way that every edge is crossed at most once. We provide constructions of 1-planar graphs with maximum degree Δ and strong chromatic index roughly 6Δ. As an upper bound, we prove that the strong chromatic index of a 1-planar graph with maximum degree Δ is at most roughly 24Δ (thus linear in Δ). In the course of proving the latter result, we prove, towards a conjecture of Hudák and Šugerek, that 1-planar graphs with minimum degree 3 have edges both of whose ends have degree at most 29.

7.3.1.12. Pushable chromatic number of graphs with degree constraints

Pushable homomorphisms and the pushable chromatic number χ_p of oriented graphs were introduced by Klostermeyer and MacGillivray in 2004. They notably observed that, for any oriented graph $\overrightarrow{G}$, we have $\chi_p(\overrightarrow{G}) \leq \chi_o(\overrightarrow{G}) \leq 2\chi_p(\overrightarrow{G})$, where $\chi_o(\overrightarrow{G})$ denotes the oriented chromatic number of $\overrightarrow{G}$. This stands as first general bounds on χ_p. This parameter was further studied in later works.

In [71], we consider the pushable chromatic number of oriented graphs fulfilling particular degree conditions. For all $\Delta \geq 29$, we first prove that the maximum value of the pushable chromatic number of an oriented graph with maximum degree Δ lies between $2^{\frac{\Delta}{2}-1}$ and $(\Delta-3)\cdot(\Delta-1)\cdot 2^{\Delta-1}+2$ which implies an improved bound on the oriented chromatic number of the same family of graphs. For subcubic oriented graphs, that is, when $\Delta \leq 3$, we then prove that the maximum value of the pushable chromatic number is 6 or 7. We also prove that the maximum value of the pushable chromatic number of oriented graphs with maximum average degree less than 3 lies between 5 and 6. The former upper bound of 7 also holds as an upper bound on the pushable chromatic number of planar oriented graphs with girth at least 6.

7.3.2. Graph and digraph decompositions

7.3.2.1. Edge-partitioning a graph into paths: beyond the Barát-Thomassen conjecture

In 2006, Barát and Thomassen conjectured that there is a function f such that, for every fixed tree T with t edges, every $f(t)$-edge-connected graph with its number of edges divisible by t has a partition of its edges into copies of T. This conjecture was recently verified in [81] by, in particular, some members of COATI. In [27], we further focus on the path case of the Barát-Thomassen conjecture. Before the aforementioned general proof was announced, several successive steps towards the path case of the conjecture were made, notably by Thomassen [94], [95], [96], until this particular case was totally solved by Botler, Mota, Oshiro and Wakabayashi [84]. Our goal in this work was to propose an alternative proof of the path case with a weaker hypothesis: Namely, we prove that there is a function f such that every 24-edge-connected graph with minimum degree $f(t)$ has an edge-partition into paths of length t whenever t divides the number of edges. We also show that 24 can be dropped to 4 when the graph is eulerian.

7.3.2.2. Constrained ear decompositions in graphs and digraphs.

Ear decompositions of graphs are a standard concept related to several major problems in graph theory like the Traveling Salesman Problem. For example, the Hamiltonian Cycle Problem, which is notoriously NP-complete, is equivalent to deciding whether a given graph admits an ear decomposition in which all ears except one are trivial (i.e. of length 1). On the other hand, a famous result of Lovász states that deciding whether a graph admits an ear decomposition with all ears of odd length can be done in polynomial time. In [38], we study the complexity of deciding whether a graph admits an ear decomposition with prescribed ear lengths. We prove that deciding whether a graph admits an ear decomposition with all ears of length at most ℓ is polynomial-time solvable for all fixed positive integer ℓ. On the other hand, deciding whether a graph admits an ear decomposition without ears of length in $\mathcal{F}$ is NP-complete for any finite set $\mathcal{F}$ of positive integers. We also prove that, for any $k \geq 2$, deciding whether a graph admits an ear decomposition with all ears of length $0 \mod k$ is NP-complete.

We also consider the directed analogue to ear decomposition, which we call handle decomposition, and prove analogous results : deciding whether a digraph admits a handle decomposition with all handles of length at most ℓ is polynomial-time solvable for all positive integer ℓ; deciding whether a digraph admits a handle decomposition without handles of length in $\mathcal{F}$ is NP-complete for any finite set $\mathcal{F}$ of positive integers (and minimizing the number of handles of length in $\mathcal{F}$ is not approximable up to $n(1-\epsilon)$); for any $k \geq 2$, deciding whether a digraph admits a handle decomposition with all handles of length $0 \mod k$ is NP-complete. Also, in contrast with the result of Lovász, we prove that deciding whether a digraph admits a handle decomposition with all handles of odd length is NP-complete. Finally, we conjecture that, for every set $\mathcal{A}$ of integers, deciding whether a digraph has a handle decomposition with all handles of length in $\mathcal{A}$ is NP-complete, unless there exists $h \in \mathbb{N}$ such that $\mathcal{A} = \{1, \cdots, h\}$.

7.3.3. Substructures in graphs and digraphs

7.3.3.1. Subdivisions in Digraphs of Large Out-Degree or Large Dichromatic Number

In 1985, Mader conjectured the existence of a function f such that every digraph with minimum out-degree at least $f(k)$ contains a subdivision of the transitive tournament of order k. This conjecture is still completely open, as the existence of $f(5)$ remains unknown. In this paper, we show that if D is an oriented path, or an in-arborescence (i.e., a tree with all edges oriented towards the root) or the union of two directed paths from x to y and a directed path from y to x, then every digraph with minimum out-degree large enough contains a subdivision of D. Additionally, we study Mader's conjecture considering another graph parameter. The dichromatic number of a digraph D is the smallest integer k such that D can be partitioned into k acyclic subdigraphs. We show in [16] that any digraph with dichromatic number greater than $4m(n-1)$ contains every digraph with n vertices and m arcs as a subdivision.

7.3.3.2. Bipartite spanning sub(di)graphs induced by 2-partitions

For a given 2-partition (V_1, V_2) of the vertices of a (di)graph G, we study properties of the spanning bipartite subdigraph $BG(V_1, V_2)$ of G induced by those arcs/edges that have one end in each $V_i, i \in \{1, 2\}$. In [20], we determine, for all pairs of non-negative integers k_1, k_2, the complexity of deciding whether G has a 2-partition (V_1, V_2) such that each vertex in V_i (for $i \in \{1, 2\}$) has at least k_i (out-)neighbours in V_{3-i}. We prove that it is NP-complete to decide whether a digraph D has a 2-partition (V_1, V_2) such that each vertex in V_1 has an out-neighbour in V_2 and each vertex in V_2 has an in-neighbour in V_1. The problem becomes polynomially solvable if we require D to be strongly connected. We give a characterization of the structure of NP-complete instances in terms of their strong component digraph. When we want higher in-degree or out-degree to/from the other set the problem becomes NP-complete even for strong digraphs. A further result is that it is NP-complete to decide whether a given digraph D has a 2-partition (V_1, V_2) such that $BD(V_1, V_2)$ is strongly connected. This holds even if we require the input to be a highly connected eulerian digraph.

7.3.3.3. Metric Dimension: from Graphs to Oriented Graphs

The metric dimension $MD(G)$ of an undirected graph G is the cardinality of a smallest set of vertices that allows, through their distances to all vertices, to distinguish any two vertices of G. Many aspects of this notion have been investigated since its introduction in the 70's, including its generalization to digraphs.

In [42], [43], we study, for particular graph families, the maximum metric dimension over all strongly-connected orientations, by exhibiting lower and upper bounds on this value. We first exhibit general bounds for graphs with bounded maximum degree. In particular, we prove that, in the case of subcubic n-node graphs, all strongly-connected orientations asymptotically have metric dimension at most $\frac{n}{2}$, and that there are such orientations having metric dimension $\frac{2n}{5}$. We then consider strongly-connected orientations of grids. For a torus with n rows and m columns, we show that the maximum value of the metric dimension of a strongly-connected Eulerian orientation is asymptotically $\frac{nm}{2}$ (the equality holding when n, m are even, which is best possible). For a grid with n rows and m columns, we prove that all strongly-connected orientations asymptotically have metric dimension at most $\frac{2nm}{3}$, and that there are such orientations having metric dimension $\frac{nm}{2}$.

7.3.4. Bio-informatics motivated problems

7.3.4.1. Overlaying a hypergraph with a graph with bounded maximum degree

A major problem in structural biology is the characterization of low resolution structures of macro-molecular assemblies. One subproblem of this very difficult question is to determine the plausible contacts between the subunits (e.g. proteins) of an assembly, given the lists of subunits involved in all the complexes. This problem can be conveniently modelled by graphs and hypergraphs. Let G and H be respectively a graph and a hypergraph defined on a same set of vertices, and let F be a fixed graph. We say that GF-overlays a hyperedge S of H if F is a spanning subgraph of the subgraph of G induced by S, and that it F-overlays H if it F-overlays every hyperedge of H. Motivated by structural biology, we study in [68] the computational complexity of two problems. The first problem, $(\Delta \leq k)F$-Overlay, consists in deciding whether there is a graph with maximum degree at most k that F-overlays a given hypergraph H. It is a particular case of the second problem Max $(\Delta \leq k)F$-Overlay, which takes a hypergraph H and an integer s as input, and consists in deciding whether there is a graph with maximum degree at most k that F-overlays at least s hyperedges of H. We give a complete polynomial/NP-complete dichotomy for the Max $(\Delta \leq k)F$-Overlay problems depending on the pairs (F, k), and establish the complexity of $(\Delta \leq k)F$-Overlay for many pairs (F, k).

DANTE Project-Team

7. New Results

7.1. Graph Signal Processing and Machine Learning

Participants: Paulo Gonçalves, Rémi Gribonval, Marion Foare, Thomas Begin, Esteban Bautista Ruiz, Gaetan Frusque, Amélie Barbe, Mikhail Tsitsvero, Marija Stojanova, Márton Karsai, Sébastien Lerique, Jacobo Levy Abitbol.

7.1.1. L^γ -PageRank for Semi-Supervised Learning

Participants: Paulo Gonçalves, Esteban Bautista Ruiz.

PageRank for Semi-Supervised Learning has shown to leverage data structures and limited tagged examples to yield meaningful classification. Despite successes, classification performance can still be improved, particularly in cases of fuzzy graphs or unbalanced labeled data. To address such limitations, a novel approach based on powers of the Laplacian matrix L^γ ($\gamma > 0$), referred to as L^γ-PageRank, is proposed. Its theoretical study shows that it operates on signed graphs, where nodes belonging to one same class are more likely to share positive edges while nodes from different classes are more likely to be connected with negative edges. It is shown that by selecting an optimal γ, classification performance can be significantly enhanced. A procedure for the automated estimation of the optimal γ, from a unique observation of data, is devised and assessed. Experiments on several datasets demonstrate the effectiveness of both L^γ-PageRank classification and the optimal γ estimation. [11]

7.1.2. Designing Convex Combination of Graph Filters

Participant: Paulo Gonçalves.

In this work, we studied the problem of parametric modeling of network-structured signals with graph filters. Unlike the popular polynomial graph filters, which are based on a single graph shift operator, we considered convex combinations of graph shift operators particularly adapted to directed graphs. As the resulting modeling problem is not convex, we reformulated it as a convex optimization problem which can be solved efficiently. Experiments on real-world data structured by undirected and directed graphs were conducted. The results showed the effectiveness of this method compared to other methods reported in the literature. [18]

7.1.3. Optimal transport under regularity constraints for domain adaptation between graphs with attributes

Participants: Paulo Gonçalves, Amélie Barbe.

In this work, we addresses the problem of domain adaptation between two graphs by optimal transport. We aimed at benefiting from the knowledge of a labeled source graph to improve the classification of nodes in an unlabeled target graph. We focused on the setting where a set of features is associated to each node of the graphs. We proposed an original method that optimizes a transportation plan from the source to the target that *(i)* preserves the structures transported between the graphs and *(ii)* prevents the mapping from transporting two source nodes with different labels to the same destination. [30]

7.1.4. Sparse tensor dimensionality reduction with application to the clustering of functional connectivity in the brain

Participants: Paulo Gonçalves, Gaetan Frusque.

Functional connectivity (FC) is a graph-like data structure commonly used by neuroscientists to study the dynamic behaviour of the brain activity. However, these analyses rapidly become complex and time-consuming, as the number of connectivity components to be studied is quadratic with the number of electrodes. In our work, we addressed the problem of clustering FC into relevant ensembles of simultaneously activated components that reveal characteristic patterns of the epileptic seizures of a given patient. While $k-$means is certainly the most popular method for data clustering, it is known to perform badly on large dimensional data sets, and to be highly sensitive to noise. To overcome the co-called curse of dimensionality, we proposed a new tensor decomposition to reduce the size of the data set formed by FC time series recorded for several seizures, before applying k-means. Our contribution is twofold: First, we derived a method that we compared to the state of the art, emphasizing one variant that imposes sparsity constraints. Second, we conducted a real case study, applying the proposed sparse tensor decomposition to epileptic data in order to infer the functional connectivity graph dynamics corresponding to the different stages of an epileptic seizure. [31], [47]

7.1.5. Graph signal processing to model WLANs performances

Participants: Paulo Gonçalves, Thomas Begin, Marija Stojanova.

As WLANs have become part of our everyday life, there is an increasing need for more transmission capacity and wireless coverage. In response to this growing need, network administrators tend to intensify the deployment of Access Points (APs). However, if not correctly done, this AP densification may lead to badly planned and uncoordinated networks with sub-optimal use of the available resources. In this work, we propose a data-driven approach using graph signal processing and a set of input/output signals to capture the behavior of a WLAN and derive a predictive performance model. Given the simplicity and the novelty of the proposed model, we believe that its relative error of around 10-20% in modeling and 25% in prediction may represent a promising start for new approaches in the modeling of WLANs. [33]

7.1.6. Joint embedding of structure and features via graph convolutional networks

Participants: Márton Karsai, Sébastien Lerique.

We propose *AN2VEC*, a node embedding method which ultimately aims at disentangling the information shared by the structure of a network and the features of its nodes. Building on the recent developments of Graph Convolutional Networks (GCN), we develop a multitask GCN Variational Autoencoder where different dimensions of the generated embeddings can be dedicated to encoding feature information, network structure, and shared feature-network information. We explore the interaction between these disentangled characters by comparing the embedding reconstruction performance to a baseline case where no shared information is extracted. We use synthetic datasets with different levels of interdependency between feature and network characters and show (i) that shallow embeddings relying on shared information perform better than the corresponding reference with unshared information, (ii) that this performance gap increases with the correlation between network and feature structure, and (iii) that our embedding is able to capture joint information of structure and features. Our method can be relevant for the analysis and prediction of any featured network structure ranging from online social systems to network medicine. [51]

7.2. Computational Human Dynamics and Temporal Networks

Participants: Márton Karsai, Sébastien Lerique, Jacobo Levy Abitbol, Samuel Unicomb, Sicheng Dai.

7.2.1. Optimal Proxy Selection for Socioeconomic Status Inference on Twitter

Participants: Márton Karsai, Jacobo Levy Abitbol.

The socioeconomic status of people depends on a combination of individual characteristics and environmental variables, thus its inference from online behavioral data is a difficult task. Attributes like user semantics in communication, habitat, occupation, or social network are all known to be determinant predictors of this feature. In this paper we propose three different data collection and combination methods to first estimate and, in turn, infer the socioeconomic status of French Twitter users from their online semantics. Our methods are based on open census data, crawled professional profiles, and remotely sensed, expert annotated information on living environment. Our inference models reach similar performance of earlier results with the advantage of relying on broadly available datasets and of providing a generalizable framework to estimate socioeconomic status of large numbers of Twitter users. These results may contribute to the scientific discussion on social stratification and inequalities, and may fuel several applications. [19]

7.2.2. Randomized reference models for temporal networks

Participant: Márton Karsai.

In this paper we propose a unified framework for classifying and understanding microcanonical RRMs (MRRMs). Focusing on temporal networks, we use this framework to build a taxonomy of MRRMs that proposes a canonical naming convention, classifies them, and deduces their effects on a range of important network features. We furthermore show that certain classes of compatible MRRMs may be applied in sequential composition to generate over a hundred new MRRMs from the existing ones surveyed in this article. We provide two tutorials showing applications of the MRRM framework to empirical temporal networks: 1) to analyze how different features of a network affect other features and 2) to analyze how such features affect a dynamic process in the network. We finally survey applications of MRRMs found in literature. [48]

7.2.3. Reentrant phase transitions in threshold driven contagion on multiplex networks

Participants: Márton Karsai, Samuel Unicomb.

Models of threshold driven contagion explain the cascading spread of information, behavior, systemic risk, and epidemics on social, financial and biological networks. At odds with empirical observation, these models predict that single-layer unweighted networks become resistant to global cascades after reaching sufficient connectivity. We investigate threshold driven contagion on weight heterogeneous multiplex networks and show that they can remain susceptible to global cascades at any level of connectivity, and with increasing edge density pass through alternating phases of stability and instability in the form of reentrant phase transitions of contagion. Our results provide a novel theoretical explanation for the observation of large scale contagion in highly connected but heterogeneous networks. [23]

7.2.4. Interactional and informational attention on Twitter

Twitter may be considered as a decentralized social information processing platform whose users constantly receive their followees' information feeds, which they may in turn dispatch to their followers. This decentralization is not devoid of hierarchy and heterogeneity, both in terms of activity and attention. In particular, we appraise the distribution of attention at the collective and individual level, which exhibits the existence of attentional constraints and focus effects. We observe that most users usually concentrate their attention on a limited core of peers and topics, and discuss the relationship between interactional and informational attention processes – all of which, we suggest, may be useful to refine influence models by enabling the consideration of differential attention likelihood depending on users, their activity levels and peers' positions. [10]

7.2.5. Efficient limited time reachability estimation in temporal networks

Participant: Márton Karsai.

In this paper we propose a probabilistic counting algorithm, which gives simultaneous and precise estimates of the in- and out-reachability (with any chosen waiting-time limit) for every starting event in a temporal network. Our method is scalable allowing measurements for temporal networks with hundreds of millions of events. This opens up the possibility to analyse reachability, spreading processes, and other dynamics in large temporal networks in completely new ways; to compute centralities based on global reachability for all events; or to find with high probability the exact node and time, which could lead to the largest epidemic outbreak. [52]

7.2.6. *weg2vec: Event embedding for temporal networks*

Participant: Márton Karsai.

Network embedding techniques are powerful to capture structural regularities in networks and to identify similarities between their local fabrics. However, conventional network embedding models are developed for static structures, commonly consider nodes only and they are seriously challenged when the network is varying in time. Temporal networks may provide an advantage in the description of real systems, but they code more complex information, which could be effectively represented only by a handful of methods so far. Here, we propose a new method of event embedding of temporal networks, called weg2vec, which builds on temporal and structural similarities of events to learn a low dimensional representation of a temporal network. This projection successfully captures latent structures and similarities between events involving different nodes at different times and provides ways to predict the final outcome of spreading processes unfolding on the temporal structure. [53]

7.3. Communication Networks

Participants: Thomas Begin, Anthony Busson, Isabelle Guérin Lassous, Marion Foare, Philippe Nain, Lafdal Abdelwedoud, Marija Stojanova, Rémy Grünblatt, Juan Pablo Astudillo.

7.3.1. *Quantum communications*

In [29] we investigate the performance of a quantum switch serving a set of users. The function of the switch is to convert bipartite entanglement generated over individual links connecting each user to the switch, into bipartite or tripartite entangled states among (pairs or groups of) users at the highest possible rates at a fixed ratio. Such entanglement can then be converted to quantum-secure shared secret bits among pairs or triples of users using E91-like Quantum Key Distribution (QKD) protocols. The switch can store a certain number of qubits in a quantum memory for a certain length of time, and can make two-qubit Bell-basis measurements or three-qubit GHZ-basis projective measurements on qubits held in the memory. We model a set of randomized switching policies. Discovering that some are better than others, we present analytical results for the case where the switch stores one qubit per user at a given time step, and find that the best policies outperform a time division multiplexing (TDM) policy for sharing the switch between bipartite and tripartite entanglement generation. This performance improvement decreases as the number of users grows. The model is easily augmented to study the capacity region in the presence of qubit decoherence, obtaining similar results. Moreover, decoherence appears to have little effect on capacity. We also study a smaller class of policies when the switch can store two qubits per user.

7.3.2. *Resource Allocation*

In [28] we consider assignment policies that allocate resources to users, where both resources and users are located on a one-dimensional line $[0, \infty\infty)$. First, we consider unidirectional assignment policies that allocate resources only to users located to their left. We propose the Move to Right (MTR) policy, which scans from left to right assigning nearest rightmost available resource to a user, and contrast it to the Unidirectional Gale-Shapley (UGS) matching policy. While both policies among all unidirectional policies minimize the expected distance traveled by a request (request distance), MTR is fairer. Moreover, we show that when user and resource locations are modeled by statistical point processes, and resources are allowed to satisfy more than one user, the spatial system under unidirectional policies can be mapped into bulk service queueing systems, thus allowing the application of many queueing theory results that yield closed-form expressions. As we consider a case where different resources can satisfy different numbers of users, we also generate new results for bulk service queues. We also consider bidirectional policies where there are no directional restrictions on resource allocation and develop an algorithm for computing the optimal assignment which is more efficient than known algorithms in the literature when there are more resources than users. Finally, numerical evaluation of performance of unidirectional and bidirectional allocation schemes yields design guidelines beneficial for resource placement.

7.3.3. VoD broadcasting over vehicular networks

Participants: Thomas Begin, Anthony Busson, Isabelle Guérin Lassous.

We consider a VoD (Video on-Demand) platform designed for vehicles traveling on a highway or other major roadway. Typically, cars or buses would subscribe to this delivery service so that their passengers get access to a catalog of movies and series stored on a back-end server. The network infrastructure comprises IEEE 802.11p RSUs (Road Side Units) that are deployed along the highway and deliver video content to traveling vehicles. In this paper, we propose a simple analytical and yet accurate solution to estimate two key performance parameters for a VoD platform: (i) the average download data rate experienced by vehicles over their journey and (ii) the average "interruption time", which corresponds to the fraction of time the video playback of a given vehicle is interrupted because of an empty buffer. Through multiple examples, we investigate the influence of several parameters (e.g., the video bit rate, the number of vehicles, the distance between RSUs, the vehicle velocity) on these two performance parameters whose outcome may help the sizing of an IEEE 802.11p-based VoD platform [12].

7.3.4. Performance Evaluation of Channel Bonding in IEEE 802.11ac

Participants: Thomas Begin, Anthony Busson, Marija Stojanova.

WLANs grow in popularity in home, public, and work environments, resulting in constantly increasing demands for wireless coverage and capacity. There exist two dominant strategies that help solve the problem of WLAN capacity: the deployment of more APs and enhancement of the standards in use. These policies result in WLANs containing a larger number of more complex devices, making the prediction of the network's behavior an even more elaborate problem. Because of these issues, WLANs are prone to inefficient configurations. In this paper, we propose a Markovian continuous time model that aims at predicting the throughputs achieved by all the WLAN's APs as a function of the network's topology and the AP's throughput demands. By means of simulation, we show that our model achieves mean relative errors of less than 10% for networks of different sizes and with diverse node configurations. The model is adapted to the specificities of the IEEE 802.11ac standard amendment and can be used to solve problems such as channel assignment or channel bonding. We derive guidelines on the best practice in channel bonding given a performance metric and for different MCS indexes, frame aggregation rates, saturation levels, and network topologies. We then put our findings to the test by identifying the optimal channel bonding combination in a WLAN containing a diverse set of nodes.

7.3.5. Distributed Congestion Control mechanism for NANs

Participants: Thomas Begin, Anthony Busson, Juan Pablo Astudillo.

The need for significant improvements in the management and efficient use of electrical energy has led to the evolution from the traditional electrical infrastructures towards modern Smart Grid networks. Taking into account the critical importance of this type of networks, multiple research groups focus their work on issues related to the generation, transport and consumption of electrical energy. One of the key research points is the data communication network associated with the electricity transport infrastructure, and specifically the network that interconnects the devices in consumers' homes, the so-called Neighborhood Area Networks (NANs). In this paper, a new distributed congestion control mechanism is proposed, implemented and evaluated for NANs. Besides, different priorities have been considered for the traffic flows transmitted by different applications. The main goal is to provide with the needed Quality of Service (QoS) to all traffic flows, especially in high traffic load situations. The proposal is evaluated in the context of a wireless ad hoc network made up by a set of smart meter devices, using the Ad hoc On-Demand Distance Vector (AODV) routing protocol and the IEEE 802.11ac physical layer standard. The application of the proposed congestion control mechanism, together with the necessary modifications made to the AODV protocol, lead to performance improvements in terms of packet delivery ratio, network throughput and transit time, fairness between different traffic sources and QoS provision [35].

7.3.6. Simulation and Performance Evaluation of the Intel Rate Adaptation Algorithm

Participants: Rémy Grünblatt, Isabelle Guérin-Lassous.

With the rise of the complexity of the IEEE 802.11 standard, rate adaptation algorithms have to deal with a large set of values for all the different parameters having an impact on the network throughput. Simple trial-and-error algorithms can no longer explore solution space in reasonable time and smart solutions are required. Most of the WiFi controllers rely on proprietary code and the used rate adaptation algorithms in these controllers are unknown. Very few WiFi controllers expose their rate adaptation algorithms if they do not rely on the MINSTREL-HT algorithm which is implemented in the mac80211 component of the Linux kernel. Intel WiFi controllers come with their own rate adaptation algorithms that are implemented in the Intel IWLWIFI Linux Driver which is open-source.

In this work, we have reverse-engineered the Intel rate adaptation mechanism from the source code of the IWLWIFI Linux driver, and we give, in a comprehensive form, the underlying rate adaptation algorithm named IWL-MVM-RS. We describe the different mechanisms used to seek the best throughput adapted to the network conditions. We have also implemented the IWL-MVM-RS algorithm in the NS-3 simulator. Thanks to this implementation, we can evaluate the performance of IWL-MVM-RS in different scenarios (static and with mobility, with and without fast fading). We also compare the performances of IWL-MVM-RS with the ones of MINSTREL-HT and IDEALWIFI, also implemented in the NS-3 simulator [26], [32].

7.3.7. A Passive Method to Infer the Weighted Conflict Graph of a IEEE 802.11 Network

Participants: Lafdal Abdelwedoud, Anthony Busson, Isabelle Guérin-Lassous, Marion Foare.

Wi-Fi networks often consist of several Access Points (APs) to form an Extended Service Set. These APs may interfere with each other as soon as they use the same channel or overlapping channels. A classical model to describe interference is the conflict graph. As the interference level varies in the network and in time, we consider a weighted conflict graph. In this work, we propose a method to infer the weights of the conflict graph of a Wi-Fi network.

Weights represent the proportion of activity from a neighbor detected by the Clear Channel Assessment mechanism. Our method relies on a theoretical model based on Markov networks applied to a decomposition of the original conflict graph. The input of our solution is the activity measured at each AP, measurements available in practice. The proposed method is validated through ns-3 simulations performed for different scenarios. Results show that our solution is able to accurately estimate the weights of the conflict graph. [24], [34].

DIANA Project-Team

6. New Results

6.1. Service Transparency

6.1.1. From Network Traffic Measurements to QoE for Internet Video

Participants: Muhammad Jawad Khokhar, Thibaut Ehlinger, Chadi Barakat.

Video streaming is a dominant contributor to the global Internet traffic. Consequently, monitoring video streaming Quality of Experience (QoE) is of paramount importance to network providers. Monitoring QoE of video is a challenge as most of the video traffic of today is encrypted. In this work, we consider this challenge and present an approach based on controlled experimentation and machine learning to estimate QoE from encrypted video traces using network level measurements only. We consider a case of YouTube and play out a wide range of videos under realistic network conditions to build ML models (classification and regression) that predict the subjective MOS (Mean Opinion Score) based on the ITU P.1203 model along with the QoE metrics of startup delay, quality (spatial resolution) of playout and quality variations, and this is using only the underlying network Quality of Service (QoS) features. We comprehensively evaluate our approach with different sets of input network features and output QoE metrics. Overall, our classification models predict the QoE metrics and the ITU MOS with an accuracy of 63-90% while the regression models show low error; the ITU MOS (1-5) and the startup delay (in seconds) are predicted with a root mean square error of 0.33 and 2.66 respectively. The results of this work were published in [26] and can be found with further details in the PhD manuscript of Muhammad Jawad Khokhar graduated in October 2019.

6.1.2. When Deep Learning meets Web Measurements to infer Network Performance

Participants: Imane Taibi, Chadi Barakat.

Web browsing remains one of the dominant applications of the internet, so inferring network performance becomes crucial for both users and providers (access and content) so as to be able to identify the root cause of any service degradation. Recent works have proposed several network troubleshooting tools, e.g, NDT, MobiPerf, SpeedTest, Fathom. Yet, these tools are either computationally expensive, less generic or greedy in terms of data consumption. The main purpose of this work funded by the IPL BetterNet is to leverage passive measurements freely available in the browser and machine learning techniques (ML) to infer network performance (e.g., delay, bandwidth and loss rate) without the addition of new measurement overhead. To enable this inference , we propose a framework based on extensive controlled experiments where network configurations are artificially varied and the Web is browsed, then ML is applied to build models that estimate the underlying network performance. In particular, we contrast classical ML techniques (such as random forest) to deep learning models trained using fully connected neural networks and convolutional neural networks (CNN). Results of our experiments show that neural networks have a higher accuracy compared to classical ML approaches. Furthermore, the model accuracy improves considerably using CNN. These results were published in [28].

6.1.3. On Accounting for Screen Resolution in Adaptive Video Streaming: A QoE-Driven Bandwidth Sharing Framework

Participants: Othmane Belmoukadam, Muhammad Jawad Khokhar, Chadi Barakat.

Screen resolution along with network conditions are main objective factors impacting the user experience, in particular for video streaming applications. Terminals on their side feature more and more advanced characteristics resulting in different network requirements for good visual experience. Previous studies tried to link MOS (Mean Opinion Score) to video bit rate for different screen types (e.g., CIF, QCIF, and HD). We leverage such studies and formulate a QoE-driven resource allocation problem to pinpoint the optimal bandwidth allocation that maximizes the QoE (Quality of Experience) over all users of a provider located behind the same bottleneck link, while accounting for the characteristics of the screens they use for video playout. For our optimization problem, QoE functions are built using curve fitting on data sets capturing the relationship between MOS, screen characteristics, and bandwidth requirements. We propose a simple heuristic based on Lagrangian relaxation and KKT (Karush Kuhn Tucker) conditions for a subset of constraints. Numerical simulations show that the proposed heuristic is able to increase overall QoE up to 20% compared to an allocation with TCP look-alike strategies implementing max-min fairness. Later, we use a MPEG/DASH implementation in the context of ns-3 and show that coupling our approach with a rate adaptation algorithm can help increasing QoE while reducing both resolution switches and number of interruptions. Our framework and the first validation results were published in [20].

6.1.4. Tuning optimal traffic measurement parameters in virtual networks with machine learning

Participants: Karyna Gogunska, Chadi Barakat.

With the increasing popularity of cloud networking and the widespread usage of virtualization as a way to offer flexible and virtual network and computing resources, it becomes more and more complex to monitor this new virtual environment. Yet, monitoring remains crucial for network troubleshooting and analysis. Controlling the measurement footprint in the virtual network is one of the main priorities in the process of monitoring as resources are shared between the compute nodes of tenants and the measurement process itself. In this paper, first, we assess the capability of machine learning to predict measurement impact on the ongoing traffic between virtual machines; second, we propose a data-driven solution that is able to provide optimal monitoring parameters for virtual network measurement with minimum traffic interference. These results were published in [25] and are part of the PhD manuscript of Karyna Gogunska graduated in December 2019.

6.1.5. Collaborative Traffic Measurement in Virtualized Data Center Networks

Participants: Houssam Elbouanani, Chadi Barakat.

Data center network monitoring can be carried out at hardware networking equipment (e.g. physical routers) and/or software networking equipment (e.g. virtual switches). While software switches offer high flexibility to deploy various monitoring tools, they have to utilize server resources, esp. CPU and memory, that can no longer be reserved fully to service users' traffic. In this work we closely examine the costs of (i) sampling packets ; (ii) sending them to a user-space program for measurement; and (iii) forwarding them to a remote server where they will be processed in case of lack of resources locally. Starting from empirical observations, we derive an analytical model to accurately predict ($R^2 = 99.5\%$) the three aforementioned costs, as a function of the sampling rate. We next introduce a collaborative approach for traffic monitoring and sampling that maximizes the amount of collected traffic without impacting the data center's operation. We analyze, through numerical simulations, the performance of our collaborative solution. The results show that it is able to take advantage of the uneven loads on the servers to maximize the amount of traffic that can be sampled at the scale of a data center. The resulting gain can reach 200% compared to a non collaborative approach. These results were published in [23].

6.1.6. Distributed Privacy Preserving Platform for Ridesharing Services

Participants: Damien Saucez, Yevhenii Semenko.

The sharing economy fundamentally changed business and social interactions. Interestingly, while in essence this form of collaborative economy allows people to directly interact with each other, it is also at the source of the advent of eminently centralized platforms and marketplaces, such as Uber and Airbnb. One may be concerned with the risk of giving the control of a market to a handful of actors that may unilaterally fix their own rules and threaten privacy. Within the Data Privacy project of the UCAJedi Idex Academy 5 and House of Human and Social Sciences, Technologies and Uses Theme, we have proposed a holistic solution to address privacy issues in the sharing economy. We considered the case of ridesharing and proposed a decentralized architecture which gives the opportunity to shift from centralized platforms to decentralized ones. Digital communications in our proposition are specifically designed to preserve data privacy and avoid any form of centralization. A blockchain is used in our proposition to guarantee the essential roles of a marketplace, but in a decentralized way. Our evaluation shows that privacy protection without trusted entities comes at the cost of harder scalability than an approach with a trusted third party. However, our numerical evaluation on real data and our Android prototype shows the practical feasibility of our approach. The results obtained in this activity are published in 12th International Conference on Security, Privacy, and Anonymity in Computation, Communication, and Storage (SpaCCS) 2019, Atlanta [31] and documented in a research report [35].

6.1.7. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels

Participants: Imane Fouad, Arnaud Legout, Natasa Sarafijanovic-Djukic.
Web tracking has been extensively studied over the last decade. To detect tracking, previous studies and user tools rely on filter lists. However, it has been shown that filter lists miss trackers. In this paper, we propose an alternative method to detect trackers inspired by analyzing behavior of invisible pixels. By crawling 84,658 webpages from 8,744 domains, we detect that third-party invisible pixels are widely deployed: they are present on more than 94.51% of domains and constitute 35.66% of all third-party images. We propose a fine-grained behavioral classification of tracking based on the analysis of invisible pixels. We use this classification to detect new categories of tracking and uncover new collaborations between domains on the full dataset of 4,216,454 third-party requests. We demonstrate that two popular methods to detect tracking, based on EasyList & EasyPrivacy and on Disconnect lists respectively miss 25.22% and 30.34% of the trackers that we detect. Moreover, we find that if we combine all three lists, 379,245 requests originated from 8,744 domains still track users on 68.70% of websites. This work will appear in PETS 2020 [24].

6.1.8. Privacy implications of switching ON a light bulb in the IoT world

Participants: Mathieu Thiery, Arnaud Legout.
The number of connected devices is increasing every day, creating smart homes and shaping the era of the Internet of Things (IoT), and most of the time, end-users are unaware of their impacts on privacy. In this work, we analyze the ecosystem around a Philips Hue smart white bulb in order to assess the privacy risks associated to the use of different devices (smart speaker or button) and smartphone applications to control it. We show that using different techniques to switch ON or OFF this bulb has significant consequences regarding the actors involved (who mechanically gather information on the user's home) and the volume of data sent to the Internet (we measured differences up to a factor 100, depending on the control technique we used). Even when the user is at home, these data flows often leave the user's country, creating a situation that is neither privacy friendly (and the user is most of the time ignorant of the situation), nor sovereign (the user depends on foreign actors), nor sustainable (the extra energetic consumption is far from negligible). We therefore advocate a complete change of approach, that favors local communications whenever sufficient. The preprint documenting this work has been published as research report [40].

6.1.9. ElectroSmart

Participants: Arnaud Legout, Mondi Ravi, David Migliacci, Abdelhakim Akodadi, Yanis Boussad.
We are currently evaluating the relevance to create a startup for the ElectroSmart project. We are quite advanced in the process and the planned creation is June 2020. There is a "contrat de transfer" ready between Inria and ElectroSmart to transfer the PI from Inria to the ElectroSmart company (when it will be created). Arnaud Legout the future CEO of the company obtained the "autorisation de création d'entreprise" from Inria. ElectroSmart has been incubated in PACA Est in December 2018.

The three future co-founder of ElectroSmart (Arnaud Legout, Mondi Ravi, David Migliacci) followed the Digital Startup training from Inria/EM Lyon.

The goal of ElectroSmart is to help people reduce their exposure to EMF and offer a solution to reduce symptoms associated with exposure to EMF. Electrosensitivity, is known to be a complex and multifactorial syndrome that impacts hundreds of millions of persons worldwide. We aim to commercialize the first treatment of electrosensitivity based on non-deceptive placebo (called open-label placebo). It is known today that placebo are an effective treatment to subjective symptoms (which is the case for several symptoms associated with electrosensitivity). The problem with placebo was that is was assumed that it must be deceptive to be efficient. Kaptchuk et al. showed recently that non-deceptive placebo are as effective as deceptive placebo, so the ethical usage of placebo is now possible. ElectroSmart want to be the first company to commercialize non-deceptive placebo for electrosensitive persons. For details, see https://electrosmart.app/.

6.2. Open Network Architecture

6.2.1. Constrained Software Defined Networks

Participant: Damien Saucez.

The objective of the ANR JCJC DET4ALL project was to offer the ability to multiplex constrained networks with real time and safety requirements on Ethernet network not initially thought for strict constraints. The reason for this move to Ethernet is to reduce the cost of networking solutions in automotive and industrial applications. We advocate that this move requires to rely on Software Defined Networking (SDN) that enables a programmatic approach to networking, hence offering modularity and flexibility. The challenge with SDN is to be able to certify the behaviour of the system while keeping the solution generic. Within DET4ALL we put the first element in place to show that the previous works that proposed programming languages and abstractions for best-effort network could be leveraged to offering safety properties and determinism in real-time industrial and automotive networks. More precisely, we have demonstrated that Linear Temporal Logic (LTL) can be used in real-time networks to demonstrate the that real-time constraints are always respected. We built a strawman to show that the Temporal NetKat language was adapted to express real-time constraints of networks even though it was not initially design for that purpose. Given that Temporal NetKat relies on LTL and an algebra, it is a good candidate to prove the correct behaviour of a SDN network which logic would be implemented with such a language. In the continuation of this work, we have determined what would be necessary to be able to provide provable live network updates in real time network without service degradation. This work is published in [30] and will be detailed in the next subsection. Due the leave of Damien Saucez to Safran for one year starting October 1st 2019, the activity on this project had to be stopped as it was in the context of an ANR JCJC project.

6.2.2. NUTS: Network Updates in Real Time Systems

Participants: Damien Saucez, Walid Dabbous.

Recent manufacturing trends have highlighted the need to adapt to volatile, fast-moving, and customer-driven markets. To keep pace with ever quicker product lifecycles, shorter order lead times and growing product variants, factories will become distributed modular cyber-physical systems interconnected by complex communication networks. We advocate that the Software Define Networking (SDN) concept with its programmatic approach to networking is a key enabler for the so-called Industry 4.0 because it provides flexibility and the possibility to formally reason on networks. We have identified that a critical point to address is how to support safe network updates of deterministic real-time communication SDN. To achieve this goal 4 elements are required. First a declarative language with LTL support is needed to express the constraints. Second, a programmable data-plane with the ability to provide real-time constraints indications must be provided in order to assess the behaviour of the forwarding elements. Such language does not exist yet however among the data-plane languages currently on the market some provide the ability to add annotations that could be used to reach our objective. Third, we have identified that deterministic algorithms had to be used to provide a verifiable sequence of network updates in order to make live updates without service degradations. Finally, mathematical techniques must be used to provide bounds on the network updates. Network Calculus can be used for that objective. This study was published as a poster in SOSR'19 [30].

6.2.3. A Joint range extension and localization for LPWAN

Participants: Mohamed Naoufal Mahfoudi, Gayatri Sivadoss, Othmane Bensouda Korachi, Thierry Turletti, Walid Dabbous.

We have proposed Snipe, a novel system offering joint localization and range extensions for LPWANs. Although LPWAN systems such as Long Range (LoRa) are designed to achieve high communication range with low energy consumption, they suffer from fading in obstructed environments with dense multipath components, and their localization system is sub-par in terms of accuracy. In this work, MIMO techniques are leveraged to achieve a higher signal-to-noise ratio at both the end device and the gateway while providing an opportunistic accurate radar-based system for localization with limited additional cost. This work has been published at Internet Technology Letters [15].

6.2.4. Online Robust Placement of Service Chains for Large Data Center Topologies

Participants: Ghada Moualla, Thierry Turletti, Damien Saucez.

The trend today is to deploy applications and more generally Service Function Chains (SFCs) in public clouds. However, before being deployed in the cloud, chains were deployed on dedicated infrastructures where software, hardware, and network components were managed by the same entity, making it straightforward to provide robustness guarantees. By moving their services to the cloud, the users lose their control on the infrastructure and hence on the robustness. We propose an online algorithm for robust placement of service chains in data centers. Our placement algorithm determines the required number of replicas for each function of the chain and their placement in the data center. Our simulations on large data-center topologies with up to 30,528 nodes show that our algorithm is fast enough such that one can consider robust chain placements in real time even in a very large data center and without the need of prior knowledge on the demand distribution. This work has been published at IEEE Access [16].

6.2.5. Bandwidth-optimal Failure Recovery Scheme for Robust Programmable Networks

Participants: Giuseppe Di Lena, Damien Saucez, Thierry Turletti.

With the emergence of Network Function Virtualization (NFV) and Software Defined Networking (SDN), efficient network algorithms considered too hard to be put in practice in the past now have a second chance to be considered again. In this context, we rethink the network dimensioning problem with protection against Shared Risk Link Group (SLRG) failures. In this work, we consider a path-based protection scheme with a global rerouting strategy, in which, for each failure situation, there may be a new routing of all the demands. Our optimization task is to minimize the needed amount of bandwidth. After discussing the hardness of the problem, we develop a scalable mathematical model that we handle using the Column Generation technique. Through extensive simulations on real-world IP network topologies and on random generated instances, we show the effectiveness of our method. Finally, our implementation in OpenDaylight demonstrates the feasibility of the approach and its evaluation with Mininet shows that technical implementation choices may have a dramatic impact on the time needed to reestablish the flows after a failure takes place. This work has been presented at the IEEE International Conference on Cloud Networking (CloudNet), November 2019, at Coimbra in Portugal [29] and documented in a research report [36]. A poster version is published in IFIP-Networking in Warsaw [41].

6.2.6. Efficient Pull-based Mobile Video Streaming leveraging In-Network Functions

Participants: Indukala Naladala, Thierry Turletti.

There has been a considerable increase in the demand for high quality mobile video streaming services, while at the same time, the video traffic volume is expected to grow exponentially. Consequently, maintaining high quality of experience (QoE) and saving network resources are becoming crucial challenges to solve. In this work, we propose a name-based mobile streaming scheme that allows efficient video content delivery by exploiting a smart pulling mechanism designed for information-centric networks (ICNs). The proposed mechanism enables fast packet loss recovery by leveraging in-network caching and coding. Through an experimental evaluation of our mechanism over an open wireless testbed and the Internet, we demonstrate that the proposed scheme leads to higher QoE levels than classical ICN and TCP-based streaming mechanisms.

This work will be presented at the IEEE Consumer Communications & Networking Conference (CCNC), in January 2020 at Las Vegas, USA [27]. The following link https://github.com/fit-r2lab/demo-cefore includes the artefacts that allows to reproduce performance results shown in the paper.

6.2.7. Low Cost Video Streaming through Mobile Edge Caching: Modelling and Optimization

Participants: Luigi Vigneri, Chadi Barakat.

Caching content at the edge of mobile networks is considered as a promising way to deal with the data tsunami. In addition to caching at fixed base stations or user devices, it has been recently proposed that an architecture with public or private transportation acting as mobile relays and caches might be a promising middle ground. While such mobile caches have mostly been considered in the context of delay tolerant networks, in this work done in collaboration with Eurecom with the support of the UCN@Sophia Labex, we argue that they could be used for low cost video streaming without the need to impose any delay on the user. Users can prefetch video chunks into their playout buffer from encountered vehicle caches (at low cost) or stream from the cellular infrastructure (at higher cost) when their playout buffer empties while watching the content. Our main contributions are: (i) to model the playout buffer in the user device and analyze its idle periods which correspond to bytes downloaded from the infrastructure; (ii) to optimize the content allocation to mobile caches, to minimize the expected number of non-offloaded bytes. We perform trace-based simulations to support our findings showing that up to 60 percent of the original traffic could be offloaded from the main infrastructure. These contributions were published in the IEEE Transactions on Mobile Computing journal [18].

6.2.8. Quality of Experience-Aware Mobile Edge Caching through a Vehicular Cloud

Participants: Luigi Vigneri, Chadi Barakat.

Densification through small cells and caching in base stations have been proposed to deal with the increasing demand for Internet content and the related overload on the cellular infrastructure. However, these solutions are expensive to install and maintain. Instead, using vehicles acting as mobile caches might represent an interesting alternative. In this work, we assume that users can query nearby vehicles for some time, and be redirected to the cellular infrastructure when the deadline expires. Beyond reducing costs, in such an architecture, through vehicle mobility, a user sees a much larger variety of locally accessible content within only few minutes. Unlike most of the related works on delay tolerant access, we consider the impact on the user experience by assigning different retrieval deadlines per content. We provide the following contributions: (i) we model analytically such a scenario; (ii) we formulate an optimization problem to maximize the traffic offloaded while ensuring user experience guarantees; (iii) we propose two variable deadline policies; (iv) we perform realistic trace-based simulations, and we show that, even with low technology penetration rate, more than 60% of the total traffic can be offloaded which is around 20% larger compared to existing allocation policies. These results were published in the IEEE Transactions on Mobile Computing journal [19].

6.2.9. Machine Learning for Next-Generation Intelligent Transportation Systems

Participants: Tingting Yuan, Thierry Turletti, Chadi Barakat.

Intelligent Transportation Systems, or ITS for short, includes a variety of services and applications such as road traffic management, traveler information systems, public transit system management , and autonomous vehicles, to name a few. It is expected that ITS will be an integral part of urban planning and future cities as it will contribute to improved road and traffic safety, transportation and transit efficiency, as well as to increased energy efficiency and reduced environmental pollution. On the other hand, ITS poses a variety of challenges due to its scalability and diverse quality-of-service needs, as well as the massive amounts of data it will generate. In this survey, we explore the use of Machine Learning (ML), which has recently gained significant traction, to enable ITS. In the context of the Drive associated team, we did a comprehensive survey of the current state-of-the-art of how ML technology has been applied to a broad range of ITS applications and services, such as cooperative driving and road hazard warning, and identify future directions for how ITS can use and benefit from ML technology. The survey is documented in [42].

6.3. Experimental Evaluation

6.3.1. Exploiting the cloud for Mininet performance

Participants: Giuseppe Di Lena, Damien Saucez, Thierry Turletti.

Networks have become complex systems that combine various concepts, techniques, and technologies. As a consequence , modelling or simulating them is now extremely complicated and researchers massively resort to prototyping techniques. Among other tools, Mininet is the most popular when it comes to evaluate SDN propositions. It allows to emulate SDN networks on a single computer. However, under certain circumstances experiments (e.g., resource intensive ones) may overload the host running Mininet. To tackle this issue, we propose Distrinet, a way to distribute Mininet over multiple hosts. Distrinet uses the same API than Mininet, meaning that it is compatible with Mininet programs. Distrinet is generic and can deploy experiments in Linux clusters or in the Amazon EC2 cloud. Thanks to optimization techniques, Distrinet minimizes the number of hosts required to perform an experiment given the capabilities of the hosting infrastructure, meaning that the experiment is run in a single host (as Mininet) if possible. Otherwise, it is automatically deployed on a platform using a minimum amount of resources in a Linux cluster or with a minimum cost in Amazon EC2. This work has been presented at the IEEE International Conference on Cloud Networking (CloudNet) [22]. Distrinet has been demonstrated both at the IEEE CloudNet conference and at the ACM CoNEXT conference in Orlando USA in December 2019 [39].

6.3.2. Distributed Network Experiment Emulation

Participants: Giuseppe Di Lena, Damien Saucez, Thierry Turletti, Walid Dabbous.

With the ever growing complexity of networks, researchers have to rely on test-beds to be able to fully assess the quality of their propositions. In the meanwhile, Mininet offers a simple yet powerful API, the goldilocks of network emulators. We advocate that the Mininet API is the right level of abstraction for network experiments. Unfortunately it is designed to be run on a single machine. To address this issue we developed a distributed version of Mininet-Distrinet-that can be used to perform network experiments in any Linux-based testbeds, either public or private. To properly use testbed resources and avoid over-commitment that would lead to inaccurate results, Distrinet uses optimization techniques that determine how to orchestrate the experiments within the testbed. Its programmatic approach, its ability to work on various testbeds, and its optimal management of resources make Distrinet a key element to reproducible research. This work has been presented at the Global Experimentation for Future Internet - Workshop (GeFi) workshop November 2019, at Coimbra in Portugal [38].

6.3.3. Evaluating smartphone performance for cellular power measurement. Under submission

Participants: Yanis Boussad, Arnaud Legout.

From crowdsource data collection to automation and robotics, mobile smartphones are well suited for various use cases given the rich hardware components they feature. Researchers can now have access to various sensors such as barometers, magnetometers, orientation sensors, in addition to multiple wireless technologies all on a single and relatively cheap mobile smartphone. In this work, we study the performance of smartphones to measure cellular wireless power. We performed our experiments inside an anechoic chamber in order to compare the measurements of smartphone to the ones obtained with professional spectrum analyzer. We first evaluate the effect of orientation on the received power, then we propose a way to improve the accuracy of smartphone power measurements by using the orientation sensors. We improve the accuracy of the measurements from 25 dBm RMSE to no more than 6 dBm RMSE. We also show how we can exploit the characteristics of the reception pattern of the smartphone to determine the angle of arrival of the signal. The results of this work are described in a research report under submission [32].

6.3.4. Towards Reproducible Wireless Experiments Using R2lab

Participants: Mohamed Naoufal Mahfoudi, Thierry Parmentelat, Thierry Turletti, Walid Dabbous.

Reproducibility is key in designing wireless systems and evaluating their performance. Trying to reproduce wireless experiments allowed us to identify some pitfalls and possible ways to simplify the complex task of avoiding them. In this research report, we expose a few considerations that we learned are instrumental for ensuring the reproducibility of wireless experiments. Then we describe the steps we have taken to make our experiments easy to reproduce. We specifically address issues related to wireless hardware, as well as varying propagation channel conditions. We show that extensive knowledge of the used hardware and of its design is required to guarantee that the inner state of the system has no negative impact on performance evaluation and experimental results. As for variability of channel conditions, we make the case that a special setup or testbed is necessary so that one can control the ambient wireless propagation environment, using for instance, an anechoic chamber like R2lab. This work is published as research report [33].

6.3.5. A step towards runnable papers using R2lab

Participants: Thierry Parmentelat, Mohamed Naoufal Mahfoudi, Thierry Turletti, Walid Dabbous.
In this research report, we present R2lab, an open, electromagnetically insulated research testbed dedicated to wireless networking. We describe the hardware capabilities currently available in terms of Software Defined Radio, and the software suite made available to deploy experiments. Using a generic experiment example, we show how it all fits into a notebook-based approach to getting closer to runnable papers. This work is published as research report [34].

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation

Participants: Gerardo Rubino, Bruno Sericola.

Fluid Queues. Stochastic fluid flow models and, in particular, those driven by Markov chains, have been intensively studied in the last two decades. Not only they have been proven to be efficient tools to mimic Internet traffic flows at a macroscopic level but they are useful tools in many areas of applications such as manufacturing systems or in actuarial sciences, to cite but a few. We propose in [61] a chapter which focus on such a model in the context of performance analysis of a potentially congested system. The latter is modeled by means of a finite-capacity system whose content is described by a Markov driven stable fluid flow. We describe step-by-step a methodology to compute exactly the loss probability of the system. Our approach is based on the computation of hitting probabilities jointly with the peak level reached during a busy period, both in the infinite and finite buffer case. Accordingly we end up with differential Riccati equations that can be solved numerically. Moreover, we are able to characterize the complete distribution of both the duration of congestion and of the total information lost during such a busy period.

Connecting irreducible and absorbing Markov chains. Irreducible Markov chains in continuous time are the basic tool for instance in performance evaluation (typically, a queuing model), where in a large majority of cases, we are interested in the behavior of the modeled system in steady-state. Most metrics used are based on the stationary distribution of the model, under unicity natural conditions. Absorbing Markov chains, also in continuous time, play the equivalent role in dependability evaluation, because realistic models must have a finite lifetime, which corresponds here to the absorption time of the chain. In this case, the object of interest is this lifetime, steady-state gives no useful information about the system, and most of the used metrics are defined based on that object. In [30] with describe different connections between the two worlds together with some consequences of those relations in both areas, that is, both in performance and in dependability.

Transient analysis of Markov queueing models. Analyzing the transient behavior of a queueing system is much harder than studying its steady state, the difference being basically that of moving from a linear system to a linear differential system. However, a huge amount of efforts has been put on the former problem, from all kinds of points of view: trials to find closed-forms of the main state distributions, algorithms for numerical evaluations, approximations of different types, exploration of other transient metrics than the basic state distributions, etc. In [62] we focus on the first two elements, the derivation of closed-forms for the main transient state distributions, and the development of numerical techniques. The chapter is organized as a survey, and the main goal is to position and to underline the role of the uniformization technique, for both finding closed-forms and for developing efficient numerical evaluation procedures. In some cases, we extend the discussion to other related transient metrics that are relevant for applications.

7.2. Distributed Systems

Participants: Hamza Ben Ammar, Yann Busnel, Yassine Hadjadj-Aoul, Yves Mocquard, Frédérique Robin, Bruno Sericola.

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

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies.

Membership management is a classic and fundamental problem in many use cases. In networking for instance, it is useful to check if a given IP address belongs to a black list or not, in order to allow access to a given server. This has also become a key issue in very large-scale distributed systems, or in massive databases. Formally, from a subset belonging to a very large universe, the problem consists in answering the question "Given any element of the universe, does it belong to a given subset?". Since the access of a perfect oracle answering the question is commonly admitted to be very costly, it is necessary to provide efficient and inexpensive techniques in the context where the elements arrive continuously in a data stream (for example, in network metrology, log analysis, continuous queries in massive databases, etc.). In [36], we propose a simple but efficient solution to answer membership queries based on a couple of Bloom filters. In a nutshell, the idea is to contact the oracle only if an item is seen for the first time. We use a classical Bloom filter to remember an item occurrence. For the next occurrences, we answer the membership query using a second Bloom filter, which is dynamically populated only when the database is queried. We provide theoretical bounds on the false positive and negative probabilities and we illustrate through extensive simulations the efficiency of our solution, in comparison with standard solutions such as classic Bloom filters.

Shuffle grouping is a technique used by stream processing frameworks to share input load among parallel instances of stateless operators. With shuffle grouping each tuple of a stream can be assigned to any available operator instance, independently from any previous assignment. A common approach to implement shuffle grouping is to adopt a Round-Robin policy, a simple solution that fares well as long as the tuple execution time is almost the same for all the tuples. However, such an assumption rarely holds in real cases where execution time strongly depends on tuple content. As a consequence, parallel stateless operators within stream processing applications may experience unpredictable unbalance that, in the end, causes undesirable increase in tuple completion times. We consider recently an application to continuous queries, which are processed by a stream processing engine (SPE) to generate timely results given the ephemeral input data. Variations of input data streams, in terms of both volume and distribution of values, have a large impact on computational resource requirements. Dynamic and Automatic Balanced Scaling for Storm (DABS-Storm) [21] is an original solution for handling dynamic adaptation of continuous queries processing according to evolution of input stream properties, while controlling the system stability. Both fluctuations in data volume and distribution of values within data streams are handled by DABS-Storm to adjust the resources usage that best meets processing needs. To achieve this goal, the DABS-Storm holistic approach combines a proactive auto-parallelization algorithm with a latency-aware load balancing strategy.

Sampling techniques constitute a classical method for detection in large-scale data streams. We have proposed a new algorithm that detects on the fly the k most frequent items in the sliding window model [52]. This algorithm is distributed among the nodes of the system. It is inspired by a recent approach, which consists in associating a stochastic value correlated with the item's frequency instead of trying to estimate its number of occurrences. This stochastic value corresponds to the number of consecutive heads in coin flipping until the first tail occurs. The original approach was to retain just the maximum of consecutive heads obtained by an item, since an item that often occurs will have a higher probability of having a high value. While effective for very skewed data distributions, the correlation is not tight enough to robustly distinguish items with comparable frequencies. To address this important issue, we propose to combine the stochastic approach with a deterministic counting of items. Specifically, in place of keeping the maximum number of consecutive heads obtained by an item, we count the number of times the coin flipping process of an item has exceeded a given threshold. This threshold is defined by combining theoretical results in leader election and coupon collector problems. Results on simulated data show how impressive is the detection of the top-k items in a large range of distributions.

Health Big Data Analysis. The aim of the study was to build a proof-of-concept demonstrating that big data technology could improve drug safety monitoring in a hospital and could help pharmacovigilance professionals to make data-driven targeted hypotheses on adverse drug events (ADEs) due to drug-drug interactions (DDI). In [17], we developed a DDI automatic detection system based on treatment data and laboratory tests from the electronic health records stored in the clinical data warehouse of Rennes academic hospital. We also used OrientDb, a graph database to store informations from five drug knowledge databases and Spark to perform analysis of potential interactions between drugs taken by hospitalized patients. Then, we developed a Machine Learning model to identify the patients in whom an ADE might have occurred because of a DDI. The DDI detection system worked efficiently and the computation time was manageable. The system could be routinely employed for monitoring.

Probabilistic analysis of population protocols. 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. In [23] we 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 the paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant, where at random times, nodes successively interact by pairs exchanging their information on the rumor. In a previous paper, we performed a study of the total number of interactions needed for all the nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow us to compare different protocols, we provided a thorough analysis of the distribution of this total number of interactions together with its asymptotic behavior. In this paper we extend this discrete time analysis by solving a conjecture proposed previously and we consider the continuous time case, where a Poisson process is associated to each node to determine the instants at which interactions occur. The rumor spreading time is thus more realistic since it is the real time needed for all the nodes of the network to discover the rumor. Once again, as most of the existing results involve huge constants, we provide tight bound and equivalent of the complementary distribution of the rumor spreading time. We also give the exact asymptotic behavior of the complementary distribution of the rumor spreading time around its expected value when the number of nodes tends to infinity.

Among the different problems addressed in the model of population protocols, average-based problems have been studied for the last few years. In these problems, agents start independently from each other with an initial integer state, and at each interaction with another agent, keep the average of their states as their new state. In [45] and [63], using a well chosen stochastic coupling, we considerably improve upon existing results by providing explicit and tight bounds of the time required to converge to the solution of these problems. We apply these general results to the proportion problem, which consists for each agent to compute the proportion of agents that initially started in one predetermined state, and to the counting population size problem, which aims at estimating the size of the system. Both protocols are uniform, i.e., each agent's local algorithm for computing the outputs, given the inputs, does not require the knowledge of the number of agents. Numerical simulations illustrate our bounds of the convergence time, and show that these bounds are tight in the sense that among extensive simulations, numerous ones fit very well with our bounds.

Organizing both transactions and blocks in a distributed ledger. We propose in [53] a new way to organize both transactions and blocks in a distributed ledger to address the performance issues of permissionless ledgers. In contrast to most of the existing solutions in which the ledger is a chain of blocks extracted from a tree or a graph of chains, we present a distributed ledger whose structure is a balanced directed acyclic graph of blocks. We call this specific graph a SYC-DAG. We show that a SYC-DAG allows us to keep all the remarkable properties of the Bitcoin blockchain in terms of security, immutability, and transparency, while enjoying higher throughput and self-adaptivity to transactions demand. To the best of our knowledge, such a design has never been proposed.

Performance of caching systems. Several studies have focused on improving the performance of caching systems in the context of Content-Centric Networking (CCN). In [16], we propose a fairly generic model of caching systems that can be adapted very easily to represent different caching strategies, even the most advanced ones. Indeed, the proposed model of a single cache, named MACS, which stands for Markov chain-based Approximation of CCN Caching Systems, can be extended to represent an interconnection of caches under different schemes. In order to demonstrate the accuracy of our model, we proposed to derive models of the two most effective techniques in the literature, namely LCD and LRU-K, which may adapt to changing patterns of access.

One of the most important concerns when dealing with the performance of caching systems is the static or dynamic (on-demand) placement of caching resources. This issue is becoming particularly important with the upcoming advent of 5G. In [33] we propose a new technique exploiting the model previously proposed model in [16], in order to achieve the best trade-off between the centralization of resources and their distribution, through an efficient placement of caching resources. To do so, we model the cache resources allocation problem as a multi-objective optimization problem, which is solved using Greedy Randomized Adaptive Search Procedures (GRASP). The obtained results confirm the quality of the outcomes compared to an exhaustive search method and show how a cache allocation solution depends on the network's parameters and on the performance metrics that we want to optimize.

7.3. Machine learning

Participants: Yassine Hadjadj-Aoul, Corentin Hardy, Quang Pham Tran Anh, Gerardo Rubino, Bruno Sericola, Imane Taibi, César Viho.

Distributed deep learning on edge-devices. A recently celebrated type of deep neural network is the Generative Adversarial Network (GAN). GANs are generators of samples from a distribution that has been learned; they are up to now centrally trained from local data on a single location. We question in [37] the performance of training GANs using a spread dataset over a set of distributed machines, following a gossip approach shown to work on standard neural networks. This performance is compared to the federated learning distributed method, that has the drawback of sending model data to a server. We also propose a gossip variant, where GAN components are gossiped independently. Experiments are conducted with Tensorflow with up to 100 emulated machines, on the canonical MNIST dataset. The position of the paper is to provide a first evidence that gossip performances for GAN training are close to the ones of federated learning, while operating in a fully decentralized setup. Second, to highlight that for GANs, the distribution of data on machines is critical (i.e., i.i.d. or not). Third, to illustrate that the gossip variant, despite proposing data diversity to the learning phase, brings only marginal improvements over the classic gossip approach.

This work is a part of the thesis [14].

Deep reinforcement learning for network slicing. Recent achievements in Deep Reinforcement Learning (DRL) have shown the potential of these approaches to solve combinatorial optimization problems. However, the Deep Deterministic Policy Gradient algorithm (DDPG), which is one of the most effective techniques, is not suitable to deal with large-scale discrete action space, which is the case of the Virtual Network Function-Forwarding Graph (VNF-FG) placement. To deal with this problem, we propose several enhancements to improve DDPG efficiency [25][47]. The conventional DDPG generates only one action per iteration; thus, it slowly explores the action space especially in a large action space. Thus, we propose to enhance the exploration by considering multiple noisy actions. In order to avoid getting stuck at a local minimum, we propose to multiply the number of critic (for Q-value) neural networks [25]. In order to improve further the exploration, we propose in [47] an evolutionary algorithm to evolve these neural networks in order to discover better ones.

The techniques presented above are generic and can be applied to a variety of problems. To make them even more effective for network slicing problems, we have also proposed to combine them with a proposed First-Fit heuristic that allows for even more interesting results.

Machine learning for Indoor Outdoor detection. Detecting whether a mobile user is indoor or outdoor is an important issue which significantly impacts user behavior contextualization and mobile network resource management. In [59] we investigate hybrid/semi-supervised Deep Learning-based methods for detecting the environment of an active mobile phone user. They are based on both labeled and unlabeled large real radio data obtained from inside the network and from 3GPP signal measurements. We have empirically evaluated the effectiveness of the semi-supervised learning methods using new real-time radio data, with partial ground truth information, gathered massively from multiple typical and diversified locations (indoor and outdoor) of mobile users. We also presented an analysis of such schemes as compared to the existing supervised classification methods including SVM and Deep Learning [57].

Cognition of user behavior can be seen as an efficient tool for automation of future mobile networks. The work presented in [51] deals with the user behaviour modeling. The model includes the prediction of two main features related to mobile user context: the environment and the mobility. We investigate Deep Learning based methods for simultaneously detecting the environment and the mobility state. We empirically evaluate the effectiveness of the proposed techniques using real-time radio data, which has been massively gathered from multiple diversified situations of mobile users.

Predicting the future Perceived Quality level with PSQA. PSQA is a technology developed by Dionysos during a period of several years, whose aim is quantifying the Quality of Experience (more precisely, the Perceived Quality) of an application or service built on the Internet around the transport of audio or video-audio signals. The main properties of PSQA are the its accuracy (indistinguishable from a subjective testing session), the fact that it is fully automatic, with no reference, and able to operate in real time. PSQA is based on supervised learning (the tool learns from subjective testing panels); once trained and validated, it works with no human intervention. In the PSQA project we selected the Random Neural Network tool for the supervised learning associated tasks, after a comparison with the available techniques at the beginning of the project. In [31] we recall all these elements, including the numerical aspects on the optimization side of the learning process, and then, we focus in the current developments where the goal is to predict the Perceived Quality in the close future. This includes the description of the Reservoir Computing models for time series forecasting, and of a tool we proposed, called Echo State Queueing Network, which is a mix between Reservoir Computing and Random Neural Networks.

7.4. Future networks and architectures

Participants: Yassine Hadjadj-Aoul, Gerardo Rubino, Quang Pham Tran Anh, Anouar Rkhami.

Machine learning for network slicing. Network Function Virtualization (NFV) provides a simple and effective mean to deploy and manage network and telecommunications' services. A typical service can be expressed in the form of a Virtual Network Function-Forwarding Graph (VNF-FG). Allocating a VNF-FG is equivalent to placing VNFs and virtual links onto a given substrate network considering resources and quality of service (QoS) constraints. The deployment of VNF-FGs in large-scale networks, such that QoS measures and deployment cost are optimized, is an emerging challenge. Single-objective VNF-FGs allocation has been addressed in existing literature; however, there is still a lack of studies considering multi-objective VNF-FGs allocation. In addition, it is not trivial to obtain optimal VNF-FGs allocation due to its high computational complexity even in the single-objective case. Genetic algorithms (GAs) have proved their ability in coping with multi-objective optimization problems, thus we propose, in [26], a GA-based scheme to solve multi-objective VNF-FGs allocation problem. The numerical results confirm that the proposed scheme can provide near Pareto-optimal solutions within a short execution time.

In [25], we explore the potential of deep reinforcement learning techniques for the placement of VNF-FGs. However, it turns out that even the most well-known learning technique is ineffective in the context of a large-scale action space. In this respect, we propose approaches to find out feasible solutions while improving significantly the exploration of the action space. The simulation results clearly show the effectiveness of the proposed learning approach for this category of problems. Moreover, thanks to the deep learning process, the performance of the proposed approach is improved over time.

The placement of services, as described above, is extremely complex. The issue is even more complex when it comes to placing a service on several non-cooperative domains, where the network operators hide their infrastructure to other competing domains. In [56], we address these problems by proposing a deep reinforcement learning based VNF-FG embedding approach. The results provide insights into the behaviors of non-cooperative domains. They also show the efficiency of the proposed VNF-FG deployment approach having automatic inter-domain load balancing.

Consistent QoS routing in SDN networks. The Software Defined Networking (SDN) paradigm proposes to decouple the control plane (decision-making process) and the data plane (packet forwarding) to overcome the limitations of traditional network infrastructures, which are known to be difficult to manage, especially at scale. Although there are previous works focusing on the problem of Quality of Service (QoS) routing in SDN networks, only few solutions have taken into consideration the network consistency, which reflects the adequacy between the decisions made and the decisions that should be taken. Therefore, we propose, in [19], a network architecture that guarantees the consistency of the decisions to be taken in an SDN network. A consistent QoS routing strategy is, then, introduced in a way to avoid any quality degradation of prioritized traffic, while optimizing resources usage. Thus, we proposed a traffic dispersion heuristic in order to achieve this goal. We compared our approach to several existing framework in terms of best-effort flows average throughput, average video bitrate and video Quality of Experience (QoE). The emulations results, which are performed using the Mininet environment, clearly demonstrate the effectiveness of the proposed methodology that outperforms existing frameworks.

Optical networks. In [20] we attack the so called *Capacity Crunch* crisis announced for optical networks infrastructures. This problem refers to the facts that (i) the transmission capacity of an optical fiber is not limitless, (ii) the bandwidth demand continues to increase exponentially and (iii) the limits are getting dangerously close. The cheapest and shortest-term solution is to increase efficiency, because there are several possibilities to do so. This work is a contribution in that direction. We focus on strongly improving the wavelength assignment procedure by moving to an heterogeneous and flexible process, adapting the dimensioning to the individual users' needs in QoS. In the paper we demonstrate that a non-uniform dimensioning strategy and a tighten QoS provision allows to save significant networks capacity, while simultaneously provisioning to each user the QoS established in its Service Level Agreement.

Survivability of internet services is a significant and crucial challenge in designing future optical networks. A robust infrastructure and transmission protocols are needed to handle such a situation so that the users can maintain communication despite the existence of one or more failed components in the network. For this reason, we present in [40] a generalized approach able to tolerate any failure scenario, to the extent the user can still communicate with the remaining components, where a scenario corresponds to an arbitrary set of links in a non-operational state. To assess the survivability problem, we propose a joint solution to the problems listed next. We show how to find a set of primary routes, a set of alternate routes associated with each failure scenario, and the capacity required on the network to allow communication between all users, in spite of the links' failures, while satisfying for each user a specific predefined quality of service threshold, defined in the Service Level Agreement (SLA). Numerical results show that the proposed approach not only enjoys the advantages of low complexity and ease of implementation but is also able to achieve significant resource savings compared to existing methods. The savings are higher than 30% on single link failures and more than a 100% on two simultaneous link failures scenarios as well as in more complex situations.

Network tomography. Internet tomography studies the inference of the internal network performances from end-to-end measurements. For this problem, Unicast probing can be advantageous due to the wide support of unicast and the easy deployment of unicast probing paths. In [48] we propose two statistical generic methods for the inference of additive metrics using unicast probing. Our solutions give more flexibility in the choice of the collection points placement. Moreover, the probed paths are not limited to specific topologies. Firstly, we propose the k-paths method that extends the applicability of a previously proposed solution called Flexicast for tree topologies. It is based on the Expectation-Maximization (EM) algorithm which is characterized by high computational and memory complexities. Secondly, we propose the Evolutionary Sampling Algorithm (ESA) that enhances the accuracy and the computing time but following a different approach. In [49] we present a different approach, targeted at link metrics inference in an SDN/NFV environment (even if it can be exported outside this field) that we called TOM (Tomography for Overlay networks Monitoring). In such an environment, we are particularly interested in supervising network slicing, a recent tool enabling to create multiple virtual networks for different applications and QoS constraints on a Telco infrastructure. The goal is to infer the underlay resources states from the measurements performed in the overlay structure. We model the inference task as a regression problem that we solve following a Neural Network approach. Since getting labeled data for the training phase can be costly, our procedure generates artificial data instead. By creating a large set of random training examples, the Neural Network learns the relations between the measures done at path and link levels. This approach takes advantage of efficient Machine Learning solutions to solve a classic inference problem. Simulations with a public dataset show very promising results compared to statistical-based methods.We explored mainly additive metrics such as delays or logs of loss rates, but the approach can also be used for non-additive ones such as bandwidth.

7.5. Wireless Networks

Participants: Yann Busnel, Yassine Hadjadj-Aoul, Ali Hodroj, Bruno Sericola, César Viho.

Self-organized UAV-based Supervision and Connectivity. The use of drones has become more widespread in recent years. Many use cases have developed involving these autonomous vehicles, ranging from simple delivery of packages to complex emergency situations following catastrophic events. The miniaturization and very low cost of these machines make it possible today to create large meshes to ensure network coverage in disaster areas, for instance. However, the problems of scaling up and self-organization are still open in these use cases. In [35], we propose a position paper that first presents different new requirements for the deployment of unmanned aerial vehicles (UAV) networks, involving the use of many drones. Then, it introduces solutions from distributed algorithms and real-time data processing to ensure quasi-optimal solutions to the raised problems.

More specifically, providing network services access anytime and anywhere is nowadays a critical issue, especially in disaster emergency situations. A natural response to such a need is the use of autonomous flying drones to help finding survivors and provide network connectivity to the rescue teams. In [34], we propose VESPA, a distributed algorithm using only one-hop information of the drones, to discover targets with unknown location and auto-organize themselves to ensure connectivity between them and the sink in a multi-hop aerial wireless network. We prove that connectivity, termination and coverage are preserved during all stages of our algorithm, and we evaluate the algorithm performances through simulations. Comparison with a prior work shows the efficiency of VESPA both in terms of discovered targets and number of used drones.

Enhancing dynamic adaptive streaming over HTTP for multi-homed users. Mobile video traffic accounted for more than half of all mobile data traffic over the past two years. Due to the limited bandwidth, users demand for high-quality video streaming becomes a challenge, which could be addressed by exploiting the emerging diversity of access network and adaptive video streaming. In [39], a network selection algorithm is proposed for Dynamic Adaptive Streaming over HTTP (DASH),the famous international standard on video streaming, to enhance the received video quality to a “multi-homed user” equipped with multiple interfaces. A Multi-Armed Bandit (MAB) heuristic is proposed for a dynamic selection of the best interface at each step. While the Adaptive Bit rate Rules (ABR) used in DASH allow the video player client to dynamically pick the bit rate level according to the perceived network conditions, at each switching step a quality degradation may occur due to the difference in network conditions of the available interfaces. This paper aims to close this gap by (i) designing a MAB algorithm over DASH for a multi-homed user, (ii) evaluating the proposed mechanism through a test-bed implementation, (iii) extending the classic MAB model and (iv) discussing some open issues.

Vehicular networks. According to recent forecasts, constant population growth and urbanization will bring an additional load of 2.9 billion vehicles to road networks by 2050. This will certainly lead to increased air pollution concerns, highly congested roads putting more strain on an already deteriorated infrastructure, and may increase the risk of accidents on the roads as well. Therefore, to face these issues we need not only to promote the usage of smarter and greener means of transportation but also to design advanced solutions that leverage the capabilities of these means along with modern cities’ road infrastructure to maximize its utility. In [38], we explore novel ways of utilizing inter-vehicle and vehicle to infrastructure communication technology to achieve a safe and efficient lane change manoeuvre for Connected and Autonomous Vehicles (CAVs). The need for such new protocols is due to the risk that every lane change manoeuvre brings to drivers and passengers lives in addition to its negative impact on congestion level and resulting air pollution, if not performed at the right time and using the appropriate speed. To avoid this risk, we design two new protocols; one is built upon and extends an existing one, and aims at ensuring a safe and efficient lane change manoeuvre, while the second is an original solution inspired from the mutual exclusion concept used in operating systems. This latter complements the former by exclusively granting lane change permission in a way that avoids any risk of collision.

7.6. Network Economics

Participants: Bruno Tuffin, Patrick Maillé.

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team.

In 2019, we have had a particular focus on network neutrality issues, but trying to look at them from original perspectives, and investigating so-called grey zones not yet addressed in the debate.

What implications of a global Internet with neutral and non-neutral portions? Network neutrality is being discussed worldwide, with different countries applying different policies, some imposing it, others acting against regulation or even repealing it as recently in the USA. The goal of [43] is to model and analyze the interactions of users, content providers, and Internet service providers (ISPs) located in countries with different rules. To do so, we build a simple two-regions game-theoretic model and focus on two scenarios of net neutrality relaxation in one region while it remains enforced in the other one. In a first scenario, from an initial situation where both regions offer the same basic quality, one region allows ISPs to offer fast lanes for a premium while still guaranteeing the basic service; in a second scenario the ISPs in both regions play a game on quality, with only one possible quality in the neutral region, and two in the non-neutral one but with a regulated quality ratio between those. Our numerical experiments lead to very different outcomes, with the first scenario benefiting to all actors (especially the ones in the relaxed-neutrality region) and the second one mainly benefiting mostly to ISPs while Content Providers are worse off, suggesting that regulation should be carefully designed.

Investigating a grey zone: sponsored data. Sponsored data, where content providers have the possibility to pay wireless providers for the data consumed by customers and therefore to exclude it from the data cap, is getting widespread in many countries, but is forbidden in others for concerns of infringing the network neutrality principles. We present in [44] a game-theoretic model analyzing the consequences of sponsored data in presence of competing wireless providers, where sponsoring decided by the content provider can be different at each provider. We also discuss the impact on the proportion of advertising on the displayed content. We show that, surprisingly, the possibility of sponsored data may actually reduce the benefits of content providers and on the other hand increase the revenue of ISPs in competition, with a very limited impact on user welfare.

Search engines, bias, consensus, and search neutrality debate. Different search engines provide different outputs for the same keyword. This may be due to different definitions of relevance, and/or to different knowledge/anticipation of users' preferences, but rankings are also suspected to be biased towards own content, which may prejudicial to other content providers. In [41], we make some initial steps toward a rigorous comparison and analysis of search engines, by proposing a definition for a consensual relevance of a page with respect to a keyword, from a set of search engines. More specifically, we look at the results of several search engines for a sample of keywords, and define for each keyword the visibility of a page based on its ranking over all search engines. This allows to define a score of the search engine for a keyword, and then its average score over all keywords. Based on the pages visibility, we can also define the consensus search engine as the one showing the most visible results for each keyword. We have implemented this model and present an analysis of the results in [41].

7.7. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino.

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 of regenerative systems. Rare events occur by definition with a very small probability but are important to analyze because of potential catastrophic consequences. In [32], we focus on rare event for so-called regenerative processes, that are basically processes such that portions of them are statistically independent of each other. For many complex and/or large models, simulation is the only tool at hand but it requires specific implementations to get an accurate answer in a reasonable time. There are two main families of rare-event simulation techniques: Importance Sampling (IS) and Splitting. In a first part, we briefly remind them and compare their respective advantages but later (somewhat arbitrarily) devote most of the work to IS. We then focus on the estimation of the mean hitting time of a rarely visited set. A natural and direct estimator consists in averaging independent and identically distributed copies of simulated hitting times, but an alternative standard estimator uses the regenerative structure allowing to represent the mean as a ratio of quantities. We see that in the setting of crude simulation, the two estimators are actually asymptotically identical in a rare-event context, but inefficient for different, even if related, reasons: the direct estimator requires a large average computational time of a single run whereas the ratio estimator faces a small probability computation. We then explain that the ratio estimator is advised when using IS. In the third part, we discuss the estimation of the distribution, not just the mean, of the hitting time to a rarely visited set of states. We exploit the property that the distribution of the hitting time divided by its expectation converges weakly to an exponential as the target set probability decreases to zero. The problem then reduces to the extensively studied estimation of the mean described previously. It leads to simple estimators of a quantile and conditional tail expectation of the hitting time. Some variants are presented and the accuracy of the estimators is illustrated on numerical examples.

In [46], we introduce and analyze a new regenerative estimator. A classical simulation estimator of this class is based on a ratio representation of the mean hitting time, using crude simulation to estimate the numerator and importance sampling to handle the denominator, which corresponds to a rare event. But the estimator of the numerator can be inefficient when paths to the set are very long. We thus introduce a new estimator that expresses the numerator as a sum of two terms to be estimated separately. We provide theoretical analysis of a simple example showing that the new estimator can have much better behavior than the classical estimator. Numerical results further illustrate this.

Randomized Quasi-Monte Carlo for Quantile Estimation. Quantile estimation is a key issue in many application domains, but has been proved difficult to efficiently estimate. In [42], we compare two approaches for quantile estimation via randomized quasi-Monte Carlo (RQMC) in an asymptotic setting where the number of randomizations for RQMC grows large but the size of the low-discrepancy point set remains fixed. In the first method, for each randomization, we compute an estimator of the cumulative distribution function (CDF), which is inverted to obtain a quantile estimator, and the overall quantile estimator is the sample average of the quantile estimators across randomizations. The second approach instead computes a single quantile estimator by inverting one CDF estimator across all randomizations. Because quantile estimators are generally biased, the first method leads to an estimator that does not converge to the true quantile as the number of randomizations goes to infinity. In contrast, the second estimator does, and we establish a central limit theorem for it. Numerical results further illustrate these points.

Reliability analysis with dependent components. In the reliability area, the Marshall-Olkin copula model has emerged as the standard tool for capturing dependence between components in failure analysis. In this model, shocks arise at exponential random times, affecting one or several components, thus inducing a natural correlation in the failure process. However, because the number of parameter of the model grows exponentially with the number of components, the tool suffers from the "curse of dimensionality." These models are usually intended to be applied to design a network before its construction; therefore, it is natural to assume that only partial information about failure behavior can be gathered, mostly from similar existing networks. To construct them, we propose in [22] an optimization approach to define the shock's parameters in the copula, in order to match marginal failures probabilities and correlations between these failures. To deal with the exponential number of parameters of the problem, we use a column-generation technique. We also discuss additional criteria that can be incorporated to obtain a suitable model. Our computational experiments show that the resulting tool produces a close estimation of the network reliability, especially when the correlation between component failures is significant.

The Creation Process is an algorithm that transforms a static network model into a dynamic one. It is the basis of different variance reduction methods designed to make efficient reliability estimations on highly reliable networks in which links can only assume two possible values, operational or failed. In [18] the Creation Process is extended to let it operate on network models in which links can assume more than two values. The proposed algorithm, that we called Multi-Level Creation Process, is the basis of a method, also introduced here, to make efficient reliability estimations of highly reliable stochastic flow networks. The method proposed, which consists in an application of Splitting over the Multi-Level Creation Process, is empirically shown to be accurate, efficient, and robust. This work was the first step towards a way to implement an efficient estimation procedure for the problem of flow reliability analysis. Our first solution in that direction was presented in [54], where not only we could develop a procedure providing a significant variance reduction but that allows a direct extension to the final target, the solution to the same estimation problem in the more general case of models where the components are dependent. The idea is an original way of implementing a splitting procedure that leads simultaneously to these two properties.

Rare events in risk analysis. One of the main tasks when dealing with critical systems (systems where specific classes of failures can deal to human losses, or to huge financial losses) is the ability to quantify the associated risks, which is the door that, when opened, leads to paths towards understanding what can happen and why, and towards capturing the relationships existing between the different parts of the system, with respect to those risks. This is also the necessary preliminary work allowing to evaluate the relative importance of different factors, always from the viewpoint of the considered risks, an important component of any disaster management system. Identifying the dominant ones is important to know which parts of the system we must reinforce. The keynote [29] described different tools available for these tasks, and how they can be used depending on the objectives to reach. The focus was on Monte Carlo techniques, the only available ones in general, because the only ones able to evaluate any kind of system, and how they deal with rare events. It also discussed the main related open research problems. The tutorial [64] is closely related to previous talk, but the presentation explores more in general the estimation problem and the main families of techniques available for its solution (Importance Sampling, and the particular case of Zero-Variance methods, Splitting, Recursive Variance Reduction techniques, etc.).

DYOGENE Project-Team

7. New Results

7.1. Distributed network control and smart-grids

1. Distributed Control of Thermostatically Controlled Loads: Kullback-Leibler Optimal Control in Continuous Time [20] The paper develops distributed control techniques to obtain grid services from flexible loads. The Individual Perspective Design (IPD) for local (load level) control is extended to piecewise deterministic and diffusion models for thermostatically controlled load models. The IPD design is formulated as an infinite horizon average reward optimal control problem, in which the reward function contains a term that uses relative entropy rate to model deviation from nominal dynamics. In the piecewise deterministic model, the optimal solution is obtained via the solution to an eigenfunction problem, similar to what is obtained in prior work. For a jump diffusion model this simple structure is absent. The structure for the optimal solution is obtained, which suggests an ODE technique for computation that is likely far more efficient than policy-or value-iteration.

2. Optimal Control of Dynamic Bipartite Matching Models [23] A dynamic bipartite matching model is given by a bipartite matching graph which determines the possible matchings between the various types of supply and demand items. Both supply and demand items arrive to the system according to a stochastic process. Matched pairs leave the system and the others wait in the queues, which induces a holding cost. We model this problem as a Markov Decision Process and study the discounted cost and the average cost case. We first consider a model with two types of supply and two types of demand items with an N-shaped matching graph. For linear cost function, we prove that an optimal matching policy gives priority to the end edges of the matching graph and is of threshold type for the diagonal edge. In addition, for the average cost problem, we compute the optimal threshold value. According to our numerical experiments, threshold-type policies perform also very well for more general bipartite graphs.

3. Kullback-Leibler-Quadratic Optimal Control of Flexible Power Demand [24] A new stochastic control methodology is introduced for distributed control, motivated by the goal of creating virtual energy storage from flexible electric loads, i.e. Demand Dispatch. In recent work, the authors have introduced Kullback-Leibler-Quadratic (KLQ) optimal control as a stochastic control methodology for Markovian models. This paper develops KLQ theory and demonstrates its applicability to demand dispatch. In one formulation of the design, the grid balancing authority simply broadcasts the desired tracking signal, and the heterogeneous population of loads ramps power consumption up and down to accurately track the signal. Analysis of the Lagrangian dual of the KLQ optimization problem leads to a menu of solution options, and expressions of the gradient and Hessian suitable for Monte-Carlo-based optimization. Numerical results illustrate these theoretical results.

4. Bike sharing systems: a new incentive rebalancing method based on spatial outliers detection [8] Since its launch, Velib' (the Bike Sharing System-BSS-in Paris) has emerged in the Parisian landscape and has been a model for similar systems in many cities. A major problem with BSS is the stations' heterogeneity caused by the attractivity of some stations located in particular areas. In this paper, we focus on spatial outliers defined as stations having a behavior significantly different from their neighboring stations. First, we propose an improved version of Moran scatterplot to exploit the similarity between neighbors, and we test it on a real dataset issued from Velib' system to identify outliers. Then, we design a new method that globally improves the resources' availability in bike stations by adapting the users' trips to the resources' availability. Results show that with a partial collaboration of the users or a limitation to the rush hours, the proposed method enhances significantly the resources' availability in Velib' system.

5. Stochastic Battery Operations using Deep Neural Networks [25] In this paper, we introduce a scenario-based optimal control framework to account for the forecast uncertainty in battery arbitrage problems. Due to the uncertainty of prices and variations of forecast errors, it is challenging for battery operators to design profitable strategies in electricity markets. Without any explicit assumption or model for electricity price forecasts' uncertainties, we generate future price scenarios via a data-driven, learning-based approach. By aiding the predictive control with such scenarios representing possible realizations of future markets, our proposed real-time controller seeks the optimal charge/discharge levels to maximize profits. Simulation results on a case-study of California-based batteries and prices show that our proposed method can bring higher profits for different battery parameters.

6. Aggregate capacity for TCLs providing virtual energy storage with cycling constraints [26] The coordination of thermostatically controlled loads (TCLs) is challenging due to the need to meet individual loads quality of service (QoS), such as indoor temperature constraints. Since these loads are usually on/off type, cycling rate is one of their QoS metrics; frequent cycling between on and off states is detrimental to them. While significant prior work has been done on the coordination of air conditioning TCLs, the question of cycling QoS has not been investigated in a principled manner. In this work we propose a method to characterize aggregate capacity of a collection of air conditioning TCLs that respects the loads cycling rate constraints (maximum number of cycles in a given time period). The development is done within the framework of randomized local control in which a load makes on/off decisions probabilistically. This characterization allows us to propose a reference planning problem to generate feasible reference trajectories for the ensemble that respect cycling constraints. The reference planning problem manifests itself in the form a Nonlinear Programming problem (NLP), that can be efficiently solved. Our proposed method is compared to previous methods in the literature that do not enforce aggregate cycling. Enforcing individual cycling constraint without taking that into account in reference generation leads to poor reference tracking.

7. Optimal Storage Arbitrage under Net Metering using Linear Programming [29] We formulate the optimal energy arbitrage problem for a piecewise linear cost function for energy storage devices using linear programming (LP). The LP formulation is based on the equivalent minimization of the epigraph. This formulation considers ramping and capacity constraints, charging and discharging efficiency losses of the storage, inelastic consumer load and local renewable generation in presence of net-metering which facilitates selling of energy to the grid and incentivizes consumers to install renewable generation and energy storage. We consider the case where the consumer loads, electricity prices, and renewable generations at different instances are uncertain. These uncertain quantities are predicted using an Auto-Regressive Moving Average (ARMA) model and used in a model predictive control (MPC) framework to obtain the arbitrage decision at each instance. In numerical results we present the sensitivity analysis of storage performing arbitrage with varying ramping batteries and different ratio of selling and buying price of electricity.

8. Energy Storage in Madeira, Portugal: Co-optimizing for Arbitrage, Self-Sufficiency, Peak Shaving and Energy Backup [30] Energy storage applications are explored from a prosumer (consumers with generation) perspective for the island of Madeira in Portugal. These applications could also be relevant to other power networks. We formulate a convex co-optimization problem for performing arbitrage under zero feed-in tariff, increasing self-sufficiency by increasing self-consumption of locally generated renewable energy, provide peak shaving and act as a backup power source during anticipated and scheduled power outages. Using real data from Madeira we perform short and long timescale simulations in order to select end-user contract which maximizes their gains considering storage degradation based on operational cycles. We observe energy storage ramping capability decides peak shaving potential, fast ramping batteries can significantly reduce peak demand charge. The numerical experiment indicates that storage providing backup does not significantly reduce gains performing arbitrage and peak demand shaving. Furthermore, we also use AutoRegressive Moving Average (ARMA) forecasting along with Model Predic-tive Control (MPC) for real-time implementation of the proposed optimization problem in the presence of uncertainty.

9. Sensitivity to forecast errors in energy storage arbitrage for residential consumers [34] With the massive deployment of distributed energy resources, there has been an increase in the number of end consumers that own photovoltaic panels and storage systems. The optimal use of such storage when facing Time of Use (ToU) prices is directly related to the quality of the load and generation forecasts as well as the algorithm that controls the battery. The sensitivity of such control to different forecasts techniques is studied in this paper. It is shown that good and bad forecasts can result in losses in particularly bad days. Nevertheless, it is observed that performing Model Predictive Control with a simple forecast that is representative of the pasts can be profitable under different price and battery scenarios. We use real data from Pecan Street and ToU price levels with different buying and selling price for the numerical experiments.

10. Sizing and Profitability of Energy Storage for Prosumers in Madeira, Portugal [47] This paper proposes a framework to select the best-suited battery for co-optimizing for peak demand shaving, energy arbitrage and increase self-sufficiency in the context of power network in Madeira, Portugal. Feed-in-tariff for electricity network in Madeira is zero, which implies consumers with excess production should locally consume the excess generation rather than wasting it. Further, the power network operator applies a peak power contract for consumers which imposes an upper bound on the peak power seen by the power grid interfaced by energy meter. We investigate the value of storage in Madeira, using four different types of prosumers, categorized based on the relationship between their inelastic load and renewable generation. We observe that the marginal increase in the value of storage deteriorates with increase in size and ramping capabilities. We propose the use of profit per cycle per unit of battery capacity and expected payback period as indices for selecting the best-suited storage parameters to ensure profitability. This mechanism takes into account the consumption and generation patterns, profit, storage degradation, and cycle and calendar life of the battery. We also propose the inclusion of a friction coefficient in the original co-optimization formulation to increase the value of storage by reducing the operational cycles and eliminate low returning transactions.

11. Arbitrage with Power Factor Correction using Energy Storage [48] The importance of reactive power compensation for power factor (PF) correction will significantly increase with the large-scale integration of distributed generation interfaced via inverters producing only active power. In this work, we focus on co-optimizing energy storage for performing energy arbitrage as well as local power factor corrections. The joint optimization problem is non-convex, but can be solved efficiently using a McCormick relaxation along with penalty-based schemes. Using numerical simulations on real data and realistic storage profiles, we show that energy storage can correct PF locally without reducing arbitrage gains. It is observed that active and reactive power control is largely decoupled in nature for performing arbitrage and PF correction (PFC). Furthermore, we consider a stochastic online formulation of the problem with uncertain load, renewable and pricing profiles. We develop a model predictive control based storage control policy using ARMA forecast for the uncertainty. Using numerical simulations we observe that PFC is primarily governed by the size of the converter and therefore, look-ahead in time in the online setting does not affect PFC noticeably. However, arbitrage gains are more sensitive to uncertainty for batteries with faster ramp rates compared to slow ramping batteries.

12. A Utility Optimization Approach to Network Cache Design [11] 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.

13. Rapid Mixing of Dynamic Graphs with Local Evolution Rules [15] Dynamic graphs arise naturally in many contexts. In peer-to-peer networks, for instance, a participating peer may replace an existing connection with one neighbor by a new connection with a neighbor of that neighbor. Several such local rewiring rules have been proposed to ensure that peer-to-peer networks achieve good connectivity properties (e.g. high expansion) at equilibrium. However, the question of whether there exists such a rule that converges rapidly to equilibrium has remained open. In this work, we provide an affirmative answer: we exhibit a local rewiring rule that converges to equilibrium after each participating node has undergone only a number of changes that is at most poly-logarithmic in the system size. As a byproduct, we derive new results for random walks on graphs, bounding the spread of their law throughout the transient phase, i.e. prior to mixing. These rely on an extension of Cheeger's inequality, based on generalized isoperimetric constants, and may be of independent interest.

7.2. Reinforcement learning

14. On Matrix Momentum Stochastic Approximation and Applications to Q-learning [27] Stochastic approximation (SA) algorithms are recursive techniques used to obtain the roots of functions that can be expressed as expectations of a noisy parameterized family of functions. In this paper two new SA algorithms are introduced: 1) PolSA, an extension of Polyak's momentum technique with a specially designed matrix momentum, and 2) NeSA, which can either be regarded as a variant of Nesterov's acceleration method, or a simplification of PolSA. The rates of convergence of SA algorithms is well understood. Under special conditions, the mean square error of the parameter estimates is bounded by $\sigma^2/n + o(1/n)$, where $\sigma^2 \geq 0$ is an identifiable constant. If these conditions fail, the rate is typically sub-linear. There are two well known SA algorithms that ensure a linear rate, with minimal value of variance, σ^2: the Ruppert-Polyak averaging technique, and the stochastic Newton-Raphson (SNR) algorithm. It is demonstrated here that under mild technical assumptions, the PolSA algorithm also achieves this optimality criteria. This result is established via novel coupling arguments: It is shown that the parameter estimates obtained from the PolSA algorithm couple with those of the optimal variance (but computationally more expensive) SNR algorithm, at a rate $O(1/n^2)$. The newly proposed algorithms are extended to a reinforcement learning setting to obtain new Q-learning algorithms, and numerical results confirm the coupling of PolSA and SNR.

15. Zap Q-Learning - A User's Guide [28] There are two well known Stochastic Approximation techniques that are known to have optimal rate of convergence (measured in terms of asymptotic variance): the Stochastic Newton-Raphson (SNR) algorithm (a matrix gain algorithm that resembles the deterministic Newton-Raphson method), and the Ruppert-Polyak averaging technique. This paper surveys new applications of these concepts for Q-learning: (i)The Zap Q-Learning algorithm was introduced by the authors in a NIPS 2017 paper. It is based on a variant of SNR, designed to more closely mimic its deterministic cousin. The algorithm has optimal rate of convergence under general assumptions, and showed astonishingly quick convergence in numerical examples. These algorithms are surveyed and illustrated with numerical examples. A potential difficulty in implementation of the Zap-Q-Learning algorithm is the matrix inversion required in each iteration. (ii)Remedies are proposed based on stochastic approximation variants of two general deterministic techniques: Polyak's momentum algorithms and Nesterov's acceleration technique. Provided the hyper-parameters are chosen with care, the performance of these algorithms can be comparable to the Zap algorithm, while computational complexity per iteration is far lower.

16. Zap Q-Learning With Nonlinear Function Approximation [44] The Zap stochastic approximation (SA) algorithm was introduced recently as a means to accelerate convergence in reinforcement learning algorithms. While numerical results were impressive, stability (in the sense of boundedness of parameter estimates) was established in only a few special cases. This class of algorithms is generalized in this paper, and stability is established under very general conditions. This general result can be applied to a wide range of algorithms found in reinforcement learning. Two classes are considered in this paper: (i)The natural generalization of Watkins' algorithm is not always stable in function approximation settings. Parameter estimates may diverge to infinity even in the *linear* function approximation setting with a simple finite state-action MDP. Under mild conditions, the Zap SA algorithm provides a stable algorithm, even in the case of *nonlinear* function approximation. (ii) The GQ algorithm of Maei et. al. 2010 is designed to address the stability challenge. Analysis is provided to explain why the algorithm may be very slow to converge in practice. The new Zap GQ algorithm is stable even for nonlinear function approximation.

17. Zap Q-Learning for Optimal Stopping Time Problems [43] The objective in this paper is to obtain fast converging reinforcement learning algorithms to approximate solutions to the problem of discounted cost optimal stopping in an irreducible, uniformly ergodic Markov chain, evolving on a compact subset of IR^n. We build on the dynamic programming approach taken by Tsitsikilis and Van Roy, wherein they propose a Q-learning algorithm to estimate the optimal state-action value function, which then defines an optimal stopping rule. We provide insights as to why the convergence rate of this algorithm can be slow, and propose a fast-converging alternative, the "Zap-Q-learning" algorithm, designed to achieve optimal rate of convergence. For the first time, we prove the convergence of the Zap-Q-learning algorithm under the assumption of linear function approximation setting. We use ODE analysis for the proof, and the optimal asymptotic variance property of the algorithm is reflected via fast convergence in a finance example.

7.3. Mathematics of wireless cellular networks

18. Performance analysis of cellular networks with opportunistic scheduling using queueing theory and stochastic geometry [6] Combining stochastic geometric approach with some classical results from queuing theory, in this paper we propose a comprehensive framework for the performance study of large cellular networks featuring opportunistic scheduling. Rapid and verifiable with respect to real data, our approach is particularly useful for network dimensioning and long term economic planning. It is based on a detailed network model combining an information-theoretic representation of the link layer, a queuing-theoretic representation of the users' scheduler, and a stochastic-geometric representation of the signal propagation and the network cells. It allows one to evaluate principal characteristics of the individual cells, such as loads (defined as the fraction of time the cell is not empty), the mean number of served users in the steady state, and the user throughput. A simplified Gaussian approximate model is also proposed to facilitate study of the spatial distribution of these metrics across the network. The analysis of both models requires only simulations of the point process of base stations and the shadowing field to estimate the expectations of some stochastic-geometric functionals not admitting explicit expressions. A key observation of our approach , bridging spatial and temporal analysis, relates the SINR distribution of the typical user to the load of the typical cell of the network. The former is a static characteristic of the network related to its spectral efficiency while the latter characterizes the performance of the (generalized) processor sharing queue serving the dynamic population of users of this cell.

19. Two-tier cellular networks for throughput maximization of static and mobile users [10] 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, 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, 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.

20. Location Aware Opportunistic Bandwidth Sharing between Static and Mobile Users with Stochastic Learning in Cellular Networks [9] 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.

21. Per-Link Reliability and Rate Control: Two Facets of the SIR Meta Distribution [13] The meta distribution (MD) of the signal-to-interference ratio (SIR) provides fine-grained reliability performance in wireless networks modeled by point processes. In particular, for an ergodic point process, the SIR MD yields the distribution of the per-link reliability for a target SIR. Here we reveal that the SIR MD has a second important application, which is rate control. Specifically, we calculate the distribution of the SIR threshold (equivalently, the distribution of the transmission rate) that guarantees each link a target reliability and show its connection to the distribution of the per-link reliability. This connection also permits an approximate calculation of the SIR MD when only partial (local) information about the underlying point process is available.

22. Simple Approximations of the SIR Meta Distribution in General Cellular Networks [14] Compared to the standard success (coverage) probability , the meta distribution of the signal-to-interference ratio (SIR) provides much more fine-grained information about the network performance. We consider general heterogeneous cellular networks (HCNs) with base station tiers modeled by arbitrary stationary and ergodic non-Poisson point processes. The exact analysis of non-Poisson network models is notoriously difficult, even in terms of the standard success probability, let alone the meta distribution. Hence we propose a simple approach to approximate the SIR meta distribution for non-Poisson networks based on the ASAPPP ("approximate SIR analysis based on the Poisson point process") method. We prove that the asymptotic horizontal gap G_0 between its standard success probability and that for the Poisson point process exactly characterizes the gap between the bth moment of the conditional success probability, as the SIR threshold goes to 0. The gap G_0 allows two simple approximations of the meta distribution for general HCNs: 1) the per-tier approximation by applying the shift G_0 to each tier and 2) the effective gain approximation by directly shifting the meta distribution for the homogeneous independent Poisson network. Given the generality of the model considered and the fine-grained nature of the meta distribution, these approximations work surprisingly well.

23. Interference Queueing Networks [16] This work features networks of coupled processor sharing queues in the Euclidean space, where customers arrive according to independent Poisson point processes at every queue, are served, and then leave the network. The coupling is through service rates. In any given queue, this rate is inversely proportional the interference seen by this queue, which is determined by the load in neighboring queues, attenuated by some distance-based path-loss function. The main focus is on the infinite grid network and translation invariant path-loss case. The model is a discrete version of a spatial birth and death process where customers arrive to the Euclidean space according to Poisson rain and leave it when they have transferred an exponential file, assuming that the instantaneous rate of each transfer is determined through information theory by the signal to interference and noise ratio experienced by the user. The stability condition is identified. The minimal stationary regime is built using coupling from the past techniques. The mean queue size of this minimal stationary regime is determined in closed form using the rate conservation principle of Palm calculus. When the stability condition holds, for all bounded initial conditions, there is weak convergence to this minimal stationary regime; however, there exist translation invariant initial conditions for which all queue sizes converge to infinity.

24. Statistical learning of geometric characteristics of wireless networks [19] Motivated by the prediction of cell loads in cellular networks, we formulate the following new, fundamental problem of statistical learning of geometric marks of point processes: An unknown marking function, depending on the geometry of point patterns, produces characteristics (marks) of the points. One aims at learning this function from the examples of marked point patterns in order to predict the marks of new point patterns. To approximate (interpolate) the marking function, in our baseline approach, we build a statistical regression model of the marks with respect some local point distance representation. In a more advanced approach, we use a global data representation via the scattering moments of random measures, which build informative and stable to deformations data representation, already proven useful in image analysis and related application domains. In this case, the regression of the scattering moments of the marked point patterns with respect to the non-marked ones is combined with the numerical solution of the inverse problem, where the marks are recovered from the estimated scattering moments. Considering some simple, generic marks, often appearing in the modeling of wireless networks, such as the shot-noise values, nearest neighbour distance, and some characteristics of the Voronoi cells, we show that the scattering moments can capture similar geometry information as the baseline approach, and can reach even better performance, especially for non-local marking functions. Our results motivate further development of statistical learning tools for stochastic geometry and analysis of wireless networks, in particular to predict cell loads in cellular networks from the locations of base stations and traffic demand.

25. Determinantal thinning of point processes with network learning applications [21] A new type of dependent thinning for point processes in continuous space is proposed, which leverages the advantages of determinantal point processes defined on finite spaces and, as such, is particularly amenable to statistical, numerical, and simulation techniques. It gives a new point process that can serve as a network model exhibiting repulsion. The properties and functions of the new point process, such as moment measures, the Laplace functional, the void probabilities, as well as conditional (Palm) characteristics can be estimated accurately by simulating the underlying (non-thinned) point process, which can be taken, for example, to be Poisson. This is in contrast (and preference to) finite Gibbs point processes, which, instead of thinning, require weighting the Poisson realizations, involving usually intractable normalizing constants. Models based on determinantal point processes are also well suited for statistical (supervised) learning techniques, allowing the models to be fitted to observed network patterns with some particular geometric properties. We illustrate this approach by imitating with determinantal thinning the well-known Matérn II hard-core thinning, as well as a soft-core thinning depending on nearest-neighbour triangles. These two examples demonstrate how the proposed approach can lead to new, statistically optimized, probabilistic transmission scheduling schemes.

26. Analyzing LoRa long-range, low-power, wide-area networks using stochastic geometry [22] In this paper we present a simple, stochastic-geometric model of a wireless access network exploiting the LoRA (Long Range) protocol, which is a non-expensive technology allowing for long-range, single-hop connectivity for the Internet of Things. We assume a space-time Poisson model of packets transmitted by LoRA nodes to a fixed base station. Following previous studies of the impact of interference, we assume that a given packet is successfully received when no interfering packet arrives with similar power before the given packet payload phase. This is as a consequence of LoRa using different transmission rates for different link budgets (transmissions with smaller received powers use larger spreading factors) and LoRa intra-technology interference treatment. Using our model, we study the scaling of the packet reception probabilities per link budget as a function of the spatial density of nodes and their rate of transmissions. We consider both the parameter values recommended by the LoRa provider, as well as proposing LoRa tuning to improve the equality of performance for all link budgets. We also consider spatially non-homogeneous distributions of LoRa nodes. We show also how a fair comparison to non-slotted Aloha can be made within the same framework.

27. Reliability and Local Delay in Wireless Networks: Does Bandwidth Partitioning Help? [33] In a series of papers initiated through a collaboration with Nokia Bell Labs, we study the effect of bandwidth partitioning (BWP) on the reliability and delay performance in infrastructureless wireless networks. The reliability performance is characterized by the density of concurrent transmissions that satisfy a certain reliability (outage) constraint and the delay performance by so-called local delay, defined as the average number of time slots required to successfully transmit a packet. We concentrate on the ultrareliable regime where the target outage probability is close to 0. BWP has two conflicting effects: while the interference is reduced as the concurrent transmissions are divided over multiple frequency bands, the signal-to-interference ratio (SIR) requirement is increased due to smaller allocated bandwidth if the data rate is to be kept constant. Instead, if the SIR requirement is to be kept the same, BWP reduces the data rate and in turn increases the local delay. For these two approaches with adaptive and fixed SIR requirements, we derive closed-form expressions of the local delay and the maximum density of reliable transmissions in the ultrareliable regime. Our analysis shows that, in the ultrareliable regime, BWP leads to the reliability-delay tradeoff.

28. The Influence of Canyon Shadowing on Device-to-Device Connectivity in Urban Scenario [35] In this work, we use percolation theory to study the feasibility of large-scale connectivity of relay-augmented device-to-device (D2D) networks in an urban scenario, featuring a haphazard system of streets and canyon shadowing allowing only for line-of-sight (LOS) communications in a limited finite range. We use a homogeneous Poisson-Voronoi tessellation (PVT) model of streets with homogeneous Poisson users (devices) on its edges and independent Bernoulli relays on the vertices. Using this model, we demonstrated the existence of a minimal threshold for relays below which large-scale connectivity of the network is not possible, regardless of all other network parameters. Through simulations, we estimated this threshold to 71.3%. Moreover, if the mean street length is not larger than some threshold (predicted to 74.3% of the communication range; which might be the case in a typical urban scenario) then any (whatever small) density of users can be compensated by equipping more crossroads with relays. Above this latter threshold, good connectivity requires some minimal density of users, compensated by the relays in a way we make explicit. The existence of the above regimes brings interesting qualitative arguments to the discussion on the possible D2D deployment scenarios.

29. Relay-assisted Device-to-Device Networks: Connectivity and Uberization Opportunities [46] It has been shown that deploying device-to-device (D2D) networks in urban environments requires equipping a considerable proportion of crossroads with relays. This represents a necessary economic investment for an operator. In this work, we tackle the problem of the economic feasibility of such relay-assisted D2D networks. First, we propose a stochastic model taking into account a positive surface for streets and crossroads, thus allowing for a more realistic estimation of the minimal number of needed relays. Secondly, we introduce a cost model for the deployment of relays, allowing one to study operators' D2D deployment strategies. We investigate the example of an uberizing neo-operator willing to set up a network entirely relying on D2D and show that a return on the initial investment in relays is possible in a realistic period of time, even if the network is funded by a very low revenue per D2D user. Our results bring quantitative arguments to the discussion on possible uberization scenarios of telecommunications networks.

30. Continuum Line-of-Sight Percolation on Poisson-Voronoi Tessellations [45] In this work, we study a new model for continuum line-of-sight percolation in a random environment given by a Poisson-Voronoi tessellation. The edges of this tessellation are the support of a Cox point process, while the vertices are the support of a Bernoulli point process. Taking the superposition of these two processes, two points of are linked by an edge if and only if they are sufficiently close and located on the same edge of the supporting tessellation. We study the percolation of the random graph arising from this construction and prove that a subcritical phase as well as a supercritical phase exist under general assumptions. Our proofs are based on a renormalization argument with some notion of stabilization and asymptotic essential connectedness to investigate continuum percolation for Cox point processes. We also give numerical estimates of the critical parameters of the model. Our model can be seen as a good candidate for modelling telecommunications networks in a random environment with obstructive conditions for signal propagation.

7.4. High-dimensional statistical inference

31. Discrete Mean Field Games: Existence of Equilibria and Convergence [12] We consider mean field games with discrete state spaces (called discrete mean field games in the following) and we analyze these games in continuous and discrete time, over finite as well as infinite time horizons. We prove the existence of a mean field equilibrium assuming continuity of the cost and of the drift. These conditions are more general than the existing papers studying finite state space mean field games. Besides, we also study the convergence of the equilibria of N -player games to mean field equilibria in our four settings. On the one hand, we define a class of strategies in which any sequence of equilibria of the finite games converges weakly to a mean field equilibrium when the number of players goes to infinity. On the other hand, we exhibit equilibria outside this class that do not converge to mean field equilibria and for which the value of the game does not converge. In discrete time this non-convergence phenomenon implies that the Folk theorem does not scale to the mean field limit.

32. Modularity-based Sparse Soft Graph Clustering [32] Clustering is a central problem in machine learning for which graph-based approaches have proven their efficiency. In this paper, we study a relaxation of the modularity maxi-mization problem, well-known in the graph partitioning literature. A solution of this relaxation gives to each element of the dataset a probability to belong to a given cluster, whereas a solution of the standard modularity problem is a partition. We introduce an efficient optimization algorithm to solve this relaxation, that is both memory efficient and local. Furthermore, we prove that our method includes, as a special case, the Louvain optimization scheme, a state-of-the-art technique to solve the traditional modularity problem. Experiments on both synthetic and real-world data illustrate that our approach provides meaningful information on various types of data.

33. Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models [41] We consider generalized linear models where an unknown n-dimensional signal vector is observed through the successive application of a random matrix and a non-linear (possibly probabilistic) componentwise function. We consider the models in the high-dimensional limit, where the observation consists of m points, and $m/n \to \alpha$ where α stays finite in the limit $m, n \to \infty$. This situation is ubiquitous in applications ranging from supervised machine learning to signal processing. A substantial amount of work suggests that both the inference and learning tasks in these problems have sharp intrinsic limitations when the available data become too scarce or too noisy. Here, we provide rigorous asymptotic predictions for these thresholds through the proof of a simple expression for the mutual information between the observations and the signal. Thanks to this expression we also obtain as a consequence the optimal value of the generalization error in many statistical learning models of interest, such as the teacher-student binary perceptron, and introduce several new models with remarquable properties. We compute these thresholds (or "phase transitions") using ideas from statistical physics that are turned into rigorous methods thanks to a new powerful smart-path interpolation technique called the stochastic interpolation method, which has recently been introduced by two of the authors. Moreover we show that a polynomial-time algorithm refered to as generalized approximate message-passing reaches the optimal generalization performance for a large set of parameters in these problems. Our results clarify the difficulties and challenges one has to face when solving complex high-dimensional statistical problems.

34. Efficient inference in stochastic block models with vertex labels [18] We study the stochastic block model with two communities where vertices contain side information in the form of a vertex label. These vertex labels may have arbitrary label distributions, depending on the community memberships. We analyze a version of the popular belief propagation algorithm. We show that this algorithm achieves the highest accuracy possible whenever a certain function of the network parameters has a unique fixed point. When this function has multiple fixed points, the belief propagation algorithm may not perform optimally, where we conjecture that a non-polynomial time algorithm may perform better than BP. We show that increasing the information in the vertex labels may reduce the number of fixed points and hence lead to optimality of belief propagation.

35. Planting trees in graphs, and finding them back [36] In this paper we study detection and reconstruction of planted structures in Erdős-Rényi random graphs. Motivated by a problem of communication security, we focus on planted structures that consist in a tree graph. For planted line graphs, we establish the following phase diagram. In a low density region where the average degree λ of the initial graph is below some critical value $\lambda_c = 1$, detection and reconstruction go from impossible to easy as the line length K crosses some critical value $f(\lambda)\ln(n)$, where n is the number of nodes in the graph. In the high density region $\lambda > \lambda_c$, detection goes from impossible to easy as K goes from $o(\sqrt{n})$ to $\omega(\sqrt{n})$, and reconstruction remains impossible so long as $K = o(n)$. For D-ary trees of varying depth h and $2 \leq D \leq O(1)$, we identify a low-density region $\lambda < \lambda_D$, such that the following holds. There is a threshold $h* = g(D)\ln(\ln(n))$ with the following properties. Detection goes from feasible to impossible as h crosses $h*$. We also show that only partial reconstruction is feasible at best for $h \geq h*$. We conjecture a similar picture to hold for D-ary trees as for lines in the high-density region $\lambda > \lambda_D$, but confirm only the following part of this picture: Detection is easy for D-ary trees of size $\omega(\sqrt{n})$, while at best only partial reconstruction is feasible for D-ary trees of any size $o(n)$. These results are in contrast with the corresponding picture for detection and reconstruction of *low rank* planted structures, such as dense subgraphs and block communities: We observe a discrepancy between detection and reconstruction, the latter being impossible for a wide range of parameters where detection is easy. This property does not hold for previously studied low rank planted structures.

36. Robustness of spectral methods for community detection [37] This work is concerned with community detection. Specifically, we consider a random graph drawn according to the stochastic block model: its vertex set is partitioned into blocks, or communities, and edges are placed randomly and independently of each other with probability depending only on the communities of their two endpoints. In this context, our aim is to recover the community labels better than by random guess, based only on the observation of the graph.

In the sparse case, where edge probabilities are in $O(1/n)$, we introduce a new spectral method based on the distance matrix D, where $D_{ij} = 1$ iff the graph distance between i and j, noted $d(i,j)$ is equal to ℓ. We show that when $\ell \sim c\log(n)$ for carefully chosen c, the eigenvectors associated to the largest eigenvalues of D provide enough information to perform non-trivial community recovery with high probability, provided we are above the so-called Kesten-Stigum threshold. This yields an efficient algorithm for community detection, since computation of the matrix D can be done in $O(n^{1+\kappa})$ operations for a small constant κ.

We then study the sensitivity of the eigendecomposition of D when we allow an adversarial perturbation of the edges of G. We show that when the considered perturbation does not affect more than $O(n^{\varepsilon})$ vertices for some small $\varepsilon > 0$, the highest eigenvalues and their corresponding eigenvectors incur negligible perturbations, which allows us to still perform efficient recovery.

Our proposed spectral method therefore: i) is robust to larger perturbations than prior spectral methods, while semi-definite programming (or SDP) methods can tolerate yet larger perturbations; ii) achieves non-trivial detection down to the KS threshold, which is conjectured to be optimal and is beyond reach of existing SDP approaches; iii) is faster than SDP approaches.

7.5. Distributed optimization for machine learning

37. Optimal Convergence Rates for Convex Distributed Optimization in Networks [17] This work proposes a theoretical analysis of distributed optimization of convex functions using a network of computing units. We investigate this problem under two communication schemes (centralized and decentralized) and four classical regularity assumptions: Lipschitz continuity, strong convexity, smoothness, and a combination of strong convexity and smoothness. Under the decentralized communication scheme, we provide matching upper and lower bounds of complexity along with algorithms achieving this rate up to logarithmic constants. For non-smooth objective functions, while the dominant term of the error is in $O(1/\sqrt{t})$, the structure of the communication network only impacts a second-order term in $O(1/t)$, where tt is time. In other words, the error due to limits in communication resources decreases at a fast rate even in the case of non-strongly convex objective functions. Such a convergence rate is achieved by the novel multi-step primal-dual (MSPD) algorithm. Under the centralized communication scheme, we show that the naive distribution of standard optimization algorithms is optimal for smooth objective functions, and provide a simple yet efficient algorithm called distributed randomized smoothing (DRS) based on a local smoothing of the objective function for non-smooth functions. We then show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where d is the underlying dimension.

38. Accelerated Decentralized Optimization with Local Updates for Smooth and Strongly Convex Objectives [31] In this paper, we study the problem of minimizing a sum of smooth and strongly convex functions split over the nodes of a network in a decentralized fashion. We propose the algorithm $ESDACD$, a decentralized accelerated algorithm that only requires local synchrony. Its rate depends on the condition number κ of the local functions as well as the network topology and delays. Under mild assumptions on the topology of the graph, $ESDACD$ takes a time $O((\tau_{\max} + \Delta_{\max})\sqrt{\kappa/\gamma}\ln(\epsilon^{-1}))$ to reach a precision ϵ where γ is the spectral gap of the graph, $\tau_{\max}$ the maximum communication delay and $\Delta_{\max}$ the maximum computation time. Therefore, it matches the rate of $SSDA$, which is optimal when $\tau_{\max} = \Omega\left(\Delta_{\max}\right)$. Applying $ESDACD$ to quadratic local functions leads to an accelerated randomized gossip algorithm of rate $O(\sqrt{\theta_{\text{gossip}}/n})$ where θ_{gossip} is the rate of the standard randomized gossip. To the best of our knowledge, it is the first asynchronous gossip algorithm with a provably improved rate of convergence of the second moment of the error. We illustrate these results with experiments in idealized settings.

39. An Accelerated Decentralized Stochastic Proximal Algorithm for Finite Sums [49] Modern large-scale finite-sum optimization relies on two key aspects: distribution and stochastic updates. For smooth and strongly convex problems, existing decentralized algorithms are slower than modern accelerated variance-reduced stochastic algorithms when run on a single machine, and are therefore not efficient. Centralized algorithms are fast, but their scaling is limited by global aggregation steps that result in communication bottlenecks. In this work, we propose an efficient **A**ccelerated, **D**ecentralized stochastic algorithm for **F**inite**S**ums named ADFS, which uses local stochastic proximal updates and randomized pairwise communications between nodes. On machines, ADFS learns from samples in the same time it takes optimal algorithms to learn from samples on one machine. This scaling holds until a critical network size is reached, which depends on communication delays, on the number of samples , and on the network topology. We provide a theoretical analysis based on a novel augmented graph approach combined with a precise evaluation of synchronization times and an extension of the accelerated proximal coordinate gradient algorithm to arbitrary sampling. We illustrate the improvement of ADFS over state-of-the-art decentralized approaches with experiments.

7.6. Stochastic Geometry

40. On the Dimension of Unimodular Discrete Spaces, Part I: Definitions and Basic Properties [39] This work introduces two new notions of dimension, namely the *unimodular Minkowski and Hausdorff dimensions*, which are inspired from the classical analogous notions. These dimensions are defined for *unimodular discrete spaces*, introduced in this work, which provide a common generalization to stationary point processes under their Palm version and unimodular random rooted graphs. The use of unimodularity in the definitions of dimension is novel. Also, a toolbox of results is presented for the analysis of these dimensions. In particular, analogues of Billingsley's lemma and Frostman's lemma are presented. These lemmas are instrumental in deriving upper bounds on dimensions, whereas lower bounds are obtained from specific coverings. The notions of unimodular Hausdorff measure and unimodular dimension function are also introduced. This toolbox is used to connect the unimodular dimensions to various other notions such as growth rate, scaling limits, discrete dimension and amenability. It is also used to analyze the dimensions of a set of examples pertaining to point processes, branching processes, random graphs, random walks, and self-similar discrete random spaces.

41. On the Dimension of Unimodular Discrete Spaces, Part II: Relations with Growth Rate [40] The notions of unimodular Minkowski and Hausdorff dimensions are defined in [39] for unimodular random discrete metric spaces. This work is focused on the connections between these notions and the polynomial growth rate of the underlying space. It is shown that bounding the dimension is closely related to finding suitable equivariant weight functions (i.e., measures) on the underlying discrete space. The main results are unimodular versions of the mass distribution principle and Billingsley's lemma, which allow one to derive upper bounds on the unimodular Hausdorff dimension from the growth rate of suitable equivariant weight functions. Also, a unimodular version of Frostman's lemma is provided, which shows that the upper bound given by the unimodular Billingsley lemma is sharp. These results allow one to compute or bound both types of unimodular dimensions in a large set of examples in the theory of point processes, unimodular random graphs, and self-similarity. Further results of independent interest are also presented, like a version of the max-flow min-cut theorem for unimodular one-ended trees.

42. Doeblin trees [4] This work is centered on the random graph generated by a Doeblin-type coupling of discrete time processes on a countable state space whereby when two paths meet, they merge. This random graph is studied through a novel subgraph, called a bridge graph, generated by paths started in a fixed state at any time. The bridge graph is made into a unimodular network by marking it and selecting a root in a specified fashion. The unimodularity of this network is leveraged to discern global properties of the larger Doeblin graph. Bi-recurrence, i.e., recurrence both forwards and backwards in time, is introduced and shown to be a key property in uniquely distinguishing paths in the Doeblin graph, and also a decisive property for Markov chains indexed by $\mathbb{Z}$. Properties related to simulating the bridge graph are also studied.

43. The Stochastic Geometry of Unconstrained One-Bit Compression [5] A stationary stochastic geometric model is proposed for analyzing the data compression method used in one-bit compressed sensing. The data set is an unconstrained stationary set, for instance all of $\mathbb{R}^n$ or a stationary Poisson point process in $\mathbb{R}^n$. It is compressed using a stationary and isotropic Poisson hyperplane tessellation, assumed independent of the data. That is, each data point is compressed using one bit with respect to each hyperplane, which is the side of the hyperplane it lies on. This model allows one to determine how the intensity of the hyperplanes must scale with the dimension n to ensure sufficient separation of different data by the hyperplanes as well as sufficient proximity of the data compressed together. The results have direct implications in compressed sensing and in source coding.

44. Limit theory for geometric statistics of point processes having fast decay of correlations [7] We develop a limit theory (Laws of Large Numbers and Central Limit Theorems) for functionals of spatially correlated point processes. The "strength" of data correlation is captured and controlled by the speed of decay of the additive error in the asymptotic factorization the correlation functions, when the separation distance increases. In this way, the classical theory of Poisson and Bernoulli processes is extended to a larger class of data inputs, such as determinantal point processes with fast decreasing kernels, including the α-Ginibre ensembles, permanental point processes as well as the zero set of Gaussian entire functions. Both linear (U-statistics) and 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) are considered.

7.7. Information theory

45. Error Exponents for MAC Channelss [3] This work analyzes a class of Multiple Access Channels (MAC) where the sum of the dimensions of the transmitted signals matches that of the received signal. This channel is a classical object of information theory in the power constrained case. We first focus on the Poltyrev regime, namely the case without power constraint. Using point process techniques, we derive the capacity under general stationarity and ergodicity noise assumptions as well as a representation of the error probability. We use this to derive bounds on the error exponent in the Gaussian case. This also leads to new results on the power constrained error exponents.

EVA Project-Team

7. New Results

7.1. Falco startup launched!

Participants: Elsa Nicol, Keoma Brun-Laguna, Thomas Watteyne.

The Falco startup (https://wefalco.com/) was launched on 14 January 2019. During 2019, it developed a complete technical solution (PCB, assembly, networking, back-end) and completed a large market development campaign. Falco was selected to join the Parisian Incubator Agoranov. It was then awarded the prestigous Netva "Deeptech North America" program, and won the Favorite Startup Pitch battle at MassChallenge, Boston, as well as the Amplify Pitch battle. It had a booth at the Cap d'Agde and Paris Nautic shows. On 14 December 2019, Falco wins the Innovation Competition at the Paris Nautic Show.

7.2. 6TiSCH Standardization

Participants: Malisa Vucinic, Jonathan Muñoz, Tengfei Chang, Yasuyuki Tanaka, Thomas Watteyne.

The standardization work at 6TiSCH remains a strong federator of the work done in the team. In 2019, the working group finalized the work on the draft-ietf-6tisch-minimal-security and draft-ietf-6tisch-architecture specification, which are both in the editor queue. The draft-ietf-6tisch-msf has also passed the working group last call. This standardization work has resulted in several papers on 6TiSCH, including a tutorial [9], [11], [12] fragmentation in 6TiSCH [8], implementation details [17], simulating 6TiSCH [24], experimental approaches [21], [26], localization [25], multi-PHY extensions [10]. The HDR of Thomas Watteyne [2] reports on the work on 6TiSCH over the past years. Some work has started on implementing 6TiSCH on single-chip micro-motes [19], [23], [7], [18].

7.3. 6TiSCH Security

Participants: Malisa Vucinic, Thomas Watteyne.

The security work of Inria-EVA is a continuation of the efforts started during the H2020 ARMOUR project. The work focused on stabilizing the "Minimal Security" solution that has now been approved to be published as an RFC [13]. The solution that is standardized enables secure network access and configuration of 6TiSCH devices under the assumption that they have been provisioned with a secret key. Ongoing work extends this solution to support true zero-configuration network setup, under the assumption that the devices have been provisioned with certificates at manufacturing time.

7.4. 6TiSCH Benchmarking

Participants: Malisa Vucinic, Tengfei Chang, Yasuyuki Tanaka, Thomas Watteyne.

With the pure 6TiSCH standardizes coming to an end, the focus of the group is moving towards benchmarking how well it works. This has resulted in the following action. Although seemingly different, they all contribute to the overall goal of better understanding (the performance of) 6TiSCH.

We have built and put online the OpenTestbed, a collection of 80 OpenMote B boards deployed in 20 "pods". These allow us to test the performance of the OpenWSN firmware in a realistic setting. You can access its management interface at http://testbed.openwsn.org/.

A tool complementary to the testbed is the 6TiSCH simulator (https://bitbucket.org/6tisch/simulator) which Yatsuyiki Tanaka is leading. The simulator now represents exactly the behavior of the 6TiSCH protocol stack, and has been a catalyst for benchmarking activities around 6TiSCH.

Beyond Inria, the benchmarking activity around 6TiSCH is a hot topic, with projects such as the 6TiSCH Open Data Action [26] (SODA, http://www.soda.ucg.ac.me/), the IoT Benchmarks Initiative (https://www.iotbench.ethz.ch/), and the Computer and Networking Experimental Research using Testbeds (CNERT) workshop at INFOCOM, all of which Inria-EVA is very involved in.

7.5. LAKE Standardization

Participants: Malisa Vucinic, Timothy Claeys, Thomas Watteyne.

In October 2019, a new working group was formed in the IETF with the goal of standardizing a lightweight authenticated key exchange protocol for IoT use cases. The group is co-chaired by Malisa Vucinic of Inria-EVA. Through our work in 6TiSCH and the requirements for the follow up work of the “Minimal Security Framework for 6TiSCH”, we directly contributed to the creation of this working group whose expected output is the key exchange protocol for the IoT. The document we lead in the LAKE working group [37] compiles the requirements for a lightweight authenticated key exchange protocol for OSCORE. OSCORE (RFC8613) is a lightweight communication security protocol providing end-to-end security on application layer for constrained IoT settings. It is expected to be deployed with standards and frameworks using CoAP such as 6TiSCH, LPWAN, OMA Specworks LwM2M, Fairhair Alliance and Open Connectivity Foundation.

7.6. IoT and Low-Power Wireless Meshed Networks

More than 50 billion devices will be connected in 2020. This huge infrastructure of devices, which is managed by highly developed technologies, is called the Internet of Things (IoT). The IoT provides advanced services, and brings economic and societal benefits. This is the reason why engineers and researchers in both industry and scientific communities are interested in this area. The Internet of Things enables the interconnection of smart physical and virtual objects, managed by highly developed technologies. Low-Power Wireless Meshed Network is an essential part of this paradigm. It uses smart, autonomous and usually limited capacity devices in order to sense and monitor their environment.

7.6.1. Centralized or Distributed Scheduling for IEEE 802.15.4e TSCH networks

Participants: Yasuyuki Tanaka, Pascale Minet, Thomas Watteyne, Malisa Vucinic, Tengfei Chang, Keoma Brun-Laguna.

The wireless TSCH (Time Slotted Channel Hopping) network specified in the e amendment of the IEEE 802.15.4 standard has many appealing properties. Its schedule of multichannel slotted data transmissions ensures the absence of collisions. Because there is no retransmission due to collisions, communication is faster. Since the devices save energy each time they do not take part in a transmission, the power autonomy of nodes is prolonged. Furthermore, channel hopping mitigates multipath fading and interferences.

All communication in a TSCH network is orchestrated by the communication schedule it is using. The scheduling algorithm used hence drives the latency and capacity of the network, and the power consumption of the nodes. To increase the flexibility and the self-organizing capacities required by IoT, the networks have to be able to adapt to changes. These changes may concern the application itself, the network topology by adding or removing devices, the traffic generated by increasing or decreasing the device sampling frequency, for instance. That is why flexibility of the schedule ruling all network communications is needed. We have designed a number of scheduling algorithms for TSCH networks, answering different needs. For instance, the centralized Load-based scheduler that assigns cells per flow, starting with the flow originating from the most loaded node has proved optimal for many configurations. Simulations with the 6TiSCH simulator showed that it gets latencies close to the optimal. They also highlighted that end-to-end latencies are positively impacted by message prioritization (i.e. each node transmits the oldest message first) at high loads, and negatively impacted by unreliable links, as presented at GlobeCom 2019 [30].

Among the distributed scheduling algorithms proposed in the literature, many rely on assumptions that may be violated by real deployments. This violation usually leads to conflicting transmissions of application data, decreasing the reliability and increasing the latency of data delivery. Others require a processing complexity that cannot be provided by sensor nodes of limited capabilities. Still others are unable to adapt quickly to traffic or topology changes, or are valid only for small traffic loads. We have designed MSF and YSF, two distributed scheduling algorithms that are adaptive and compliant with the standardized protocols used in the 6TiSCH working group at IETF. The Minimal Scheduling Function (MSF) is a distributed scheduling algorithm in which neighbor nodes locally negotiate adding and removing cells. MSF was evaluated by simulation and experimentation, before becoming the default scheduling algorithm of the IETF 6TiSCH working group, and now an official standard. We also designed LLSF, a scheduling algorithm focused on low latency communication. We proposed a full-featured 6TiSCH scheduling function called YSF, that autonomously takes into account all the aspects of network dynamics, including the network formation phase and parent switches. YSF aims at minimizing latency and maximizing reliability for data gathering applications. Simulation results obtained with the 6TiSCH simulator show that YSF yields lower end-to-end latency and higher end-to-end reliability than MSF, regardless of the network topology. Unlike other top-down scheduling functions, YSF does not rely on any assumption regarding network topology or traffic load, and is therefore more robust in real network deployments. An intensive simulation campaign made with the 6TiSCH simulator has provided comparative performance results. Our proposal outperforms MSF, the 6TiSCH Minimal Scheduling Function, in terms of end-to-end latency and end-to-end packet delivery ratio.

Furthermore we published additional research on computing the upper bounds on the end-to-end latency, finding the best trade-off between latency and network lifetime.

7.6.2. *Modeling and Improving Named Data Networking over IEEE 802.15.4*

Participants: Amar Abane, Samia Bouzefrane (Cnam), Paul Muhlethaler.

Enabling Named Data Networking (NDN) in real world Internet of Things (IoT) deployments becomes essential to benefit from Information Centric Networking (ICN) features in current IoT systems. One objective of the model is to show that caching can attenuate the number of transmissions generated by broadcast to achieve a reasonable overhead while keeping the data dissemination power of NDN. To design realistic NDN-based communication solutions for IoT, revisiting mainstream technologies such as low-power wireless standards may be the key. We explore the NDN forwarding over IEEE 802.15.4 by modeling a broadcast-based forwarding [27]. Based on the observations, we adapt the Carrier-Sense Multiple Access (CSMA) algorithm of 802.15.4 to improve NDN wireless forwarding while reducing broadcast effects in terms of packet redundancy, round-trip time and energy consumption. As future work, we aim to explore more complex CSMA adaptations for lightweight forwarding to make the most of NDN and design a general-purpose Named-Data CSMA.

7.6.3. *Evaluation of LORA with stochastic geometry*

Participants: Bartek Blaszczyszyn (Dyogene), Paul Muhlethaler.

We present a simple, stochastic-geometric model of a wireless access network exploiting the LoRA (Long Range) protocol, which is a non-expensive technology allowing for long-range, single-hop connectivity for the Internet of Things. We assume a space-time Poisson model of packets transmitted by LoRA nodes to a fixed base station. Following previous studies of the impact of interference, we assume that a given packet is successfully received when no interfering packet arrives with similar power before the given packet payload phase, see [16]. This is as a consequence of LoRa using different transmission rates for different link budgets (transmissions with smaller received powers use larger spreading factors) and LoRa intra-technology interference treatment. Using our model, we study the scaling of the packet reception probabilities per link budget as a function of the spatial density of nodes and their rate of transmissions. We consider both the parameter values recommended by the LoRa provider, as well as proposing LoRa tuning to improve the equality of performance for all link budgets. We also consider spatially non-homogeneous distributions of LoRa nodes. We show how a fair comparison to non-slotted Aloha can be made within the same framework.

7.6.4. *Position Certainty Propagation: A location service for MANETs*

Participants: Abdallah Sobehy, Paul Muhlethaler, Eric Renault (Telecom Sud-Paris).

A location method based on triangulation (via Channel State Information (CSI) based localization method is proposed [6]. A known method of triangulation is adopted to deduce the location of a node from 3 reference nodes (anchor nodes). We propose an optimized energy-aware and low computational solution, requiring 3-GPS equipped nodes (anchor nodes) in the network. Moreover, the computations are lightweight and can be implemented distributively among nodes. Knowing the maximum range of communication for all nodes and distances between 1-hop neighbors, each node localizes itself and shares its location with the network in an efficient manner. We simulate our proposed algorithm on a NS-3 simulator, and compare our solution with state-of-the-art methods. Our method is capable of localizing more nodes i.e. $\simeq 90\%$ of nodes in a network with an average degree $\simeq 10$.

7.7. Industry 4.0 and Low-Power Wireless Meshed Networks

The Internet of Things (IoT) connects tiny electronic devices able to measure a physical value (temperature, humidity, etc.) and/or to actuate on the physical world (pump, valve, etc). Due to their cost and ease of deployment, battery-powered wireless IoT networks are rapidly being adopted.

The promise of wireless communication is to offer wire-like connectivity. Major improvements have been made in that direction, but many challenges remain as industrial applications have strong operational requirements. This section of the IoT application is called Industrial IoT (IIoT).

By the year 2020, it is expected that the number of connected objects will exceed several billion devices. These objects will be present in everyday life for a smarter home and city as well as in future smart factories that will revolutionize the industry organization. This is actually the expected fourth industrial revolution, better known as Industry 4.0. In which, the Internet of Things (IoT) is considered as a key enabler for this major transformation. The IoT will allow more intelligent monitoring and self-organizing capabilities than traditional factories. As a consequence, the production process will be more efficient and flexible with products of higher quality.

To produce better quality products and improve monitoring in Industry 4.0, strong requirements in terms of latency, robustness and power autonomy have to be met by the networks supporting the Industry 4.0 applications.

7.7.1. *Reliability for the Industrial Internet of Things (IIoT) and Industry 4.0*

Participants: Yasuyuki Tanaka, Pascale Minet, Keoma Brun-Laguna, Thomas Watteyne.

The main IIoT requirement is reliability. Every bit of information that is transmitted in the network must not be lost. Current off-the-shelf solutions offer over 99.999% reliability.

To provide the end-to-end reliability targeted by industrial applications, we investigate an approach based on message retransmissions (on the same path). We propose two methods to compute the maximum number of transmissions per message and per link required to achieve the targeted end-to-end reliability. The MFair method is very easy to compute and provides the same reliability over each link composing the path, by means of different maximum numbers of transmissions, whereas the MOpt method minimizes the total number of transmissions necessary for a message to reach the sink. MOpt provides a better reliability and a longer lifetime than MFair, which provides a shorter average end-to-end latency. This study [5] was published in the Sensors journal in 2019.

7.8. Machine Learning applied to Networking

7.8.1. *Machine Learning for energy-efficient and QoS-aware Data Centers*

Participants: Ruben Milocco (Comahue University, Argentina, Invited Professor), Pascale Minet, Eric Renault (Telecom Sud-Paris), Selma Boumerdassi (Cnam).

To limit global warming, all industrial sectors must make effort to reduce their carbon footprint. Information and Communication Technologies (ICTs) alone generate 2% of global CO2 emissions every year. Due to the rapid growth in Internet services, data centers have the largest carbon footprint of all ICTs. According to ARCEP (the French telecommunications regulator), Internet data traffic multiplied by 4.5 between 2011 and 2016. In order to support such a growth and maintain this traffic, data centers'energy consumption needs to be optimized.

We determine whether resource allocation in DCs can satisfy the three following requirements: 1) meet user requirements (e.g. short response times), 2) keep the data center efficient, and 3) reduce the carbon footprint.

An efficient way to reduce the energy consumption in a DC is to turn off servers that are not used for a minimum duration. The high dynamicity of the jobs submitted to the DC requires periodically adjusting the number of active servers to meet job requests. This is called Dynamic Capacity Provisioning. This provisioning can be based on prediction. In such a case, a proactive management of the DC is performed. The goal of this study is to provide a methodology to evaluate the energy cost reduction brought by proactive management, while keeping a high level of user satisfaction.

The state-of-the art shows that appropriate proactive management improves the cost, either by improving QoS or saving energy. As a consequence, there is great interest in studying different proactive strategies based on predictions of either the energy or the resources needed to serve CPU and memory requests. The cost depends on 1) the proactive strategy used, 2) the workload requested by jobs and 3) the prediction used. The problem complexity explains why, despite its importance, the maximum cost savings have not been evaluated in theoretical studies.

We propose a method to compute the upper bound of the relative cost savings obtained by proactive management compared to a purely reactive management based on the Last Value. With this method, it becomes possible to quantitatively compare the efficiency of two predictors.

We also show how to apply this method to a real DC and how to select the value of the DC parameters to get the maximum cost savings. Two types of predictors are studied: linear predictors, represented by the ARMA model, and nonlinear predictors obtained by maximizing the conditional probability of the next sample, given the past. They are both applied to the publicly available Google dataset collected over a period of 29 days. We evaluate the largest benefit that can be obtained with those two predictors. Some of these results have been presented at HPCS 2019 [20].

7.8.2. Machine Learning applied to IoT networks

Participants: Miguel Landry Foko Sindjoung (Phd Student, Dschang University, Cameroon, Inria Internship), Pascale Minet.

Knowledge of link quality in IoT networks allows a more accurate selection of wireless links to build the routes used for data gathering. The number of re-transmissions is decreased, leading to shorter end-to-end latency, better end-to-end reliability and a longer network lifetime.

We propose to predict link quality by means of machine learning techniques applied on two metrics: the Received Signal Strength Indicator (RSSI) and the Packet Delivery Ratio (PDR). These two metrics were selected because RSSI is a hardware metric that is easily obtained and PDR takes into account packets that are not successfully received, unlike RSSI.

The data set used in this study was collected from a TSCH network deployed in the Grenoble testbed consisting of 50 nodes operating on 16 channels. Data collected by Mercator include 108659 measurements of PDR and average RSSI. We train the model over the training set and predict the link quality on the channel considered for the samples in the validation set. By comparing the predicted values with the real values, the confusion matrix is computed by evaluating the number of true-positive, true-negative, false-positive and false-negative for the link and channel considered.

Whatever the link quality estimator used, RSSI, PDR or both, the Random Forest (RF) classifier model outperforms the other models studied: Linear Regression, Linear Support Vector Machine, Support Vector Machine.

Since using Bad links that have been predicted Good strongly penalizes network performance in terms of end-to-end latency, end-to-end reliability and network lifetime, the joint use of PDR and RSSI improves the accuracy of link quality prediction. Hence, we recommend using the Random Forest classifier applied on both PDR and RSSI metrics. This work has been presented at the PEMWN 2019 conference [33].

7.9. Machine Learnig applied to Smart Farming

Participants: Jamal Ammouri (Internship Cnam), Malika Boudiaf (Ummto, Tizi-Ouzou, Algeria), Samia Bouzefrane (Cnam), Pascale Minet, Meziane Yacoub (Cnam).

Intelligent Farming System (IFS) is made possible by the use of 4 elements: sensors and actuators, the Internet of Things (IoT), edge/cloud processing, and machine learning.

Soil degradation and a hot climate explain the poor yield of olive groves in North Algeria. Edaphic, climatic and geographical data were collected from 10 olive groves over several years and analyzed by means of Self-Organizing Maps (SOMs). SOM is a non-supervised neural network that projects high-dimensional data onto a low-dimension discrete space, called a topological map, such that close data are mapped onto nearby locations on the map. In the paper [28] presented at the PEMWN 2019 conference, we have shown how to use self-organizing maps to determine olive grove clusters with similar features, characterize each cluster and show the temporal evolution of each olive grove. With the SOM, it becomes possible to alert the farmer when some specific action needs to be done in the case of hydric stress, NPK stress, pest/disease attack. As a result, the nutritional quality of the oil produced is improved. SOM can be integrated in the Intelligent Farming System (IFS) to boost conservation agriculture.

This work requires a strong collaboration with agronomists. Malika Boudiaf (Laboratoire Ressources Naturelles, UMMTO, Tizi-Ouzou, Algeria) provided the data set and gave us many explanations about soil conservation. Meziane Yacoub (Cnam) is an expert in SOMs. Jamal Ammouri (Cnam) was co-advised by Samia Bouzefrane, Pascale Minet and Meziane Yacoub.

7.10. Protocols and Models for Wireless Networks - Application to VANETs

7.10.1. Connection-less IoT - Protocol and models

Participants: Iman Hemdoush, Cédric Adjih, Paul Mühlethaler.

The goal is to construct some next-generation access protocols, for the IoT (or alternately for vehicular networks). One starting point are methods from the family of Non-Orthogonal Multiple Access (NOMA), where multiple transmissions can "collide" but can still be recovered - with sophisticated multiple access protocols (MAC) that take the physical layer/channel into account. One such example is the family of the Coded Slotted Aloha methods.Another direction is represented by some vehicular communications where vehicles communicate directly with each other without necessarily going through the infrastructure. This is also true more generally in any wireless network where the control is relaxed (such as in unlicensed IoT networks like LoRa). One observation is that in such distributed scenarios, explicit or implicit forms of signaling (with sensing, messaging, etc.), can be used for designing sophisticated protocols - including using machine learning techniques.

During this study, some of the following tools should be used: protocol/algorithm design (ensuring properties by construction), simulations (ns-2, ns-3, matlab, ...) on detailed or simplified network models, mathematical modeling (stochastic geometry, etc...) ; machine-learning techniques or modeling as code-on-graphs.

The first result we have obtained concerns Irregular Repetition Slotted Aloha (IRSA) which is a modern method of random access for packet networks that is based on repeating transmitted packets, and on successive interference cancellation at the receiver. In classical idealized settings of slotted random access protocols (where slotted ALOHA achieves $1/e$), it has been shown that IRSA could asymptotically achieve the maximal throughput of 1 packet per slot. Additionally, IRSA had previously been studied for many different variants and settings, including the case where the receiver is equipped with “multiple-packet reception” (MPR) capability. We extensively revisit the case of IRSA with MPR. We present a method to compute optimal

IRSA degree distributions with a given maximum degree n. A tighter bound for the load threshold (G/K) was proven, showing that plain K-IRSA cannot reach the asymptotic known bound $G/K = 1$ for $K > 1$, and we prove a new, lower bound for its performance. Numerical results illustrate that optimal degree distributions can approach this bound. Second, we analyze the error floor behavior of K-IRSA and provide an insightful approximation of the packet loss rate at low loads, and show its excellent performance. Third, we show how to formulate the search for the appropriate parameters of IRSA as an optimization problem, and how to solve it efficiently. By doing that for a comprehensive set of parameters, and by providing this work with simulations, we give numerical results that shed light on the performance of IRSA with MPR. A final open question is: what is the impact of introducing more structure in the slot selection (like Spatially Coupled Coded Slotted Aloha) and how best to do so?

7.10.2. Indoor positionning using Channel State Information (CSI) from a MIMO antenna

Participants: Abdallah Sobehy, Paul Muhlethaler, Eric Renault (Telecom Sud-Paris).

The channel status information is used for locating a node by applying machine learning [35] techniques. We propose a novel lightweight deep learning solution to the indoor positioning problem based on noise and dimensionality reduction of MIMO Channel State Information (CSI): real and imaginary parts of the signal received. Based on preliminary data analysis, the magnitude of the CSI is selected as the input feature for a Multilayer Perceptron (MLP) neural network. Polynomial regression is then applied to batches of data points to filter noise and reduce input dimensionality by a factor of 14. The MLP's hyper-parameters are empirically tuned to achieve the highest accuracy. The method is applied to a CSI dataset estimated at an 8 x 2 MIMO antenna that is published by the organizers of the Communication Theory Workshop Indoor Positioning Competition. The proposed solution is compared with a state-of-the-art method presented by the authors who designed the MIMO antenna that is used to generate the data-set. Our method yields a mean error which is 8 times less than that of its counterpart. We conclude that the arithmetic mean and standard deviation misrepresent the results since the errors follow a log- normal distribution. The mean of the log error distribution of our method translates to a mean error as low as 1.5 cm. We have shown that, using a K-nearest neighbor learning method an even better, indoor positioning is achieved. The input feature is the magnitude component of CSI which is pre-processed to reduce noise and allow for a quicker search. The Euclidean distance between CSI is the criterion chosen for measuring the closeness between samples. The proposed method is compared with three other methods, all based on deep learning approaches and tested with the same data-set. The K-nearest neighbor method presented in this paper achieves a Mean Square Error (MSE) of 2.4 cm, which outperforms its counterparts.

7.10.3. Predicting Vehicles Positions using Roadside Units: a Machine-Learning Approach

Participants: Samia Bouzefrane (Cnam), Soumya Banerjee (Birla Institute Of Technology, Mesra), Paul Mühlethaler, Mamoudou Sangare.

We study positioning systems using Vehicular Ad Hoc Networks (VANETs) to predict the position of vehicles. We use the reception power of the packets received by the Road Side Units (RSUs) and sent by the vehicles on the roads. In fact, the reception power is strongly influenced by the distance between a vehicle and a RSU. We have already used and compared three widely recognized techniques : K Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest. We have studied these techniques in various configurations and discuss their respective advantages and drawbacks. We revisit the positioning problem VANETs but we also consider Neural Networks (NN) to predict the position [22]. The neural scheme we have tested in this paper consists of one hidden layer with three neurons. To boost this technique we use an ensemble neural network with 50 elements built with a bagging algorithm. The numerical experiments presented in this contribution confirm that a precise prediction can only be obtained when there is a main direct path of propagation. The prediction is altered when the training is incomplete or less precise but the precision remains acceptable. In contrast, with Rayleigh fading, the accuracy obtained is much less striking. We observe that the Neural Network is nearly always the best approach. With a direct path the ranking is: Neural Network, Random Forest, KNN and SVM except in the case when we have no measurement in [30m; 105m] where the ranking is Neural Network, Random Forest, SVM and KNN. When there is no direct path, the ranking is SVM, NN, RF and KNN but the difference in performance between SVM and NN is small.

7.10.4. Combining random access TDMA scheduling strategies for vehicular ad hoc networks

Participants: Fouzi Boukhalfa, Mohamed Hadded (Vedecom), Paul Mühlethaler, Oyunchimeg Shagdar (Vedecom).

This work is based on Fouzi Boukhalfa's PhD which started in October 2018, [29],[15]. The idea is to combine TDMA protocols with random access techniques to benefit from the advantages of both techniques. Fouzi Boukhalfa proposes to combine the DTMAC protocol introduced by Mohamed Hadded with a generalization of CSMA. This generalized CSMA uses active signaling; the idea is to send signaling bursts in order to select a unique transmitter. The protocol that Fouzi Boukhalfa obtains reduces the access and merging collisions of DTMAC but can also propose access with low latency for emergency traffic. The idea is that vehicles access their slots reserved with DTMAC but the transmission slots encompass a special section at the beginning with active signaling. The transmission of the signaling burst, during a mini-slot, is organized according to a random binary key. A '1' in the key means that a signaling burst will be transmitted, while a '0' means that the vehicle senses the channel on this mini-slot to potentially find the transmission of a signaling burst by another vehicle. Fouzi Boukhalfa shows that if we use a random key to transmit the signaling burst it very significantly decreases the collision rate (both merging and access collisions) and that emergency traffic can have a very small access delay. Fouzi Boukhalfa builds an analytical model which thoroughly confirms the simulation result. This model can encompass detection error in the selection process of the signaling bursts. It is shown that with a reasonable error rate the performance is only marginaly affected.

7.10.5. Forecasting traffic accidents in VANETs

Participants: Samia Bouzefrane (Cnam), Soumya Banerjee (Birla Institute Of Technology, Mesra), Paul Mühlethaler, Mamoudou Sangare.

Road traffic accidents have become a major cause of death. With increasing urbanization and populations, the volume of vehicles has increased exponentially. As a result, traffic accident forecasting and the identification of the accident prone areas can help reduce the risk of traffic accidents and improve the overall life expectancy.

Conventional traffic forecasting techniques use either a Gaussian Mixture Model (GMM) or a Support Vector Classifier (SVC) to model accident features. A GMM on the one hand requires large amount of data and is computationally inexpensive, SVC on the other hand performs well with less data but is computationally expensive. We present a prediction model that combines the two approaches for the purpose of forecasting traffic accidents. A hybrid approach is proposed, which incorporates the advantages of both the generative (GMM) and the discriminant model (SVC). Raw feature samples are divided into three categories: those representing accidents with no injuries, accidents with non incapacitating injuries and those with incapacitating injuries. The output or the accident severity class was divided into three major categories namely: no injury in the accident, non-incapacitating injury in the accident and an incapacitating injury in the accident. A hybrid classifier is proposed which combines the descriptive strength of the baseline Gaussian mixture model (GMM) with the high performance classification capabilities of the support vector classifier (SVC). A new approach is introduced using the mean vectors obtained from the GMM model as input to the SVC. The model was supported with data pre-processing and re-sampling to convert the data points into suitable form and avoid any kind of biasing in the results. Feature importance ranking was also performed to choose relevant attributes with respect to accident severity. This hybrid model successfully takes advantage of both models and obtained a better accuracy than the baseline GMM model. The radial basis kernel outperforms the linear kernel by achieving an accuracy of 85.53%. Data analytics performed including the area under the receiver operating characteristics curve (AUC-ROC) and area under the precision/recall curve(AUC-PR) indicate the successful application of this model in traffic accident forecasting. Experimental results show that the proposed model can significantly improve the performance of accident prediction. Improvements of up to 24% are reported in the accuracy as compared to the baseline statistical model (GMM). The data about circumstances of personal injury in road accidents, the types of vehicles involved and the consequential casualties were obtained from data.govt.uk.

Although a significant improvement in accuracy has been observed, this study has several limitations. The first concerns the dataset used. This research is based on a road traffic accident dataset from the year of 2017 which contains very few data samples for the no injury and non-incapacitating injury types of accident. The data was unbalanced not just with respect to the output class but also with respect to the sub features of various attributes. Moreover, aggregating the accident severity into just three categories limits the scope of the study and the results obtained. The greater the number of severity classes, the less is the amount of extra training data required to feed in the SVC to avoid overfitting. Thus, datasets with sufficient records corresponding to each class are desirable and must be used for further study.

The second limitation concerns the dependence of the SVC model on parameters and attribute selection. In this study, the performance of SVC relies heavily on the feature selection results and the mean vectors obtained from the GMM. In order to improve the accuracy of the support vector classifier, other approaches like particle swarm optimization (PSO), ant colony optimization, genetic algorithms etc. could be used for effective parameter selection. In addition to this, more kernels like the polynomial kernel and the sigmoid kernel could be tested to improve future model performances.

FUN Project-Team

6. New Results

6.1. Security and Verification

Participants: Rehan Malak, Allan Blanchard, Antoine Gallais, Valeria Loscri, Nathalie Mitton.

6.1.1. Security

Numerous medium access control (MAC) have been proposed for Low-power Lossy Networks (LLNs) over the recent years. They aim at ensuring both energy efficiency and robustness of the communication transmissions. Nowadays, we observe deployments of LLNs for potentially critical application scenarios (e.g., plant monitoring, building automation), which require both determinism and security guarantees. They involve battery-powered devices which communicate over lossy wireless links. Radio interfaces are turned off by a node as soon as no traffic is to be sent or relayed. Denial-of-sleep attacks consist in exhausting the devices by forcing them to keep their radio on. In [21], we focus on jamming attacks whose impact can be mitigated by approaches such as time-division and channel hopping techniques. We use the IEEE 802.15.4e standard to show that such approaches manage to be resistant to jamming but yet remain vulnerable to selective jamming. We discuss the potential impacts of such onslaughts, depending on the knowledge gained by the attacker, and to what extent envisioned protections may allow jamming attacks to be handled at upper layers.

6.1.2. Verification

Modern verification projects continue to offer new challenges for formal verification. One of them is the linked list module of Contiki, a popular open-source operating system for the Internet of Things. It has a rich API and uses a particular list representation that make it different from the classical linked list implementations. Being widely used in the OS, the list module is critical for reliability and security. A recent work verified the list module using ghost arrays. In [17], [35], we report on a new verification effort for this module. Realized in the Frama-C/Wp tool, the new approach relies on logic lists. A logic list provides a convenient high-level view of the linked list. The specifications of all functions are now proved faster and almost all automatically, only a small number of auxiliary lemmas and a couple of assertions being proved interactively in Coq. The proposed specifications are validated by proving a few client functions manipulating lists. During the verification, a more efficient implementation for one function was found and verified. We compare the new approach with the previous effort based on ghost arrays, and discuss the benefits and drawbacks of both techniques.

While deductive verification is increasingly used on real-life code, making it fully automatic remains difficult. The development of powerful SMT solvers has improved the situation, but some proofs still require interactive theorem provers in order to achieve full formal verification. Auto-active verification relies on additional guiding annotations (assertions, ghost code, lemma functions, etc.) and provides an important step towards a greater automation of the proof. However, the support of this methodology often remains partial and depends on the verification tool. [18] presents an experience report on a complete functional verification of several C programs from the literature and real-life code using auto-active verification with the C software analysis platform Frama-C and its deductive verification plugin . The goal is to use automatic solvers to verify properties that are classically verified with interactive provers. Based on our experience, we discuss the benefits of this methodology and the current limitations of the tool, as well as proposals of new features to overcome them.

6.2. Visible Light Communication

Participants: Antonio Costanzo, Valeria Loscri.

Visible Light Communication (VLC) exploits optical frequencies, diffused by usual LED lamps, for adding data communication features to illuminating systems. This paradigm has attracted a growing interest in both scientific and industrial community in the latter decade. Nevertheless, classical wireless communication mechanisms for physical and Medium Access Control (MAC) layers are hardly available for VLC, due to the massive external interference caused by sunlight. Moreover, effects related to the data frames features need to be taken into account in order to improve the effectiveness of the VLC paradigm. Such as an instance, the preamble length of a packet in order to synchronize the data transmission represents an important factor in VLC. A too long preamble allows a better synchronization while impacting negatively in terms of overhead. Nevertheless, a too short preamble may be not effective for synchronizing the transmission. The more suitable selection of the preamble length is strictly related to the noise environments. In order to make an adaptive selection able to choice the more suitable preamble length, we have designed and integrated in our VLC system a machine learning algorithm based on multi-arm bandit approach, in order to dynamically select the best configuration [28]. Another important approach to face the high interference impacting on the VLC performance is represented by the treatment of the noise through a signal processing approach in order to estimate it and proceed with a mitigation of the noise [10], [29], [36], [30]. This approach has been implemented and tested by the means of a real prototype. Results obtained show the effectiveness of a similar approach.

6.3. Alternative communications

Participant: Valeria Loscri.

In the last few years, there has been an increasing interest in the study of "alternative communication" paradigms, ranging from the exploitation of the visible light as carrier information, the exploitation of a different portion of the spectrum in the THz frequency, the leverage of artificial molecules for transmitting information (i.e. artificial molecular communication). Another interesting approach is consisting on a different perspective of the interaction between the signals and the environment. Right now, the environment has always been considered as something that cannot be "changed", a kind of obstacle for the wireless transmission. A new paradigm is arising, based on the metamaterial surface, where the interaction between the signals and the environment can be adapted in order to improve the performance of the communication.

6.3.1. TeraHez communications

In [32] and [23] we have investigated a metasurface and how the design of this metasurface has to be realized in order to adapt the behavior of the optical and THz signals based on the specific application considered.

6.3.2. Molecular communications

Concerning the artificial molecular communication paradigm, a fundamental aspect to be considered is that the most of times the target application of this type of communication is a biological system. A fundamental question arising is then: how maximize the effectiveness of the communication by keeping a low impact in terms of interference for the biological system? We have tried to answer to these fundamental questions by considering a signal processing approach in [11], [19].

6.4. Long range communications

Participants: Nathalie Mitton, Brandon Foubert, Ibrahim Amadou.

In the context of smart farming, communications still pose a key challenge. Ubiquitous access to the Internet is not available worldwide, and battery capacity is still a limitation. Inria and the Sencrop company are collaborating to develop an innovative solution for wireless weather stations, based on multi-technology communications, to enable smart weather stations deployment everywhere around the globe. We discuss this model in [12] and assess the quality of a LoRA signal in different conditions [16].

6.5. Vehicular networks

Participants: Nathalie Mitton, Valeria Loscri.

6.5.1. Positioning

Typical Global Navigation Satellite System (GNSS) receivers offer precision in the order of meters. This error margin is excessive for vehicular safety applications, such as forward collision warning, autonomous intersection management, or hard braking sensing. In [14] we develop CooPS, a GNSS positioning system that uses Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communications to cooperatively determine absolute and relative position of the ego-vehicle with enough precision. To that end, we use differential GNSS through position vector differencing to acquire track and across-track axes projections, employing elliptical and spherical geometries. We evaluate CooPS performance by carrying out real experiments using off-the-shelf IEEE 802.11p equipment at the campus of the Federal University of Rio de Janeiro. We obtain an accuracy level under 1.0 and 1.5 m for track (where-in-lane) and across-track (which-lane) axes, respectively. These accuracy levels were achieved using a 2.5 m accuracy circular error probable (CEP) of 50% and a 5 Hz navigation update rate GNSS receiver.

6.5.2. Vehicular social networks

In recent years, the concept of social networking combined with the Internet of Vehicles has brought to the definition of the Social IoV (SIoV) paradigm, i.e., a social network where every vehicle is capable of establishing social relationships in an autonomous way with other vehicles or road infrastructure equipment. In SIoV, social networking is applied to vehicular networks according to how social ties are built upon, i.e., either among vehicles or humans. An analysis of the SIoT-based social relations in a vehicular network scenario for establishing a Social Internet of Vehicles and providing insights on this growing research area [33]. By considering the specific features of the Online Social Networks (ONSs) and Vehicular Social Networks (VSNs), we realize that there are limitations and advantages on both these systems. In [15] we have proposed SOVER, a hybrid OSN-VSN framework, allowing the communication between both the communities, the OSNs and VSNs. In [24] we investigate the twofold nature of SIoV, both based on human factors and relationships and as an instance of the Social Internet of Things (SIoT. Based on this twofold nature, it is possible to distinguish different applications and use-cases.)

6.6. On the use of controlled mobility

Participant: Nathalie Mitton.

Relying on controlled mobility as enabled by drones or robots could be a great asset for task management, data collection or quality of network deployment.

6.6.1. Robots

Robots and controlled mobility can help in the dynamic coverage of an area. In [22], we address the problem of defining a wireless sensor network by deploying sensors with the aim of guaranteeing the coverage of the area and the connectivity among the sensors. The wireless sensor networks are widely studied since they provide several services, e.g., environmental monitoring and target tracking. We consider several typologies of sensors characterized by different sensing and connectivity ranges. A cost is associated with each typology of sensors. In particular, the higher the sensing and connectivity ranges, the higher the cost. We formulate the problem of deploying sensors at minimum cost such that each sensor is connected to a base station with either a one-or a multi-hop and the area is full covered. We present preliminary computational results by solving the proposed mathematical model, on several instances. We provide a simulation-based analysis of the performances of such a deployment from the routing perspective.

Robots could be helpful when called upon an alert sent by sensors. But to intervene quickly, they need to locate or follow back the alert source as fast as possible. Two new algorithms (GFGF1 and GFGF2) for event finding in wireless sensor and robot networks based on the Greedy-Face-Greedy (GFG) routing are proposed in [27]. The purpose of finding the event (reported by sensors) is to allocate the task to the closest robot to act upon the event. Using two scenarios (event in or out of the network) and two topologies (random and random with hole) it is shown that GFGF1 always find the closest robot to the event but with more than twice higher communication cost compared to GFG, especially for the outside of the network scenario. GFGF2 features more than 4 times communication cost reduction compared to GFG but with percentage of finding the closest robot up to 90%.

6.6.2. Drones

Disaster scenarios are particularly devastating in urban environments, which are generally very densely populated. Disasters not only endanger the life of people, but also affect the existing communication infrastructure. In fact, such an infrastructure could be completely destroyed or damaged; even when it continues working, it suffers from high access demand to its resources within a short period of time, thereby compromising the efficiency of rescue operations. [31], [25] leverage the ubiquitous presence of wireless devices (e.g., smartphones) in urban scenarios to assist search and rescue activities following a disaster. This work considers multi-interface wireless devices and drones to collect emergency messages in areas affected by natural disasters. Specifically, it proposes a collaborative data collection protocol that organizes wireless devices in multiple tiers by targeting a fair energy consumption in the whole network, thereby extending the network lifetime. Moreover, it introduces a scheme to control the path of drones so as to collect data in a short time. Simulation results in realistic settings show that the proposed solution balances the energy consumption in the network by means of efficient drone routes, thereby effectively assisting search and rescue operations.

6.7. Self-organization, routing and orchestration

Participants: Nathalie Mitton, Valeria Loscri, Brandon Foubert.

By offering low-latency and context-aware services, fog computing will have a peculiar role in the deployment of Internet of Things (IoT) applications for smart environments. Unlike the conventional remote cloud, for which consolidated architectures and deployment options exist, many design and implementation aspects remain open when considering the latest fog computing paradigm. In [9], we focus on the problems of dynamically discovering the processing and storage resources distributed among fog nodes and, accordingly, orchestrating them for the provisioning of IoT services for smart environments. In particular, we show how these functionalities can be effectively supported by the revolutionary Named Data Networking (NDN) paradigm. Originally conceived to support named content delivery, NDN can be extended to request and provide named computation services, with NDN nodes acting as both content routers and in-network service executors. To substantiate our analysis, we present an NDN fog computing framework with focus on a smart campus scenario, where the execution of IoT services is dynamically orchestrated and performed by NDN nodes in a distributed fashion. A simulation campaign in ndnSIM, the reference network simulator of the NDN research community, is also presented to assess the performance of our proposal against state-of-the-art solutions. Results confirm the superiority of the proposal in terms of service provisioning time, paid at the expenses of a slightly higher amount of traffic exchanged among fog nodes.

[26] proposes FLY-COPE, a complete self-organization architecture that relies on cooperative communications and drone-assisted data collection, allowing a fast location of victims and rescuing operation organization in disaster relief operation. FLY-COPE mainly combines two components: i) a ground component that spontaneously emerges from any communicating devices (piece of infrastructure, mobile phone, etc) that cooperate to alert rescuers and remain all alive as long as possible and ii) an aerial component comprising UAV to communicate efficiently with ground devices. We show by simulation and/or by experimentation that each component of FLY-COPE allows substantial energy saving for efficient and fast disaster response.

The IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL) builds a Direction Oriented Directed Acyclic Graph (DODAG) rooted at one node. This node may act as a border router to provide Internet connectivity to the members of the DODAG but such a situation creates a single point of failure. Upon border router failure, all nodes connected to the DODAG are affected as all ongoing communications are instantly broken and no new communications can be initiated. Moreover, nodes close to the border router should forward traffic from farther nodes in addition to their own, which may cause congestion and energy depletion inequality. In [20], we specify a full solution to enable border router redundancy in RPL networks. To achieve this, we propose a mechanism leveraging cooperation between colocated RPL networks. It enables failover to maintain Internet connectivity and load balancing to improve the overall energy consumption and bandwidth. Our contribution has been implemented in Contiki OS and was evaluated through experiments performed on the FIT IoT-LAB testbed.

GANG Project-Team

7. New Results

7.1. Graph and Combinatorial Algorithms

7.1.1. Fast Diameter Computation within Split Graphs

When can we compute the diameter of a graph in quasi linear time? In [22], we address this question for the class of *split graphs*, that we observe to be the hardest instances for deciding whether the diameter is at most two. We stress that although the diameter of a non-complete split graph can only be either 2 or 3, under the Strong Exponential-Time Hypothesis (SETH) we cannot compute the diameter of a split graph in less than quadratic time. Therefore it is worth to study the complexity of diameter computation on *subclasses* of split graphs, in order to better understand the complexity border. Specifically, we consider the split graphs with bounded *clique-interval number* and their complements, with the former being a natural variation of the concept of interval number for split graphs that we introduce in this paper. We first discuss the relations between the clique-interval number and other graph invariants such as the classic interval number of graphs, the treewidth, the *VC-dimension* and the *stabbing number* of a related hypergraph. Then, in part based on these above relations, we almost completely settle the complexity of diameter computation on these subclasses of split graphs:

- For the k-clique-interval split graphs, we can compute their diameter in truly subquadratic time if $k = \mathcal{O}(1)$, and even in quasi linear time if $k = o(\log n)$ and in addition a corresponding ordering is given. However, under SETH this cannot be done in truly subquadratic time for any $k = \omega(\log n)$.
- For the *complements* of k-clique-interval split graphs, we can compute their diameter in truly subquadratic time if $k = \mathcal{O}(1)$, and even in time $\mathcal{O}(km)$ if a corresponding ordering is given. Again this latter result is optimal under SETH up to polylogarithmic factors.

Our findings raise the question whether a k-clique interval ordering can always be computed in quasi linear time. We prove that it is the case for $k = 1$ and for some subclasses such as bounded-treewidth split graphs, threshold graphs and comparability split graphs. Finally, we prove that some important subclasses of split graphs – including the ones mentioned above – have a bounded clique-interval number.

7.1.2. Diameter computation on H-minor free graphs and graphs of bounded (distance) VC-dimension

Under the Strong Exponential-Time Hypothesis, the diameter of general unweighted graphs cannot be computed in truly subquadratic time. Nevertheless there are several graph classes for which this can be done such as bounded-treewidth graphs, interval graphs and planar graphs, to name a few. We propose to study unweighted graphs of constant *distance VC-dimension* as a broad generalization of many such classes – where the distance VC-dimension of a graph G is defined as the VC-dimension of its ball hypergraph: whose hyperedges are the balls of all possible radii and centers in G. In particular for any fixed H, the class of H-minor free graphs has distance VC-dimension at most $|V(H)| - 1$. In [23], we show the following.

- Our first main result is a Monte Carlo algorithm that on graphs of distance VC-dimension at most d, for any fixed k, either computes the diameter or concludes that it is larger than k in time $\tilde{\mathcal{O}}(k \cdot mn^{1-\varepsilon_d})$, where $\varepsilon_d \in (0; 1)$ only depends on d. We thus obtain a *truly subquadratic-time parameterized* algorithm for computing the diameter on such graphs.
- Then as a byproduct of our approach, we get the first truly subquadratic-time randomized algorithm for *constant* diameter computation on all the *nowhere dense* graph classes. The latter classes include all proper minor-closed graph classes, bounded-degree graphs and graphs of bounded expansion.
- Finally, we show how to remove the dependency on k for *any* graph class that excludes a fixed graph H as a minor. More generally, our techniques apply to any graph with constant distance VC-dimension and *polynomial expansion* (or equivalently having strongly sublinear balanced separators). As a result for all such graphs one obtains a truly subquadratic-time randomized algorithm for computing their diameter.

We note that all our results also hold for *radius* computation. Our approach is based on the work of Chazelle and Welzl who proved the existence of spanning paths with strongly sublinear *stabbing number* for every hypergraph of constant VC-dimension. We show how to compute such paths efficiently by combining known algorithms for the stabbing number problem with a clever use of ε-nets, region decomposition and other partition techniques.

7.1.3. Approximation of eccentricites and distance using δ-hyperbolicity

In [9], we show that the eccentricities of all vertices of a δ-hyperbolic graph $G=(V,E)$ can be computed in linear time with an additive one-sided error of at most $c \cdot \delta$, i.e., after a linear time preprocessing, for every vertex v of G one can compute in $O(1)$ time an estimate $\overline{ecc_G(v)}$ of its eccentricity $ecc_G(v) := max\{d_G(u,v) : u \in V\}$ such that $ecc_G(v) \leq \overline{ecc_G(v)} \leq ecc_G(v) + c \cdot \delta$ for a small constant c. We prove that every δ-hyperbolic graph G has a shortest path tree T, constructible in linear time, such that for every vertex v of G, $ecc_G(v) \leq ecc_T(v) \leq ecc_G(v) + c \cdot \delta$, where $ecc_T(v) := max\{d_T(u,v) : u \in V\}$. These results are based on an interesting monotonicity property of the eccentricity function of hyperbolic graphs: the closer a vertex is to the center of G, the smaller its eccentricity is. We also show that the distance matrix of G with an additive one-sided error of at most $c' \cdot \delta$ can be computed in $O(|V|^2 log^2 |V|)$ time, where $c' < c$ is a small constant. Recent empirical studies show that many real world graphs (including Internet application networks, web networks, collaboration networks, social networks, biological networks, and others) have small hyperbolicity. So, we analyze the performance of our algorithms for approximating eccentricities and distance matrix on a number of real-world networks. Our experimental results show that the obtained estimates are even better than the theoretical bounds.

7.1.4. Graph and Hypergraph Decompositions

In [26], we study modular decomposition of hypergraphs and propose some polynomial algorithms to this aim. We also study several notions of approximation of modular decomposition of graphs, by relaxing the definition of modules introducing a tolerance (ϵ edges can miss) this will be presented at CALDAM 2020, Hyderabad. Both topics can be seen as the search for new models of regularity in discrete structures, as in particular bipartite graphs. In both references our polynomial algorithms have to be improved before being applied on real-world data.

7.2. Distributed Computing

7.2.1. Distributed Interactive Proofs

In a distributed locally-checkable proof, we are interested in checking the legality of a given network configuration with respect to some Boolean predicate. To do so, the network enlists the help of a *prover* — a computationally-unbounded oracle that aims at convincing the network that its state is legal, by providing the nodes with certificates that form a distributed proof of legality. The nodes then verify the proof by examining their certificate, their local neighborhood and the certificates of their neighbors.

In [24], we examine the power of a *randomized* form of locally-checkable proof, called *distributed Merlin-Arthur protocols*, or dMA for short. In a dMA protocol, the prover assigns each node a short certificate, and the nodes then exchange *random messages* with their neighbors. We show that while there exist problems for which dMA protocols are more efficient than protocols that do not use randomness, for several natural problems, including Leader Election, Diameter, Symmetry, and Counting Distinct Elements, dMA protocols are no more efficient than standard nondeterministic protocols. This is in contrast with Arthur-Merlin (dAM) protocols and Randomized Proof Labeling Schemes (RPLS), which are known to provide improvements in certificate size, at least for some of the aforementioned properties.

The study of interactive proofs in the context of distributed network computing is a novel topic, recently introduced by Kol, Oshman, and Saxena [PODC 2018]. In the spirit of sequential interactive proofs theory, we study in [20] the power of distributed interactive proofs. This is achieved via a series of results establishing trade-offs between various parameters impacting the power of interactive proofs, including the number of interactions, the certificate size, the communication complexity, and the form of randomness used. Our results also connect distributed interactive proofs with the established field of distributed verification. In general, our results contribute to providing structure to the landscape of distributed interactive proofs.

7.2.2. Topological Approach of Network Computing

More than two decades ago, combinatorial topology was shown to be useful for analyzing distributed fault-tolerant algorithms in shared memory systems and in message passing systems. In [18], we show that combinatorial topology can also be useful for analyzing distributed algorithms in networks of arbitrary structure. To illustrate this, we analyze consensus, set-agreement, and approximate agreement in networks, and derive lower bounds for these problems under classical computational settings, such as the LOCAL model and dynamic networks.

In [19], we study the number of rounds needed to solve consensus in a synchronous network G where at most t nodes may fail by crashing. This problem has been thoroughly studied when G is a complete graph, but very little is known when G is arbitrary. We define a notion of radius that considers all ways in which t nodes may crash, and present an algorithm that solves consensus in radius rounds. Then we derive a lower bound showing that our algorithm is optimal for vertex-transitive graphs, among oblivious algorithms.

7.2.3. Making Local Algorithms Wait-Free

When considering distributed computing, reliable message-passing synchronous systems on the one side, and asynchronous failure-prone shared-memory systems on the other side, remain two quite independently studied ends of the reliability/asynchrony spectrum. The concept of locality of a computation is central to the first one, while the concept of wait-freedom is central to the second one. In [8], we propose a new DECOUPLED model in an attempt to reconcile these two worlds. It consists of a synchronous and reliable communication graph of nodes, and on top a set of asynchronous crash-prone processes, each attached to a communication node. To illustrate the DECOUPLED model, the paper presents an asynchronous 3-coloring algorithm for the processes of a ring. From the processes point of view, the algorithm is wait-free. From a locality point of view, each process uses information only from processes at distance $O(log^*n)$ from it. This local wait-free algorithm is based on an extension of the classical Cole and Vishkin's vertex coloring algorithm in which the processes are not required to start simultaneously.

In [31], we show that, for any task T associated to a locally checkable labeling (lcl), if T is solvable in t rounds by a deterministic algorithm in the local model, then T remains solvable by a deterministic algorithm in $O(t)$ rounds in an asynchronous variant of the local model whenever $t = O(polylogn)$.

7.2.4. Towards Synthesis of Distributed Algorithms with SMT Solvers

In [32], we consider the problem of synthesizing distributed algorithms working on a specific execution context. We show it is possible to use the linear time temporal logic in order to both specify the correctness of algorithms and their execution contexts. We then provide a method allowing to reduce the synthesis problem of finite state algorithms to some model-checking problems. We finally apply our technique to automatically generate algorithms for consensus and epsilon-agreement in the case of two processes using the SMT solver Z3.

7.2.5. On Weakest Failure Detector

Failure detectors are devices (objects) that provide the processes with information on failures. They were introduced to enrich asynchronous systems so that it becomes possible to solve problems (or implement concurrent objects) that are otherwise impossible to solve in pure asynchronous systems where processes are prone to crash failures. The most famous failure detector (which is called “eventual leader” and denoted Ω) is the weakest failure detector which allows consensus to be solved in n-process asynchronous systems where up to $t = n - 1$ processes may crash in the read/write communication model, and up to $t < n/2$ processes may crash in the message-passing communication model.

When looking at the mutual exclusion problem (or equivalently the construction of a lock object), while the weakest failure detectors are known for both asynchronous message-passing systems and read/write systems in which up to $t < n$ processes may crash, for the starvation-freedom progress condition, it is not yet known for weaker deadlock-freedom progress condition in read/write systems. In [34], we extend the previous results, namely, it presents the weakest failure detector that allows mutual exclusion to be solved in asynchronous n-process read/write systems where any number of processes may crash, whatever the progress condition (deadlock-freedom or starvation-freedom).

In read/read/write communication model, and in the message-passing communication model, all correct processes are supposed to participate in a consensus instance and in particular the eventual leader.

In [33], we considers the case where some subset of processes that do not crash (not predefined in advance) are allowed not to participate in a consensus instance. In this context Ω cannot be used to solve consensus as it could elect as eventual leader a non-participating process. This paper presents the weakest failure detector that allows correct processes not to participate in a consensus instance.This failure detector, denoted Ω^*, is a variant of Ω. The paper presents also an Ω^*-based consensus algorithm for the asynchronous read/write model, in which any number of processes may crash, and not all the correct processes are required to participate.

7.2.6. Multi-Round Cooperative Search Games with Multiple Players

We study search in the context of competing agents. The setting we consider combines game-theoretic concepts with notions related to parallel computing. Assume that a treasure is placed in one of M boxes according to a known distribution and that k searchers are searching for it in parallel during T rounds. In [27], we study the question of how to incentivize selfish players so that group performance would be maximized. Here, this is measured by the *success probability*, namely, the probability that at least one player finds the treasure. We focus on *congestion policies*$C(\ell)$ that specify the reward that a player receives if it is one of ℓ players that (simultaneously) find the treasure for the first time. Our main technical contribution is proving that the *exclusive policy*, in which $C(1) = 1$ and $C(\ell) = 0$ for $\ell > 1$, yields a *price of anarchy* of $\left(1 - (1 - 1/k)^k\right)^{-1}$, and that this is the best possible price among all symmetric reward mechanisms. For this policy we also have an explicit description of a symmetric equilibrium, which is in some sense unique, and moreover enjoys the best success probability among all symmetric profiles. For general congestion policies, we show how to polynomially find, for any $\theta > 0$, a symmetric multiplicative $(1 + \theta)(1 + C(k))$-equilibrium.

Together with an appropriate reward policy, a central entity can suggest players to play a particular profile at equilibrium. As our main conceptual contribution, we advocate the use of symmetric equilibria for such purposes. Besides being fair, we argue that symmetric equilibria can also become highly robust to crashes of players. Indeed, in many cases, despite the fact that some small fraction of players crash (or refuse to participate), symmetric equilibria remain efficient in terms of their group performances and, at the same time, serve as approximate equilibria. We show that this principle holds for a class of games, which we call *monotonously scalable* games. This applies in particular to our search game, assuming the natural *sharing policy*, in which $C(\ell) = 1/\ell$. For the exclusive policy, this general result does not hold, but we show that the symmetric equilibrium is nevertheless robust under mild assumptions.

7.3. Models and Algorithms for Networks

7.3.1. Exploiting Hopsets: Improved Distance Oracles for Graphs of Constant Highway Dimension and Beyond

For fixed $h \geq 2$, we consider in [25] the task of adding to a graph G a set of weighted shortcut edges on the same vertex set, such that the length of a shortest h-hop path between any pair of vertices in the augmented graph is exactly the same as the original distance between these vertices in G. A set of shortcut edges with this property is called an *exact h-hopset* and may be applied in processing distance queries on graph G. In particular, a 2-hopset directly corresponds to a distributed distance oracle known as a *hub labeling*. In this work, we explore centralized distance oracles based on 3-hopsets and display their advantages in several practical scenarios. In particular, for graphs of constant highway dimension, and more generally for graphs of constant skeleton dimension, we show that 3-hopsets require *exponentially* fewer shortcuts per node than any previously described distance oracle, and also offer a speedup in query time when compared to simple oracles based on a direct application of 2-hopsets. Finally, we consider the problem of computing minimum-size h-hopset (for any $h \geq 2$) for a given graph G, showing a polylogarithmic-factor approximation for the case of unique shortest path graphs. When $h = 3$, for a given bound on the space used by the distance oracle, we provide a construction of hopset achieving polylog approximation both for space and query time compared to the optimal 3-hopset oracle given the space bound.

7.3.2. Hardness of exact distance queries in sparse graphs through hub labeling

A *distance labeling scheme* is an assignment of bit-labels to the vertices of an undirected, unweighted graph such that the distance between any pair of vertices can be decoded solely from their labels. An important class of distance labeling schemes is that of *hub labelings*, where a node $v \in G$ stores its distance to the so-called hubs $S_v \subseteq V$, chosen so that for any $u, v \in V$ there is $w \in S_u \cap S_v$ belonging to some shortest uv path. Notice that for most existing graph classes, the best distance labelling constructions existing use at some point a hub labeling scheme at least as a key building block.

In [28], our interest lies in hub labelings of sparse graphs, i.e., those with $|E(G)| = O(n)$, for which we show a lowerbound of $\frac{n}{2^{O(\sqrt{\log n})}}$ for the average size of the hubsets. Additionally, we show a hub-labeling construction for sparse graphs of average size $O(\frac{n}{RS(n)^c})$ for some $0 < c < 1$, where $RS(n)$ is the so-called Ruzsa-Szemerédi function, linked to structure of induced matchings in dense graphs. This implies that further improving the lower bound on hub labeling size to $\frac{n}{2^{(\log n)^{o(1)}}}$ would require a breakthrough in the study of lower bounds on $RS(n)$, which have resisted substantial improvement in the last 70 years.

For general distance labeling of sparse graphs, we show a lowerbound of $\frac{1}{2^{\Theta(\sqrt{\log n})}} SumIndex(n)$, where $SumIndex(n)$ is the communication complexity of the Sum-Index problem over Z_n. Our results suggest that the best achievable hub-label size and distance-label size in sparse graphs may be $\Theta(\frac{n}{2^{(\log n)^c}})$ for some $0 < c < 1$.

7.3.3. Fast Public Transit Routing with Unrestricted Walking through Hub Labeling

In [30], we propose a novel technique for answering routing queries in public transportation networks that allows unrestricted walking. We consider several types of queries: earliest arrival time, Pareto-optimal journeys regarding arrival time, number of transfers and walking time, and profile, i.e. finding all Pareto-optimal journeys regarding travel time and arrival time in a given time interval. Our techniques uses hub labeling to represent unlimited foot transfers and can be adapted to both classical algorithms RAPTOR and CSA. We obtain significant speedup compared to the state-of-the-art approach based on contraction hierarchies. A research report version is deposited on HAL with number hal-02161283.

7.3.4. Independent Lazy Better-Response Dynamics on Network Games

In [29], 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.

7.3.5. A Comparative Study of Neural Network Compression

There has recently been an increasing desire to evaluate neural networks locally on computationally-limited devices in order to exploit their recent effectiveness for several applications; such effectiveness has nevertheless come together with a considerable increase in the size of modern neural networks, which constitute a major downside in several of the aforementioned computationally-limited settings. There has thus been a demand of compression techniques for neural networks. Several proposal in this direction have been made, which famously include hashing-based methods and pruning-based ones. However, the evaluation of the efficacy of these techniques has so far been heterogeneous, with no clear evidence in favor of any of them over the others. In [36], we address this latter issue by providing a comparative study. While most previous studies test the capability of a technique in reducing the number of parameters of state-of-the-art networks , we follow [CWT

+ 15] in evaluating their performance on basic architectures on the MNIST dataset and variants of it, which allows for a clearer analysis of some aspects of their behavior. To the best of our knowledge, we are the first to directly compare famous approaches such as HashedNet, Optimal Brain Damage (OBD), and magnitude-based pruning with L1 and L2 regularization among them and against equivalent-size feed-forward neural networks with simple (fully-connected) and structural (convolutional) neural networks. Rather surprisingly, our experiments show that (iterative) pruning-based methods are substantially better than the HashedNet architecture, whose compression doesn't appear advantageous to a carefully chosen convolutional network. We also show that, when the compression level is high, the famous OBD pruning heuristics deteriorates to the point of being less efficient than simple magnitude-based techniques.

GANG Project-Team

7. New Results

7.1. Graph and Combinatorial Algorithms

7.1.1. Fast Diameter Computation within Split Graphs

When can we compute the diameter of a graph in quasi linear time? In [22], we address this question for the class of *split graphs*, that we observe to be the hardest instances for deciding whether the diameter is at most two. We stress that although the diameter of a non-complete split graph can only be either 2 or 3, under the Strong Exponential-Time Hypothesis (SETH) we cannot compute the diameter of a split graph in less than quadratic time. Therefore it is worth to study the complexity of diameter computation on *subclasses* of split graphs, in order to better understand the complexity border. Specifically, we consider the split graphs with bounded *clique-interval number* and their complements, with the former being a natural variation of the concept of interval number for split graphs that we introduce in this paper. We first discuss the relations between the clique-interval number and other graph invariants such as the classic interval number of graphs, the treewidth, the *VC-dimension* and the *stabbing number* of a related hypergraph. Then, in part based on these above relations, we almost completely settle the complexity of diameter computation on these subclasses of split graphs:

- For the k-clique-interval split graphs, we can compute their diameter in truly subquadratic time if $k = \mathcal{O}(1)$, and even in quasi linear time if $k = o(\log n)$ and in addition a corresponding ordering is given. However, under SETH this cannot be done in truly subquadratic time for any $k = \omega(\log n)$.
- For the *complements* of k-clique-interval split graphs, we can compute their diameter in truly subquadratic time if $k = \mathcal{O}(1)$, and even in time $\mathcal{O}(km)$ if a corresponding ordering is given. Again this latter result is optimal under SETH up to polylogarithmic factors.

Our findings raise the question whether a k-clique interval ordering can always be computed in quasi linear time. We prove that it is the case for $k = 1$ and for some subclasses such as bounded-treewidth split graphs, threshold graphs and comparability split graphs. Finally, we prove that some important subclasses of split graphs – including the ones mentioned above – have a bounded clique-interval number.

7.1.2. Diameter computation on H-minor free graphs and graphs of bounded (distance) VC-dimension

Under the Strong Exponential-Time Hypothesis, the diameter of general unweighted graphs cannot be computed in truly subquadratic time. Nevertheless there are several graph classes for which this can be done such as bounded-treewidth graphs, interval graphs and planar graphs, to name a few. We propose to study unweighted graphs of constant *distance VC-dimension* as a broad generalization of many such classes – where the distance VC-dimension of a graph G is defined as the VC-dimension of its ball hypergraph: whose hyperedges are the balls of all possible radii and centers in G. In particular for any fixed H, the class of H-minor free graphs has distance VC-dimension at most $|V(H)| - 1$. In [23], we show the following.

- Our first main result is a Monte Carlo algorithm that on graphs of distance VC-dimension at most d, for any fixed k, either computes the diameter or concludes that it is larger than k in time $\widetilde{\mathcal{O}}(k \cdot mn^{1-\varepsilon_d})$, where $\varepsilon_d \in (0;1)$ only depends on d. We thus obtain a *truly subquadratic-time parameterized* algorithm for computing the diameter on such graphs.
- Then as a byproduct of our approach, we get the first truly subquadratic-time randomized algorithm for *constant* diameter computation on all the *nowhere dense* graph classes. The latter classes include all proper minor-closed graph classes, bounded-degree graphs and graphs of bounded expansion.
- Finally, we show how to remove the dependency on k for *any* graph class that excludes a fixed graph H as a minor. More generally, our techniques apply to any graph with constant distance VC-dimension and *polynomial expansion* (or equivalently having strongly sublinear balanced separators). As a result for all such graphs one obtains a truly subquadratic-time randomized algorithm for computing their diameter.

We note that all our results also hold for *radius* computation. Our approach is based on the work of Chazelle and Welzl who proved the existence of spanning paths with strongly sublinear *stabbing number* for every hypergraph of constant VC-dimension. We show how to compute such paths efficiently by combining known algorithms for the stabbing number problem with a clever use of ε-nets, region decomposition and other partition techniques.

7.1.3. Approximation of eccentricites and distance using δ-hyperbolicity

In [9], we show that the eccentricities of all vertices of a δ-hyperbolic graph $G = (V, E)$ can be computed in linear time with an additive one-sided error of at most $c \cdot \delta$, i.e., after a linear time preprocessing, for every vertex v of G one can compute in $O(1)$ time an estimate $\widehat{ecc_G(v)}$ of its eccentricity $ecc_G(v) := max\{d_G(u, v) : u \in V\}$ such that $ecc_G(v) \leq \widehat{ecc_G(v)} \leq ecc_G(v) + c \cdot \delta$ for a small constant c. We prove that every δ-hyperbolic graph G has a shortest path tree T, constructible in linear time, such that for every vertex v of G, $ecc_G(v) \leq ecc_T(v) \leq ecc_G(v) + c \cdot \delta$, where $ecc_T(v) := max\{d_T(u, v) : u \in V\}$. These results are based on an interesting monotonicity property of the eccentricity function of hyperbolic graphs: the closer a vertex is to the center of G, the smaller its eccentricity is. We also show that the distance matrix of G with an additive one-sided error of at most $c' \cdot \delta$ can be computed in $O(|V|^2 log^2|V|)$ time, where $c' < c$ is a small constant. Recent empirical studies show that many real world graphs (including Internet application networks, web networks, collaboration networks, social networks, biological networks, and others) have small hyperbolicity. So, we analyze the performance of our algorithms for approximating eccentricities and distance matrix on a number of real-world networks. Our experimental results show that the obtained estimates are even better than the theoretical bounds.

7.1.4. Graph and Hypergraph Decompositions

In [26], we study modular decomposition of hypergraphs and propose some polynomial algorithms to this aim. We also study several notions of approximation of modular decomposition of graphs, by relaxing the definition of modules introducing a tolerance (ϵ edges can miss) this will be presented at CALDAM 2020, Hyderabad. Both topics can be seen as the search for new models of regularity in discrete structures, as in particular bipartite graphs. In both references our polynomial algorithms have to be improved before being applied on real-world data.

7.2. Distributed Computing

7.2.1. Distributed Interactive Proofs

In a distributed locally-checkable proof, we are interested in checking the legality of a given network configuration with respect to some Boolean predicate. To do so, the network enlists the help of a *prover* — a computationally-unbounded oracle that aims at convincing the network that its state is legal, by providing the nodes with certificates that form a distributed proof of legality. The nodes then verify the proof by examining their certificate, their local neighborhood and the certificates of their neighbors.

In [24], we examine the power of a *randomized* form of locally-checkable proof, called *distributed Merlin-Arthur protocols*, or dMA for short. In a dMA protocol, the prover assigns each node a short certificate, and the nodes then exchange *random messages* with their neighbors. We show that while there exist problems for which dMA protocols are more efficient than protocols that do not use randomness, for several natural problems, including Leader Election, Diameter, Symmetry, and Counting Distinct Elements, dMA protocols are no more efficient than standard nondeterministic protocols. This is in contrast with Arthur-Merlin (dAM) protocols and Randomized Proof Labeling Schemes (RPLS), which are known to provide improvements in certificate size, at least for some of the aforementioned properties.

The study of interactive proofs in the context of distributed network computing is a novel topic, recently introduced by Kol, Oshman, and Saxena [PODC 2018]. In the spirit of sequential interactive proofs theory, we study in [20] the power of distributed interactive proofs. This is achieved via a series of results establishing trade-offs between various parameters impacting the power of interactive proofs, including the number of interactions, the certificate size, the communication complexity, and the form of randomness used. Our results also connect distributed interactive proofs with the established field of distributed verification. In general, our results contribute to providing structure to the landscape of distributed interactive proofs.

7.2.2. Topological Approach of Network Computing

More than two decades ago, combinatorial topology was shown to be useful for analyzing distributed fault-tolerant algorithms in shared memory systems and in message passing systems. In [18], we show that combinatorial topology can also be useful for analyzing distributed algorithms in networks of arbitrary structure. To illustrate this, we analyze consensus, set-agreement, and approximate agreement in networks, and derive lower bounds for these problems under classical computational settings, such as the LOCAL model and dynamic networks.

In [19], we study the number of rounds needed to solve consensus in a synchronous network G where at most t nodes may fail by crashing. This problem has been thoroughly studied when G is a complete graph, but very little is known when G is arbitrary. We define a notion of radius that considers all ways in which t nodes may crash, and present an algorithm that solves consensus in radius rounds. Then we derive a lower bound showing that our algorithm is optimal for vertex-transitive graphs, among oblivious algorithms.

7.2.3. Making Local Algorithms Wait-Free

When considering distributed computing, reliable message-passing synchronous systems on the one side, and asynchronous failure-prone shared-memory systems on the other side, remain two quite independently studied ends of the reliability/asynchrony spectrum. The concept of locality of a computation is central to the first one, while the concept of wait-freedom is central to the second one. In [8], we propose a new DECOUPLED model in an attempt to reconcile these two worlds. It consists of a synchronous and reliable communication graph of nodes, and on top a set of asynchronous crash-prone processes, each attached to a communication node. To illustrate the DECOUPLED model, the paper presents an asynchronous 3-coloring algorithm for the processes of a ring. From the processes point of view, the algorithm is wait-free. From a locality point of view, each process uses information only from processes at distance $O(log^*n)$ from it. This local wait-free algorithm is based on an extension of the classical Cole and Vishkin's vertex coloring algorithm in which the processes are not required to start simultaneously.

In [31], we show that, for any task T associated to a locally checkable labeling (lcl), if T is solvable in t rounds by a deterministic algorithm in the local model, then T remains solvable by a deterministic algorithm in $O(t)$ rounds in an asynchronous variant of the local model whenever $t = O(polylogn)$.

7.2.4. Towards Synthesis of Distributed Algorithms with SMT Solvers

In [32], we consider the problem of synthesizing distributed algorithms working on a specific execution context. We show it is possible to use the linear time temporal logic in order to both specify the correctness of algorithms and their execution contexts. We then provide a method allowing to reduce the synthesis problem of finite state algorithms to some model-checking problems. We finally apply our technique to automatically generate algorithms for consensus and epsilon-agreement in the case of two processes using the SMT solver Z3.

7.2.5. On Weakest Failure Detector

Failure detectors are devices (objects) that provide the processes with information on failures. They were introduced to enrich asynchronous systems so that it becomes possible to solve problems (or implement concurrent objects) that are otherwise impossible to solve in pure asynchronous systems where processes are prone to crash failures. The most famous failure detector (which is called "eventual leader" and denoted Ω) is the weakest failure detector which allows consensus to be solved in n-process asynchronous systems where up to $t = n - 1$ processes may crash in the read/write communication model, and up to $t < n/2$ processes may crash in the message-passing communication model.

When looking at the mutual exclusion problem (or equivalently the construction of a lock object), while the weakest failure detectors are known for both asynchronous message-passing systems and read/write systems in which up to $t < n$ processes may crash, for the starvation-freedom progress condition, it is not yet known for weaker deadlock-freedom progress condition in read/write systems. In [34], we extend the previous results, namely, it presents the weakest failure detector that allows mutual exclusion to be solved in asynchronous n-process read/write systems where any number of processes may crash, whatever the progress condition (deadlock-freedom or starvation-freedom).

In read/read/write communication model, and in the message-passing communication model, all correct processes are supposed to participate in a consensus instance and in particular the eventual leader.

In [33], we considers the case where some subset of processes that do not crash (not predefined in advance) are allowed not to participate in a consensus instance. In this context Ω cannot be used to solve consensus as it could elect as eventual leader a non-participating process. This paper presents the weakest failure detector that allows correct processes not to participate in a consensus instance.This failure detector, denoted Ω^*, is a variant of Ω. The paper presents also an Ω^*-based consensus algorithm for the asynchronous read/write model, in which any number of processes may crash, and not all the correct processes are required to participate.

7.2.6. Multi-Round Cooperative Search Games with Multiple Players

We study search in the context of competing agents. The setting we consider combines game-theoretic concepts with notions related to parallel computing. Assume that a treasure is placed in one of M boxes according to a known distribution and that k searchers are searching for it in parallel during T rounds. In [27], we study the question of how to incentivize selfish players so that group performance would be maximized. Here, this is measured by the *success probability*, namely, the probability that at least one player finds the treasure. We focus on *congestion policies*$C(\ell)$ that specify the reward that a player receives if it is one of ℓ players that (simultaneously) find the treasure for the first time. Our main technical contribution is proving that the *exclusive policy*, in which $C(1) = 1$ and $C(\ell) = 0$ for $\ell > 1$, yields a *price of anarchy* of $\left(1 - (1 - 1/k)^k\right)^{-1}$, and that this is the best possible price among all symmetric reward mechanisms. For this policy we also have an explicit description of a symmetric equilibrium, which is in some sense unique, and moreover enjoys the best success probability among all symmetric profiles. For general congestion policies, we show how to polynomially find, for any $\theta > 0$, a symmetric multiplicative $(1 + \theta)(1 + C(k))$-equilibrium.

Together with an appropriate reward policy, a central entity can suggest players to play a particular profile at equilibrium. As our main conceptual contribution, we advocate the use of symmetric equilibria for such purposes. Besides being fair, we argue that symmetric equilibria can also become highly robust to crashes of players. Indeed, in many cases, despite the fact that some small fraction of players crash (or refuse to participate), symmetric equilibria remain efficient in terms of their group performances and, at the same time, serve as approximate equilibria. We show that this principle holds for a class of games, which we call *monotonously scalable* games. This applies in particular to our search game, assuming the natural *sharing policy*, in which $C(\ell) = 1/\ell$. For the exclusive policy, this general result does not hold, but we show that the symmetric equilibrium is nevertheless robust under mild assumptions.

7.3. Models and Algorithms for Networks

7.3.1. Exploiting Hopsets: Improved Distance Oracles for Graphs of Constant Highway Dimension and Beyond

For fixed $h \geq 2$, we consider in [25] the task of adding to a graph G a set of weighted shortcut edges on the same vertex set, such that the length of a shortest h-hop path between any pair of vertices in the augmented graph is exactly the same as the original distance between these vertices in G. A set of shortcut edges with this property is called an *exact h-hopset* and may be applied in processing distance queries on graph G. In particular, a 2-hopset directly corresponds to a distributed distance oracle known as a *hub labeling*. In this work, we explore centralized distance oracles based on 3-hopsets and display their advantages in several practical scenarios. In particular, for graphs of constant highway dimension, and more generally for graphs of constant skeleton dimension, we show that 3-hopsets require *exponentially* fewer shortcuts per node than any previously described distance oracle, and also offer a speedup in query time when compared to simple oracles based on a direct application of 2-hopsets. Finally, we consider the problem of computing minimum-size h-hopset (for any $h \geq 2$) for a given graph G, showing a polylogarithmic-factor approximation for the case of unique shortest path graphs. When $h = 3$, for a given bound on the space used by the distance oracle, we provide a construction of hopset achieving polylog approximation both for space and query time compared to the optimal 3-hopset oracle given the space bound.

7.3.2. Hardness of exact distance queries in sparse graphs through hub labeling

A *distance labeling scheme* is an assignment of bit-labels to the vertices of an undirected, unweighted graph such that the distance between any pair of vertices can be decoded solely from their labels. An important class of distance labeling schemes is that of *hub labelings*, where a node $v \in G$ stores its distance to the so-called hubs $S_v \subseteq V$, chosen so that for any $u, v \in V$ there is $w \in S_u \cap S_v$ belonging to some shortest uv path. Notice that for most existing graph classes, the best distance labelling constructions existing use at some point a hub labeling scheme at least as a key building block.

In [28], our interest lies in hub labelings of sparse graphs, i.e., those with $|E(G)| = O(n)$, for which we show a lowerbound of $\frac{n}{2^{O(\sqrt{\log n})}}$ for the average size of the hubsets. Additionally, we show a hub-labeling construction for sparse graphs of average size $O(\frac{n}{RS(n)^c})$ for some $0 < c < 1$, where $RS(n)$ is the so-called Ruzsa-Szemerédi function, linked to structure of induced matchings in dense graphs. This implies that further improving the lower bound on hub labeling size to $\frac{n}{2^{(\log n)^{o(1)}}}$ would require a breakthrough in the study of lower bounds on $RS(n)$, which have resisted substantial improvement in the last 70 years.

For general distance labeling of sparse graphs, we show a lowerbound of $\frac{1}{2^{\Theta(\sqrt{\log n})}}SumIndex(n)$, where $SumIndex(n)$ is the communication complexity of the Sum-Index problem over Z_n. Our results suggest that the best achievable hub-label size and distance-label size in sparse graphs may be $\Theta(\frac{n}{2^{(\log n)^c}})$ for some $0 < c < 1$.

7.3.3. Fast Public Transit Routing with Unrestricted Walking through Hub Labeling

In [30], we propose a novel technique for answering routing queries in public transportation networks that allows unrestricted walking. We consider several types of queries: earliest arrival time, Pareto-optimal journeys regarding arrival time, number of transfers and walking time, and profile, i.e. finding all Pareto-optimal journeys regarding travel time and arrival time in a given time interval. Our techniques uses hub labeling to represent unlimited foot transfers and can be adapted to both classical algorithms RAPTOR and CSA. We obtain significant speedup compared to the state-of-the-art approach based on contraction hierarchies. A research report version is deposited on HAL with number hal-02161283.

7.3.4. Independent Lazy Better-Response Dynamics on Network Games

In [29], 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.

7.3.5. A Comparative Study of Neural Network Compression

There has recently been an increasing desire to evaluate neural networks locally on computationally-limited devices in order to exploit their recent effectiveness for several applications; such effectiveness has nevertheless come together with a considerable increase in the size of modern neural networks, which constitute a major downside in several of the aforementioned computationally-limited settings. There has thus been a demand of compression techniques for neural networks. Several proposal in this direction have been made, which famously include hashing-based methods and pruning-based ones. However, the evaluation of the efficacy of these techniques has so far been heterogeneous, with no clear evidence in favor of any of them over the others. In [36], we address this latter issue by providing a comparative study. While most previous studies test the capability of a technique in reducing the number of parameters of state-of-the-art networks , we follow [CWT

+ 15] in evaluating their performance on basic architectures on the MNIST dataset and variants of it, which allows for a clearer analysis of some aspects of their behavior. To the best of our knowledge, we are the first to directly compare famous approaches such as HashedNet, Optimal Brain Damage (OBD), and magnitude-based pruning with L1 and L2 regularization among them and against equivalent-size feed-forward neural networks with simple (fully-connected) and structural (convolutional) neural networks. Rather surprisingly, our experiments show that (iterative) pruning-based methods are substantially better than the HashedNet architecture, whose compression doesn't appear advantageous to a carefully chosen convolutional network. We also show that, when the compression level is high, the famous OBD pruning heuristics deteriorates to the point of being less efficient than simple magnitude-based techniques.

MARACAS Team

7. New Results

7.1. Results of axis 1: fundamental limits

We worked in 2019 on the following main research directions:

1. Fundamental limits of IoT networks

Table 1.

Principal Investigators:	Malcolm Egan, Samir Perlaza, Jean-Marie Gorce
Students:	Dadja Toussaint Anade-Akbo, Lélio Chetot
Funding:	Orange Labs, ANR Arburst
Partners:	Philippe Mary (IETR, Rennes), Laurent Clavier (IRCICA, Lille) JM Kélif (Orange Labs) H. Vincent Poor (Princeton University, NJ, USA)
Publications:	[34], [48], [36], [49], [35]

One of the main figures of merit in an IoT cell is the capability to support a massive access from distributed nodes, but with very small information quantity [12]. This perspective raises fundamental questions relative to the theoretical limits and performance of this kind of very large scale deployments. Fundamental limits are neither well known nor even well formulated. What is the maximal number of IoT nodes we may deploy in a given environment? At which energetic cost? With which transmission reliability or latency? These multiple questions highlight that the problem is not unique and the capacity is not the only (and even not the main) challenge to be addressed. We aim at establishing the fundamental limits of a decentralized system in a bursty regime which includes short packets of information and impulsive interference regime. We are targeting the fundamental limits and their mathematical expression, according to the usual information theory framework capturing the capacity region by establishing converse and achievability theorems.

2. Stability and sensitivity of fundamental limits

Table 2.

Principal Investigator:	Malcolm Egan, Samir Perlaza
Students:	-
Funding:	
Partners:	H. Vincent Poor, Alex Disto, Princeton University Vyacheslav Kungurtsev, Czech Technical University
Publications:	[8],[33]

The analysis of the fundamental limits on communications systems is performed under some assumptions including Gaussian noise, channel input symbols with average power, among others. Nonetheless, despite that these constraints were well suited for describing communications systems in the early 90's, the evolution of these systems make these assumptions vacuous today. Often, noise is better described by α–stable stochastic processes in IoT networks and channel inputs are subject to constraints in the amplitude, energy harvesting etc. From this perspective, our contributions are based on the notion of capacity sensitivity to study the capacity of continuous memoryless point-to-point channels. The capacity sensitivity reflects how the capacity changes with small perturbations in any of the parameters describing the channel, e.g., cost constraints on the input distribution as well as on the noise distribution.

3. Energy self-sustained wireless networks

Table 3.

Principal Investigator:	Samir Perlaza
Students:	Nizar Khalfet
Funding:	H2020 ComMed
Partners:	I. Kikridis (U. of Cyprus)
Publications:	[29], [42], [43], [50]

The main scientific challenge is to set up a theoretical framework for designing and developing fully decentralized energy-self-sustained communications systems. The main motivation stems from the fact that wireless networks deployed in hard-to-reach places, e.g., remote geographical areas, concrete structures, human body or war zones are often limited by the lifetime of their batteries. This contrasts with the fact that hardware is built to last for very long periods. One of the solutions being considered today for solving the energy limitation problem is the use of energy harvesting (EH) techniques. Within this context, our work focuses on the study of wireless communications systems based on EH sources. EH is expected to be the enabler of energy self-sustainability by eliminating the critical dependence on manual battery recharging.

However, a solid answer on whether or not EH is a viable solution can be given only if the corresponding fundamental limits of data transmission based on EH are known. This is mainly because these limits are based on the laws of Physics and thus, determine the barrier between feasible and unfeasible systems. We study the fundamental limits of three strongly correlated problems regarding the energy supply of future wireless networks: (i) Data transmission over centralized and decentralized EH multi-user channels; (ii) Simultaneous energy and information transmission in multi-user channels; and (iii) Energy cooperation. In a near future, we expect to exploit these results to design algorithms and protocols and later to perform a proof of concept on FIT/CorteXlab. We believe that a solid theoretical framework may help to drive the future design and performance evaluation of applications involving EH based wireless communications systems within smart buildings, smart cities.

4. Security and Privacy

Table 4.

Principal Investigator:	Samir Perlaza
Students:	David Kibloff
Funding:	Inria-DGA PhD
Partners:	Guilaume Villemaud (Socrate), Ligong Wang (ETIS, Cergy) Raphael Shaeffer (TU Berlin)
Publications:	[44], [51]

Information theory is also well adapted to study the fundamental limits of privacy and secrecy. Indeed, the wiretap channel and the covert communication [53] models have been shown to be appropriate for privacy preserving communications in wireless communications. With the PhD of David Kibloff defended in October 2019, we explored the following problem. Given a code used to send a message to two receivers through a degraded discrete memoryless broadcast channel (DM-BC), the sender wishes to alter the codewords to achieve the following goals: (i) the original broadcast communication continues to take place, possibly at the expense of a tolerable increase of the decoding error probability; and (ii) an additional covert message can be transmitted to the stronger receiver such that the weaker receiver cannot detect the existence of this message. The main results are: (a) feasibility of covert communications is proven by using a random coding argument for general DM-BCs; and (b) necessary conditions for establishing covert communications are described and an impossibility (converse) result is presented for a particular class of DM-BCs. Together, these results characterize the asymptotic fundamental limits of covert communications for this particular class of DM-BCs within an arbitrarily small gap. Future extensions will concern the Gaussian and other continuous channels, or more complex scenarios where some subsets of nodes are willing to communicate while some external observers cannot even detect the existence of these messages. Covert communication allows to introduce a side constraint that prevent a network to be attacked.

5. Structured Codes for Quantization and Channel Estimation

Table 5.

Principal Investigator:	Malcolm Egan
Publications:	[25]

Finite frames are sequences of vectors in finite dimensional Hilbert spaces that play a key role in signal processing and coding theory. In this work, we study the class of tight unit-norm frames for $\mathbb{C}^d$ that also form regular schemes, which we call tight regular schemes (TRS). Many common frames that arising in vector quantization and channel state estimation, such as equiangular tight frames and mutually unbiased bases, fall in this class. We investigate characteristic properties of TRSs and prove that for many constructions, they are intimately connected to weighted 1-designs—arising from cubature rules for integrals over spheres in $\mathbb{C}^d$—with weights dependent on the Voronoi regions of each frame element. Aided by additional numerical evidence, we conjecture that all TRSs in fact satisfy this property.

7.2. Results of axis 2: algorithms

1. Massive random access in LPWAN

Table 6.

Principal Investigator:	Jean-Marie Gorce, Claire Goursaud
Students:	Diane Duchemin, Lélio Chetot
Funding:	ANR Ephyl, Inria-Nokia common lab
Partners:	Sequans, Supelec Rennes, ISEP, CEA Leti, Nokia
Publications:	[30], [31], [37], [47]

The optimization of IoT access techniques was the objective of the ANR Ephyl collaborative project, where we studied different solutions at the PHY and MAC layers as presented in [47].

The main question Maracas group addressed in this research is the detection of simultaneous random transmissions from distributed nodes. The underlying mechanism is a coded slotted Aloha allowing to avoid hand-skake mechanisms. Each node can transmit randomly and the receiver tries to detect several packets simultaneously. Our objective is to identify a good code family, and to determine the fundamental trade-off in terms of nodes density versus reliability. During this year, we focused on the detection of a small subset of simultaneous active nodes, exploiting optimal detection. We developed a MAP based iterative detector at a multi-antennas receiver in [30]. We also proposed a low complexity detector in [37].

This joint coding-decoding optimization problem will be also investigated from extensive simulations and experimental data (see section 3.4), and represents an interesting problem to evaluate deep learning based approaches.

2. Interference management

Table 7.

Principal Investigator:	Léonardo Cardoso, Jean-Marie Gorce
Students:	Hassan Khalam
Funding:	Fed4PMR (PIA)
Partners:	Thales
Publications:	[41]

Interference management and resource management is a very complex problem in wireless environment (e.g. [55]). The capacity region is known for some specific scenarios and some specific channel conditions. But the optimal performance relies on perfect feedback mechanisms, to get channel state information at the transmitters and to coordinate them. As proposed by Jafar et al, topological interference management (TIM) [56] is a seducing framework to balance performance with feedback complexity. In the context of the Fed4PMR project, we develop new algorithms to allow partial coordination between interfering transmitters [41], relying only on some partial interference information. This approach suits particularly well with the requirements of PMR networks, since their deployments is not optimized. The algorithm relies on an association of degrees of freedom evaluation, graph theory and interference alignment.

Based on this first study, we will explore the suitability of TIM in other application scenarios (especially for the standard IEE802.11ax under preparation). For short, TIM allows to build optimal graph representations of a wireless networks, with reduced coordination needs. TIM can be seen as an approach to optimally quantize a complex interfering graph and to distribute its knowledge in an optimal fashion.

3. Learning in radio systems

Table 8.

Principal Investigator:	Léonardo Cardoso, Malcolm Egan, Jean-Marie Gorce
Student:	Cyrille Morin, Mathieu Goutay
Funding:	ADR Analytics, Inria-Nokia common lab AI chair ANR program (applied)
Partners:	Jakob Hoydis, Nokia Bell Labs
Publications:	[45]

Following the artificial intelligence tsunami, the research community in wireless systems (both industry and academia) is engaged in a strong competition to determine how this revolution could change the paradigm of wireless networks. Following the preliminary studies made by Jakob Hoydis [54], we investigate in this research action, the potential of deep learning in radio communications. The central question is to identify which processing could take advantage from neural networks against classical approaches.

Our joint strategy with Nokia follows: we target the production of a huge set of experimental data with FIT/CorteXlab to facilitate the comparison of different solutions and to train neural networks on real data. We currently investigate three original problems : transmitter identification from its RF signature (Cyrille Morin PhD) [45], self-synchronization procedures based on neural networks (Cyrille Morin PhD) and dirty RF compensation (Mathieu Goutay PhD, patents submitted). Last but not least, we believe that an intelligent radio should be able to learn from its environment and to adapt its behavior. Therefore, in the future, we will explore reinforcement principles associated to neural networks and applied to learning based radio.

This topic is very hot, and most top ranked conference have special sessions on this topic. We believe that our partnership with Nokia, our data sets from FIT/CorteXlab and our experience in estimation theory let us be highly competitive.

7.3. Results of axis 3: experimental assessment

During 2019-2020, our experimental work was mostly devoted to the development of new functions of FIT/CorteXlab, and to the development of experimental evaluations with external partners.

1. Development of a user and administrative graphical interface

Table 9.

Principal Investigator:	Pascal Girard, Matthieu Imbert, Léonardo Cardoso
Funding:	FIT/CorteXlab
Partners:	FIT consortium

The objective is to develop a web-based user-friendly interface for using CorteXlab. Several modules are planned and the first module is the user management module, which aims at easing platform usage and improving the metadata that we can associate with each experimenter and experiment. This metadata aims at improving the metrics we can gather about the platform's usage.

2. Development of a docker-based experiment conducting middleware.

Table 10.

Principal Investigator:	Matthieu Imbert, Léonardo Cardoso
Funding:	FIT/CorteXlab
Partners:	FIT consortium

CorteXlab relies on Minus, an experiment conducting middleware which allows users to submit experimental tasks to the platform, handles the automatic execution of these experiments, and gathers their results. The initial design for Minus relies on a fixed toolchain (mainly composed of GNURadio, hardware drivers, and additional external or in-house software or GNURadio blocks, FPGA tools, etc.). Experimenters are supposed to use this fixed toolchain in a batch-like workflow. It is hard for experimenters to extend the limits of the fixed toolchain (e.g. to use a custom library or software, or a different version of GNURadio), and the development phase of an experiment can be painful due to the batch-like interface. To improve this, we have developed a new experimental workflow based on docker [61] images and containers which allows experimenters to use our in-house provided docker images [52], adapt them if needed, or even create completely custom ones. These images have the benefit that they can be used identically on the experimenters' workstations, on the CorteXlab platform, or another platform, and they can be used interactively if needed, even on CorteXlab. This increases greatly the ease of use of the platform, the reproducibility and share-ability of experiments, and the breadth of its usage.

3. Reference scenario for massive IoT access

Table 11.

Principal Investigator:	Othmane Oubejja, Jean-Marie Gorce Matthieu Imbert, Léonardo Cardoso
Funding:	ANR Ephyl, ANR ARburst
Partners:	CEA Leti, Supelec Rennes, Sequans
Publications:	[46]

In this work we developed an experimental setup for dense IoT access evaluation, as part of the project "Enhanced Physical Layer for Cellular IoT" (EPHYL), using FIT/CorteXlab radio testbed. The aim of this work is to provide a customizable and open source design for IoT networks prototyping in a massive multi-user, synchronized and reproducible environment thanks to the hardware and software capabilities of the testbed. The massive access feature is managed by emulating a base station and several sensors per radio nodes. As shown in Fig.4 , two categories of modular network components are used in our design: a base station unit and a multi-sensor emulator unit. These components are separately hosted in dedicated and remotely accessible radio nodes.

The features of this design can be accessed through customizable demos as documentation and resources are available online. As a result, it is possible for any interested user to plug custom algorithms, evaluate diverse communication scenarios and perform necessary physical measurements.

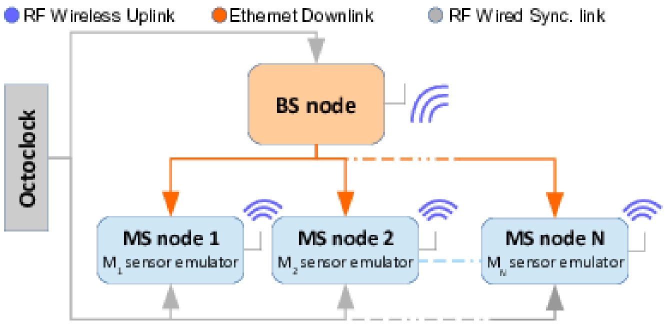

Figure 4. EPHYL IoT network representation

7.4. Results of axis 4: other application fields

1. Smart Grid

Table 12.

Principal Investigators:	Samir Perlaza
Student:	Matei Moldoveanu (visitor)
Partners:	Inaki Esnaola
Publications:	[40]

We study the recovery of missing data from multiple smart grid datasets within a matrix completion framework. The datasets contain the electrical magnitudes required for monitoring and control of the electricity distribution system. Each dataset is described by a low rank matrix. Different datasets are correlated as a result of containing measurements of different physical magnitudes generated by the same distribution system. To assess the validity of matrix completion techniques in the recovery of missing data, we characterize the fundamental limits when two correlated datasets are jointly recovered. We then proceed to evaluate the performance of Singular Value Thresholding (SVT) and Bayesian SVT (BSVT) in this setting. We show that BSVT outperforms SVT by simulating the recovery for different correlated datasets. The performance of BSVT displays the tradeoff behaviour described by the fundamental limit, which suggests that BSVT exploits the correlation between the datasets in an efficient manner.

2. Molecular Communications

Some of the most ambitious applications of molecular communications are expected to lie in nanomedicine and advanced manufacturing. In these domains, the molecular communication system is surrounded by a range of biochemical processes, some of which may be sensitive to chemical species used for communication. Under these conditions, the biological system and the molecular communication system impact each other. As such, the problem of coexistence arises, where both

Table 13.

Principal Investigators:	Malcolm Egan
Postdoc:	Bayram Akdeniz
Funding:	Inria Projet Recherche Exploratoire (PRE)
Partners:	Valeria Loscri (FUN Team, Inria) Marco Di Renzo (CNRS), Bao Tang (University of Graz, Austria) Trung Duong (Queen's University Belfast) Ido Nevat (TUMCREATE, Singapore)
Publications:	[38], [39], [24], [26]

the reliability of the molecular communication system and the function of the biological system must be ensured. In this paper, we study this problem with a focus on interactions with biological systems equipped with chemosensing mechanisms, which arises in a large class of biological systems. We motivate the problem by considering chemosensing mechanisms arising in bacteria chemo-taxis, a ubiquitous and well-understood class of biological systems. We then propose strategies for a molecular communication system to minimize disruption of biological system equipped with a chemosensing mechanism. This is achieved by exploiting tools from the theory of chemical reaction networks. To investigate the capabilities of our strategies, we obtain fundamental information theoretic limits by establishing a new connection with the problem of covert communications.

3. Intelligent Transportation

Table 14.

Principal Investigators:	Malcolm Egan
Partners:	Michel Jakob (Czech Technical University in Prague), Nir Oren (University of Aberdeen)
Publications:	[27]

Market mechanisms are now playing a key role in allocating and pricing on-demand transportion services. In practice, most such services use posted-price mechanisms, where both passengers and drivers are offered a journey price which they can accept or reject. However, providers such as Liftago and GrabTaxi have begun to adopt a mechanism whereby auctions are used to price drivers. These latter mechanisms are neither posted-price nor classical double auctions, and can instead be considered a hybrid mechanism. In this work, we develop and study the properties of a novel hybrid on-demand transport mechanism. Due to the need for incorporating statistical knowledge and communication of system state information, communication-theoretic methods can play a useful role.

In particular, as these mechanisms require knowledge of passenger demand, we analyze the data-profit tradeoff as well as how passenger and driver preferences influence mechanism performance. We show that the revenue loss for the provider scales with $\sqrt{n \log n}$ for n passenger requests under a multi-armed bandit learning algorithm with beta distributed preferences. We also investigate the effect of subsidies on both profit and the number of successful journeys allocated by the mechanism, comparing these with a posted-price mechanism, showing improvements in profit with a comparable number of successful requests.

NEO Project-Team

7. New Results

7.1. Stochastic Modeling

Participants: Sara Alouf, Eitan Altman, Konstantin Avrachenkov, Alain Jean-Marie, Giovanni Neglia.

7.1.1. Network growth models

Network growth models that embody principles such as preferential attachment and local attachment rules have received much attention over the last decade. Among various approaches, random walks have been leveraged to capture such principles. In the framework of joint team with Brazil (Thanes), G. Neglia, together with G. Iacobelli and D. Figueiredo (both from UFRJ, Brazil), has studied a simple model where network growth and a random walker are coupled [23]. In particular, they consider the No Restart Random Walk model where a walker builds its graph (tree) while moving around. The walker takes s steps (a parameter) on the current graph. A new node with degree one is added to the graph and connected to the node currently occupied by the walker. The walker then resumes, taking another s steps, and the process repeats. They have analyzed this process from the perspective of the walker and the network, showing a fundamental dichotomy between transience and recurrence for the walker as well as power law and exponential degree distribution for the network.

7.1.2. Controlled Markov chains

E. Altman in collaboration with D. Josselin and S. Boularouk (CERI/LIA, Univ Avignon) study in [26] a multiobjective dynamic program where all the criteria are in the form of total expected sum of costs till absorption in some set of states. They assume that instantaneous costs are strictly positive and make no assumption on the ergodic structure of the Markov Decision Process. Their main result is to extend the linear program solution approach that was previously derived for transient Constrained Markov Decision Processes to the general ergodic structure. Several (additive) cost metrics are defined and (possibly randomized) routing policies are sought which minimize one of the costs subject to constraints over the other objectives.

7.1.3. Escape probability estimation in large graphs

Consider large graphs as the object of study and specifically the problem of escape probability estimation. Generally, this characteristic cannot be calculated analytically nor even numerically due to the complexity and large size of the investigation object. In [32], K. Avrachenkov and A. Borodina (Karelian Institute of Applied Mathematical Research, Russia) have presented an effective method for estimating the probability that the random walk on graph first enters a node b before returning into the starting node a. Regenerative properties of the random walk allow the use of an accelerated method for the simulation of cycles based on the splitting technique. The results of numerical experiments confirm the advantages of the proposed method.

7.1.4. Random surfers and prefetching

Prefetching is a basic technique used to reduce the latency of diverse computer services. Deciding what to prefetch amounts to make a compromise between latency and the waste of resources (network bandwidth, storage, energy) if contents is mistakenly prefetched. Modeling the problem in case of web/video/gaming navigation, is done by identifying a graph of “documents” connected by links representing the possible chaining. A surfer, either random or strategic, browses this graph. The prefetching controller must make it sure that the documents browsed are always available locally. In the case where the surfer is random and/or the graph is not completely known in advance, the question is largely unexplored. Q. Petitjean, under the supervision of S. Alouf and A. Jean-Marie, has determined through extensive simulations that when the graph is a tree, neither the greedy strategy, nor the one optimal when the tree is completely known, are optimal when the tree is discovered progressively.

7.1.5. The marmoteCore platform

The development of marmoteCore (see Section 6.1) has been pursued. Its numerical features for computing stationary distributions, average hitting times and absorption probabilities have been used in a joint work with F. Cazals, D. Mazauric and G. Santa Cruz (ABS team) and J. Roux (Univ Cote d'Azur) [52]. The software has been presented to young researchers in networking at the ResCom 2019 summer school.

7.2. Random Graph and Matrix Models

Participants: Konstantin Avrachenkov, Andrei Bobu.

7.2.1. Random geometric graphs

Random geometric graphs are good examples of random graphs with a tendency to demonstrate community structure. Vertices of such a graph are represented by points in Euclid space R^d, and edge appearance depends on the distance between the points. Random geometric graphs were extensively explored and many of their basic properties are revealed. However, in the case of growing dimension $d \to \infty$ practically nothing is known; this regime corresponds to the case of data with many features, a case commonly appearing in practice. In [30], K. Avrachenkov and A. Bobu focus on the cliques of these graphs in the situation when average vertex degree grows significantly slower than the number of vertices n with $n \to \infty$ and $d \to \infty$. They show that under these conditions random geometric graphs do not contain cliques of size 4 a.s. As for the size 3, they present new bounds on the expected number of triangles in the case $\log^2(n) \ll d \ll \log^3(n)$ that improve previously known results.

Network geometries are typically characterized by having a finite spectral dimension (SD), that characterizes the return time distribution of a random walk on a graph. The main purpose of this work is to determine the SD of random geometric graphs (RGGs) in the thermodynamic regime, in which the average vertex degree is constant. The spectral dimension depends on the eigenvalue density (ED) of the RGG normalized Laplacian in the neighborhood of the minimum eigenvalues. In fact, the behavior of the ED in such a neighborhood characterizes the random walk. Therefore, in [33] K. Avrachenkov together with L. Cottatellucci (FAU, Germany and Eurecom) and M. Hamidouche (Eurecom) first provide an analytical approximation for the eigenvalues of the regularized normalized Laplacian matrix of RGGs in the thermodynamic regime. Then, we show that the smallest non zero eigenvalue converges to zero in the large graph limit. Based on the analytical expression of the eigenvalues, they show that the eigenvalue distribution in a neighborhood of the minimum value follows a power-law tail. Using this result, they find that the SD of RGGs is approximated by the space dimension d in the thermodynamic regime.

In [42] K. Avrachenkov together with L. Cottatellucci (FAU, Germany and Eurecom) and M. Hamidouche (Eurecom) have analyzed the limiting eigenvalue distribution (LED) of random geometric graphs. In particular, they study the LED of the adjacency matrix of RGGs in the connectivity regime, in which the average vertex degree scales as $\log(n)$ or faster. In the connectivity regime and under some conditions on the radius r, they show that the LED of the adjacency matrix of RGGs converges to the LED of the adjacency matrix of a deterministic geometric graph (DGG) with nodes in a grid as the size of the graph n goes to infinity. Then, for n finite, they use the structure of the DGG to approximate the eigenvalues of the adjacency matrix of the RGG and provide an upper bound for the approximation error.

7.3. Data Analysis and Learning

Participants: Konstantin Avrachenkov, Maximilien Dreveton, Giovanni Neglia, Chuan Xu.

7.3.1. Almost exact recovery in label spreading

In semi-supervised graph clustering setting, an expert provides cluster membership of few nodes. This little amount of information allows one to achieve high accuracy clustering using efficient computational procedures. Our main goal is to provide a theoretical justification why the graph-based semi-supervised learning works very well. Specifically, for the Stochastic Block Model in the moderately sparse regime, in

[34] K. Avrachenkov and M. Dreveton have proved that popular semi-supervised clustering methods like Label Spreading achieve asymptotically almost exact recovery as long as the fraction of labeled nodes does not go to zero and the average degree goes to infinity.

7.3.2. Similarities, kernels and proximity measures on graphs

In [13], K. Avrachenkov together with P. Chebotarev (RAS Trapeznikov Institute of Control Sciences, Russia) and D. Rubanov (Google) have analytically studied proximity and distance properties of various kernels and similarity measures on graphs. This helps to understand the mathematical nature of such measures and can potentially be useful for recommending the adoption of specific similarity measures in data analysis.

7.3.3. The effect of communication topology on learning speed

Many learning problems are formulated as minimization of some loss function on a training set of examples. Distributed gradient methods on a cluster are often used for this purpose. In [47], G. Neglia, together with G. Calbi (Univ Côte d'Azur), D. Towsley, and G. Vardoyan (UMass at Amherst, USA), has studied how the variability of task execution times at cluster nodes affects the system throughput. In particular, a simple but accurate model allows them to quantify how the time to solve the minimization problem depends on the network of information exchanges among the nodes. Interestingly, they show that, even when communication overhead may be neglected, the clique is not necessarily the most effective topology, as commonly assumed in previous works.

In [48] G. Neglia and C. Xu, together with D. Towsley (UMass at Amherst, USA) and G. Calbi (Univ Côte d'Azur) have investigated why the effect of the communication topology on the number of epochs needed for machine learning training to converge appears experimentally much smaller than what predicted by theory.

7.4. Game Theory

Participants: Eitan Altman, Konstantin Avrachenkov, Mandar Datar, Swapnil Dhamal, Alain Jean-Marie.

7.4.1. Resource allocation: Kelly mechanism and Tullock game

The price-anticipating Kelly mechanism (PAKM) is one of the most extensively used strategies to allocate divisible resources for strategic users in communication networks and computing systems. It is known in other communities as the Tullock game. The users are deemed as selfish and also benign, each of which maximizes his individual utility of the allocated resources minus his payment to the network operator. E. Altman, A. Reiffers-Masson (IISc Bangalore, India), D. Sadoc-Menasche (UFJR, Brazil), M. Datar, S. Dhamal, C. Touati (Inria Grenoble-Rhone-Alpes) and R. El-Azouzi (CERI/LIA, Univ Avignon) have first applied this type of games to competition in crypto-currency protocols between miners in blockchain [11]. Blockchain is a distributed synchronized secure database containing validated blocks of transactions. A block is validated by special nodes called miners and the validation of each new block is done via the solution of a computationally difficult problem, which is called the proof-of-work puzzle. The miners compete against each other and the first to solve the problem announces it, the block is then verified by the majority of miners in this network, trying to reach consensus. After the propagated block reaches the consensus, it is successfully added to the distributed database. The miner who found the solution receives a reward either in the form of crypto-currencies or in the form of a transaction reward. The authors show that the discrete version of the game is equivalent to a congestion game and thus has an equilibrium in pure strategies.

E. Altman, M. Datar, C. Touati (Inria Grenoble-Rhone-Alpes) and G. Burnside (Nokia Bell Labs) then introduce further constraints on the total amount of resources used and study pricing issues in this constrained game. They show that a normalized equilibrium (in the sense of Rosen) exists which implies that pricing can be done in a scalable way, i.e; prices can be chosen to be independent of the player. A possible way to prove this structure is to show that the utilities are strict diagonal concave (which is an extension to game setting of concavity) which they did in [27].

In [25], Y. Xu, Z. Xiao, T. Ni, X. Wang (all from Fudan Univ, China), J. H. Wang (Tsinghua Univ, China) and E. Altman formulate a non-cooperative Tullock game consisting of a finite amount of benign users and one misbehaving user. The maliciousness of this misbehaving user is captured by his willingness to pay to trade for unit degradation in the utilities of benign users. The network operator allocates resources to all the users via the price-anticipating Kelly mechanism. They present six important performance metrics with regard to the total utility and the total net utility of benign users, and the revenue of network operator under three different scenarios: with and without the misbehaving user, and the maximum. We quantify the robustness of PAKM against the misbehaving actions by deriving the upper and lower bounds of these metrics.

7.4.2. A stochastic game with non-classical information structure

In [44], V. Kavitha, M. Maheshwari (both from IIT Bombay, India) and E. Altman introduce a stochastic game with partial, asymmetric and non-classical information. They obtain relevant equilibrium policies using a new approach which allows managing the belief updates in a structured manner. Agents have access only to partial information updates, and their approach is to consider optimal open loop control until the information update. The agents continuously control the rates of their Poisson search clocks to acquire the locks, the agent to get all the locks before others would get reward one. However, the agents have no information about the acquisition status of others and will incur a cost proportional to their rate process. The authors solved the problem for the case with two agents and two locks and conjectured the results for a general number of agents. They showed that a pair of (partial) state-dependent time-threshold policies form a Nash equilibrium.

7.4.3. Zero-Sum stochastic games over the field of real algebraic numbers

In [14], K. Avrachenkov together with V. Ejov (Flinders Univ, Australia), J. Filar and A. Moghaddam (both from Univ of Queensland, Australia) have considered a finite state, finite action, zero-sum stochastic games with data defining the game lying in the ordered field of real algebraic numbers. In both the discounted and the limiting average versions of these games, they prove that the value vector also lies in the same field of real algebraic numbers. Their method supplies finite construction of univariate polynomials whose roots contain these value vectors. In the case where the data of the game are rational, the method also provides a way of checking whether the entries of the value vectors are also rational.

7.4.4. Evolutionary Markov games

I. Brunetti (CIRED), Y. Hayel (CERI/LIA, Univ Avignon) and E. Altman extend in [59] evolutionary game theory by introducing the concept of individual state. They analyze a particular simple case, in which they associate a state to each player, and suppose that this state determines the set of available actions. They consider deterministic stationary policies and suppose that the choice of a policy determines the fitness of the player and it impacts the evolution of the state. They define the interdependent dynamics of states and policies and introduce the State Policy coupled Dynamics in order to study the evolution of the population profile. They prove the relation between the rest points of the system and the equilibria of the game. Then they assume that the processes of states and policies move with different velocities: this assumption allows them to solve the system and then find the equilibria of the game with two different methods: the singular perturbation method and a matrix approach.

7.4.5. Stochastic replicator dynamics

In [12], K. Avrachenkov and V.S. Borkar (IIT Bombay, India) have considered a novel model of stochastic replicator dynamics for potential games that converts to a Langevin equation on a sphere after a change of variables. This is distinct from the models of stochastic replicator dynamics studied earlier. In particular, it is ill-posed due to non-uniqueness of solutions, but is amenable to the Kolmogorov selection principle that picks a unique solution. The model allows us to make specific statements regarding metastable states such as small noise asymptotics for mean exit times from their domain of attraction, and quasi-stationary measures. We illustrate the general results by specializing them to replicator dynamics on graphs and demonstrate that the numerical experiments support theoretical predictions.

7.4.6. Stochastic coalitional better-response dynamics for finite games with application to network formation games

In [57], K. Avrachenkov and V.V. Sing (IIT Delhi, India) have considered coalition formation among players in 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 $\mathcal{K}$-stable equilibrium or becomes stuck in a closed cycle. The authors also assume that at each time a selected coalition makes mistake in deviation with small probability that add mutations (perturbations) into CBR dynamics. They prove that all $\mathcal{K}$-stable equilibria and all action profiles from closed cycles, that have minimum stochastic potential, are stochastically stable. A similar statement holds for strict $\mathcal{K}$-stable equilibrium. They 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.

7.4.7. Strong Stackelberg equilibria in stochastic games

In a joint work with V. Bucarey López (Univ Libre de Bruxelles, Belgium and Inria team INOCS), E. Della Vecchia (Univ Nacional de Rosario, Argentina), and F. Ordóñez (Univ de Chile, Chile), A. Jean-Marie has considered Stackelberg equilibria for discounted stochastic games. The motivation originates in applications of Game Theory to security issues, but the question is of general theoretical and practical relevance. The solution concept of interest is that of Stationary Strong Stackelberg Equlibrium (SSSE) policies: both players apply state feedback policies; the leader announces her strategy and the follower plays a best response to it. Tie breaks are resolved in favor of the leader. The authors provide classes of games where the SSSE exists, and we prove via counterexamples that SSSE does not exist in the general case. They define suitable dynamic programming operators whose fixed points are referred to as Fixed Point Equilibrium (FPE). They show that the FPE and SSSE coincide for a class of games with Myopic Follower Strategy. Numerical examples shed light on the relationship between SSSE and FPE and the behavior of Value Iteration, Policy Iteration and Mathematical programming formulations for this problem. A security application illustrates the solution concepts and the efficiency of the algorithms introduced. The results are presented in [67], [50], [51].

7.4.8. Routing on a ring network

R. Burra, C. Singh and J. Kuri (IISc Bangalore, India), study in [60] with E. Altman routing on a ring network in which traffic originates from nodes on the ring and is destined to the center. The users can take direct paths from originating nodes to the center and also multihop paths via other nodes. The authors show that routing games with only one and two hop paths and linear costs are potential games. They give explicit expressions of Nash equilibrium flows for networks with any generic cost function and symmetric loads. They also consider a ring network with random number of users at nodes, all of them having same demand, and linear routing costs. They give explicit characterization of Nash equilibria for two cases: (i) General i.i.d. loads and one and two hop paths, (ii) Bernoulli distributed loads. They also analyze optimal routing in each of these cases.

7.4.9. Routing games applied to the network neutrality debate

The Network Neutrality issue has been at the center of debate worldwide lately. Some countries have established laws so that principles of network neutrality are respected. Among the questions that have been discussed in these debates there is whether to allow agreements between service and content providers, i.e. to allow some preferential treatment by an operator to traffic from some providers (identity-based discrimination). In [63], A. Reiffers-Masson (IISc Bangalore), Y. Hayel, T. Jimenez (CERI/LIA, Univ Avignon) and E. Altman, study this question using models from routing games.

7.4.10. Peering vs transit: A game theoretical model for autonomous systems connectivity

G. Accongiagioco (IMT, Italy), E. Altman, E. Gregori (Institute of Informatics and Telematics, Univ Pisa) and Luciano Lenzini (Dipartimento di Informatica, Univ Pisa) propose a model for network optimization in a non-cooperative game setting with specific reference to the Internet connectivity. The model describes the decisions taken by an Autonomous System when joining the Internet. They first define a realistic model for the interconnection costs incurred; then they use this cost model to perform a game theoretic analysis of the decisions related to link creation and traffic routing, keeping into account the peering/transit dichotomy. The proposed model does not fall into the standard category of routing games, hence they devise new tools to solve it by exploiting specific properties of the game. They prove analytically the existence of multiple equilibria.

7.4.11. Altruistic behavior and evolutionary games

Within some species like bees or ants, the one who interacts is not the one who reproduces. This implies that the Darwinian fitness is related to the entire swarm and not to a single individual and thus, standard Evolutionary Game models do not apply to these species. Furthermore, in many species, one finds altruistic behaviors, which favors the group to which the playing individual belongs, but which may hurt the single individual. In [58], [62], I. Brunetti (CIRED), R. El-Azouzi, M. Haddad, H. Gaiech, Y. Hayel (LIA/CERI, Univ Avignon) and E. Altman define evolutionary games between group of players and study the equilibrium behavior as well as convergence to equilibrium.

7.5. Applications in Telecommunications

Participants: Eitan Altman, Konstantin Avrachenkov, Giovanni Neglia.

7.5.1. Elastic cloud caching services

In [37], G. Neglia, together with D. Carra (Univ of Verona, Italy) and P. Michiardi (Eurecom), has considered in-memory key-value stores used as caches, and their elastic provisioning in the cloud. The cost associated to such caches not only includes the storage cost, but also the cost due to misses: in fact, the cache miss ratio has a direct impact on the performance perceived by end users, and this directly affects the overall revenues for content providers. The goal of their work is to adapt dynamically the number of caches based on the traffic pattern, to minimize the overall costs. They present a dynamic algorithm for TTL caches whose goal is to obtain close-to-minimal costs and propose a practical implementation with limited computational complexity: their scheme requires constant overhead per request independently from the cache size. Using real-world traces collected from the Akamai content delivery network, they show that their solution achieves significant cost savings specially in highly dynamic settings that are likely to require elastic cloud services.

7.5.2. Neural networks for caching

In [19] G. Neglia, together with V. Fedchenko (Univ Côte d'Azur) and B. Ribeiro (Purdue Univ, USA), has proposed a caching policy that uses a feedforward neural network (FNN) to predict content popularity. This scheme outperforms popular eviction policies like LRU or ARC, but also a new policy relying on the more complex recurrent neural networks. At the same time, replacing the FNN predictor with a naive linear estimator does not degrade caching performance significantly, questioning then the role of neural networks for these applications.

7.5.3. Similarity caching

In similarity caching systems, a user request for an object o that is not in the cache can be (partially) satisfied by a similar stored object o', at the cost of a loss of user utility. Similarity caching systems can be effectively employed in several application areas, like multimedia retrieval, recommender systems, genome study, and machine learning training/serving. However, despite their relevance, the behavior of such systems is far from being well understood. In [41], G. Neglia, together with M. Garetto (Univ of Turin, Italy) and E. Leonardi (Politechnic of Turin, Italy), provides a first comprehensive analysis of similarity caching in the offline, adversarial, and stochastic settings. They show that similarity caching raises significant new challenges, for which they propose the first dynamic policies with some optimality guarantees. They evaluate the performance of the proposed schemes under both synthetic and real request traces.

7.5.4. Performance evaluation and optimization of 5G wireless 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, one idea is 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. In [16], A. Chattopadhyay and B. Błaszczyszyn (Inria DYOGENE team) in collaboration with E. Altman use the classical two-tier Poisson network model with different transmit powers, 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, they calculate the average downlink data rate of the typical static and mobile (i.e., moving) users, the latter accounted for handoff outage periods. They consider also the average throughput of these two types of users.

In [15], the same authors 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, the authors 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. They 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, they propose a learning algorithm based on single timescale stochastic approximation. Also, noting that the unconstrained MDP can be used to solve a constrained problem, they 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 their scheme, and also the trade-off between performance gain and fairness.

7.5.5. The age of information

Two decades after the seminal paper on software aging and rejuvenation appeared in 1995, a new concept and metric referred to as the age of information (AoI) has been gaining attention from practitioners and the research community. In the vision paper [46], D.S. Menasche (UFRJ, Brazil), K. Trivedi (Duke Univ, USA) and E. Altman show the similarities and differences between software aging and information aging. In particular, modeling frameworks that have been applied to software aging, such as the semi Markov approach can be immediately applied in the realm of age of information. Conversely, they indicate that questions pertaining to sampling costs associated with the age of information can be useful to assess the optimal rejuvenation trigger interval for software systems.

The demand for Internet services that require frequent updates through small messages has tremendously grown in the past few years. Although the use of such applications by domestic users is usually free, their access from mobile devices is subject to fees and consumes energy from limited batteries. If a user activates his mobile device and is in the range of a publisher, an update is received at the expense of monetary and energy costs. Thus, users face a tradeoff between such costs and their messages aging. It is then natural to ask how to cope with such a tradeoff, by devising aging control policies. An aging control policy consists of deciding, based on the utility of the owned content, whether to activate the mobile device, and if so, which technology to use (WiFi or cellular). In [28] E. Altman, R. El-Azouzi (CERI/LIA, Univ Avignon), D.S. Menasche (UFRJ, Brazil) and Y. Xu (Fudan Univ, China) show the existence of an optimal strategy in the class of threshold strategies, wherein users activate their mobile devices if the age of their poadcasts surpasses a given threshold and remain inactive otherwise. The accuracy of their model is validated against traces from the UMass DieselNet bus network. The first version of this paper, among the first to introduce the age of information, appeared already in arXiv on 2010.

7.5.6. Wireless transmission vehicle routing

The Wireless Transmission Vehicle Routing Problem (WT-VRP) consists of searching for a route for a vehicle responsible for collecting information from stations. The new feature w.r.t. classical vehicle routing is the possibility of picking up information via wireless transmission, without visiting physically the stations of the network. The WT-VRP has applications in underwater surveillance and environmental monitoring. In [53], L. Flores Luyo and E. Ocaña Anaya (IMCA, Brazil), A. Agra (Univ Aveiro, Brazil), R. Figueiredo (CERI/LIA, Univ Avignon) and E. Altman, study three criteria for measuring the efficiency of a solution and propose a mixed integer linear programming formulation to solve the problem. Computational experiments were done to access the numerical complexity of the problem and to compare solutions under the three criteria proposed.

7.5.7. Video streaming in 5G cellular networks

Dynamic Adaptive Streaming over HTTP (DASH) has become the standard choice for live events and on-demand video services. In fact, by performing bitrate adaptation at the client side, DASH operates to deliver the highest possible Quality of Experience (QoE) under given network conditions. In cellular networks, in particular, video streaming services are affected by mobility and cell load variation. In this context, DASH video clients continually adapt the streaming quality to cope with channel variability. However, since they operate in a greedy manner, adaptive video clients can overload cellular network resources, degrading the QoE of other users and suffer persistent bitrate oscillations. In [40] R. El-Azouzi (CERI/LIA, Univ Avignon), A. Sunny (IIT Palakkad, India), L. Zhao (Huazhong Agricultural Univ, China), E. Altman, D. Tsilimantos (Huawei Technologies, France), F. De Pellegrini (CERI/LIA Univ Avignon), and S. Valentin (Darmstadt Univ, Germany) tackle this problem using a new scheduler at base stations, named Shadow-Enforcer, which ensures minimal number of quality switches as well as efficient and fair utilization of network resources.

While most modern-day video clients continually adapt quality of the video stream, they neither coordinate with the network elements nor among each other. Consequently, a streaming client may quickly overload the cellular network, leading to poor Quality of Experience (QoE) for the users in the network. Motivated by this problem, A. Sunny (IIT Palakkad, India), R. El-Azouzi, A. Arfaoui (both from CERI/LIA, Univ Avignon), E. Altman, S. Poojary (BITS, India), D. Tsilimantos (Huawei Technologies, France) and S. Valentin (Darmstadt Univ, Germany) introduce in [24] D-VIEWS — a scheduling paradigm that assures video bitrate stability of adaptive video streams while ensuring better system utilization. The performance of D-views is then evaluated through simulations.

In [39], R. El-Azouzi, K.V. Acharya (ENS Lyon), M. Haddad (CERI/LIA, Univ Avignon), S. Poojary (BITS, India), A. Sunny (IIT Palakkad, India), D. Tsilimantos (Huawei Technologies, France), S. Valentin (Darmstadt Univ, Germany) and E. Altman, develop an analytical framework to compute the Quality-of-Experience (QoE) metrics of video streaming in wireless networks. Their framework takes into account the system dynamics that arises due to the arrival and departure of flows. They also consider the possibility of users abandoning the system on account of poor QoE. Considering the coexistence of multiple services such as video streaming and elastic flows, they use a Markov chain based analysis to compute the user QoE metrics: probability of starvation, prefetching delay, average video quality and bitrate switching. The simulation results validate the accuracy of their model and describe the impact of the scheduler at the base station on the QoE metrics.

7.5.8. A learning algorithm for the Whittle index policy for scheduling web crawlers

In [31] K. Avrachenkov and V.S. Borkar (IIT Bombay, India) have revisited the Whittle index policy for scheduling web crawlers for ephemeral content and developed a reinforcement learning scheme for it based on LSPE(0). The scheme leverages the known structural properties of the Whittle index policy.

7.5.9. Distributed cooperative caching for VoD with geographic constraints

Consider the caching of video streams in a cellular network in which each base station is equipped with a cache. Video streams are partitioned into multiple substreams and the goal is to place substreams in caches such that the residual backhaul load is minimized. In [36] K. Avrachenkov together with J. Goseling (UTwente, The Netherlands) and B. Serbetci (Eurecom) have studied two coding mechanisms for the substreams: Layered coding (LC) mechanism and multiple description coding (MDC). They develop a distributed asynchronous algorithm for deciding which files to store in which cache to minimize the residual bandwidth, i.e., the cost for downloading the missing substreams of the user's requested video with a certain video quality from the gateway (i.e., the main server). They show that their algorithm converges rapidly. Finally, they show that MDC partitioning is better than the LC mechanism when the most popular content is stored in caches; however, their algorithm enables to use the LC mechanism as well without any performance loss.

Further, in [35], K. Avrachenkov together with J. Goseling (UTwente, The Netherlands) and B. Serbetci (Eurecom), have considered the same setting as above but maximized the expected utility. The utility depends on the quality at which a user is requesting a file and the chunks that are available. They impose alpha-fairness across files and qualities. Similarly to [36] they have developed a distributed asynchronous algorithm for deciding which chunks to store in which cache.

7.6. Applications in Social Networks

Participants: Eitan Altman, Swapnil Dhamal, Giovanni Neglia.

7.6.1. Utility from accessing an online social network

The retention of users on online social networks has important implications, encompassing economic, psychological and infrastructure aspects. In the framework of our joint team with Brazil (Thanes), G. Neglia, together with E. Hargreaves and D. Menasche (both from UFRJ, Brazil) investigated the following question: what is the optimal rate at which users should access a social network? To answer this question, they have proposed an analytical model to determine the value of an access (VoA) to the social network. In the simple setting they considered, VoA is defined as the chance of a user accessing the network and obtaining new content. Clearly, VoA depends on the rate at which sources generate content and on the filtering imposed by the social network. Then, they have posed an optimization problem wherein the utility of users grows with respect to VoA but is penalized by costs incurred to access the network. Using the proposed framework, they provide insights on the optimal access rate. Their results are parameterized using Facebook data, indicating the predictive power of the approach. This research activity led to two publications in 2019 [49], [43].

Last year, the same researchers, together with E. Altman, A. Reiffers-Masson (IISc, India), and the journalist C. Agosti (Univ of Amsterdam, Netherlands), have worked on Facebook News Feed personalization algorithm. The publication [21] complete that line of work described in NEO's 2018 technical report.

7.6.2. Optimal investment strategies for competing camps in a social network

S. Dhamal, W. Ben-Ameur (Telecom SudParis), T. Chahed (Telecom SudParis), and E. Altman have studied the problem of optimally investing in nodes of a social network in [17], wherein two camps attempt to maximize adoption of their respective opinions by the population. Several settings are analyzed, namely, when the influence of a camp on a node is a concave function of its investment on that node, when one of the camps has uncertain information regarding the values of the network parameters, when a camp aims at maximizing competitor's investment required to drive the overall opinion of the population in its favor, and when there exist common coupled constraints concerning the combined investment of the two camps on each node. Extensive simulations are conducted on real-world social networks for all the considered settings.

S. Dhamal, W. Ben-Ameur (Telecom SudParis), T. Chahed (Telecom SudParis), and E. Altman have studied a two-phase investment game for competitive opinion dynamics in social networks, in [18]. The existence of Nash equilibrium and its polynomial time computability is shown under reasonable assumptions. A simulation study is conducted on real-world social networks to quantify the effects of the initial biases and the weigh attributed by nodes to their initial biases, as well as that of a camp deviating from its equilibrium strategy. The study concludes that, if nodes attribute high weight to their initial biases, it is advantageous to have a high investment in the first phase, so as to effectively influence the biases to be harnessed in the second phase.

7.6.3. Extending the linear threshold model

S. Dhamal has proposed a generalization of the linear threshold model to account for multiple product features, in [38]. An integrated framework is presented for product marketing using multiple channels: mass media advertisement, recommendations using social advertisement, and viral marketing using social networks. An approach for allocating budget among these channels is proposed.

7.6.4. Public retention in Youtube

There exist many aspects involved in a video turning viral on YouTube. These include properties of the video such as the attractiveness of its title and thumbnail, the recommendation policy of YouTube, marketing and advertising policies and the influence that the video's creator or owner has in social networks. E. Altman and T. Jimenez (CERI/LIA, Univ Avignon), study in [29] audience retention measurements provided by YouTube to video creators, which may provide valuable information for improving the videos and for better understanding the viewers' potential interests in them. They then study the question of when is a video too long and can gain from being shortened. They examine consistency between several existing audience retention measures. They end in a proposal for a new audience retention measure and identify its advantages.

7.6.5. The medium selection game

F. Lebeau (ENS Lyon), C. Touati (Inria Grenoble-Rhone-Alpes), E. Altman and N. Abuzainab (Virginia Tech, USA) consider in [45] competition of content creators in routing their content through various media. The routing decisions may correspond to the selection of a social network (e.g. twitter versus facebook or linkedin) or of a group within a given social network. The utility for a player to send its content to some medium is given as the difference between the dissemination utility at this medium and some transmission cost. The authors model this game as a congestion game and compute the pure potential of the game. In contrast to the continuous case, they show that there may be various equilibria. They show that the potential is M-concave which allows them to characterize the equilibria and to propose an algorithm for computing it. They then introduce a learning mechanism which allows them to give an efficient algorithm to determine an equilibrium. They finally determine the asymptotic form of the equilibrium and discuss the implications on the social medium selection problem.

7.7. Applications to Environmental Issues

Participant: Alain Jean-Marie.

7.7.1. Sustainable management of water consumption

Continuing a series of game-theoretic studies on sustainable management of water resources, A. Jean-Marie, jointly with T. Jimenez (CERI/LIA, Univ Avignon) and M. Tidball (INRA), consider in [54] the basic groundwater exploitation problem, in the case where agents (farmers) have incomplete information about other agents' profit functions and about pumping cost functions. Farmers behave more or less myopically. The authors analyze two models where they assume that each agent relies on simple beliefs about the other agents' behavior. In a first model, a variation of their own extraction has a first order linear effect on the extractions of others. In a second model, agents consider that extraction of the others players is a proportion of the available water. Farmers' beliefs are updated through observations of the resource level over time. The paper also considers two models with a myopic feature and no learning. In the first one, agents do not know the profit function of the other agents and cost is announced before extraction. In the second one agents know the profit function of the other player and cost is announced after extraction. In this last case agents play a Nash equilibrium. The four behaviors are compared from the economic and environmental points of view.

7.7.2. Pollution permit trading

In a joint work with K. Fredj (Univ of Northern British Columbia, Canada), G. Martín-Herrán (Univ Valladolid, Spain) and Mabel Tidball (INRA), A. Jean-Marie investigated in [20] the strategic behaviour of two countries or firms that minimize costs facing emission standards. Emission standards can be reached through emission reduction, banking or borrowing, and emission trading in a given and fixed planning horizon. The authors extend classical models with: the introduction of transaction costs in tradeable emission markets on the one hand, and using a dynamic game setting, on the other hand. They analyze the case with and without transaction costs and the case with and without discount rate. They characterize socially optimal solutions and Nash equilibria in each case and, depending on the initial allocation, characterize the buyer and seller in the emission trading market. The main findings prove that the agents' equilibrium is not efficient when transaction costs are positive.

RESIST Team

7. New Results

7.1. Monitoring

7.1.1. Encrypted Traffic Analysis

Participants: Jérôme François [contact], Pierre-Olivier Brissaud, Pierre-Marie Junges, Isabelle Chrisment, Thibault Cholez, Olivier François, Olivier Bettan [Thales].

Nowadays, most of Web services are accessed through HTTPS. While preserving user privacy is important, it is also mandatory to monitor and detect specific users' actions, for instance, according to a security policy. Our paper [4] presents a solution to monitor HTTP/2 traffic over TLS. It highly differs from HTTP/1.1 over TLS traffic what makes existing monitoring techniques obsolete. Our solution, H2Classifier, aims at detecting if a user performs an action that has been previously defined over a monitored Web service, but without using any decryption. It is thus only based on passive traffic analysis and relies on random forest classifier. A challenge is to extract representative values of the loaded content associated to a Web page, which is actually customized based on the user action. Extensive evaluations with five top used Web services demonstrate the viability of our technique with an accuracy between 94% and 99%.

We were also interested by Internet of Things (IoT) as related devices become widely used and their control is often provided through a cloud-based web service that interacts with an IoT gateway, in particular for individual users and home automation. Therefore, we propose a technique demonstrating that is possible to infer private user information, i.e., actions performed, by considering a vantage point outside the end-user local IoT network. By learning the relationships between the user actions and the traffic sent by the web service to the gateway, we have been able to establish elementary signatures, one for each possible action, which can be then composed to discover compound actions in encrypted traffic. We evaluated the efficiency of our approach on one IoT gateway interacting with up to 16 IoT devices and showed that a passive attacker can infer user activities with an accuracy above 90%. This work has been published in [16] and is related to the H2020 SecureIoT project (section 9.3.1.2).

7.1.2. Predictive Security Monitoring for Large-Scale Internet-of-Things

Participants: Jérôme François [contact], Rémi Badonnel, Abdelkader Lahmadi, Isabelle Chrisment, Adrien Hemmer.

The Internet-of-Things has become a reality with numerous protocols, platforms and devices being developed and used to support the growing deployment of smart services. Providing new services requires the development of new functionalities, and the elaboration of complex systems that are naturally a source of potential threats. Real cases recently demonstrated that the IoT can be affected by naïve weaknesses. Therefore, security is of paramount importance.

In that context, we have proposed a process mining approach, that is capable to cope with a variety of devices and protocols, for supporting IoT predictive security [14]. We have described the underlying architecture and its components, and have formalized the different phases related to this solution, from the building of behavioral models to the detection of misbehaviors and potential attacks. The pre-processing identifies the states characterizing the IoT-based system, while process mining methods elaborate behavioral models that are compatible with the heterogeneity of protocols and devices [26]. These models are then exploited to analyze monitoring data at runtime and detect misbehaviors and potential attacks preventively. Based on a proof-of-concept prototype, we have quantified the detection performances, as well as the influence of time splitting and clustering techniques. The experimental results clearly show the benefits of our solution combining process mining and clustering techniques. As future work, we are interested in comparing it to other alternative learning techniques, as well as in evaluating to what extent the generated alerts can be exploited to drive the activation of counter-measures.

This work has been achieved in the context of the H2020 SecureIoT project (section 9.3.1.2).

7.1.3. Monitoring of Blockchains' Networking Infrastructure

Participants: Thibault Cholez [contact], Jean-Philippe Eisenbarth, Olivier Perrin.

With the raise of blockchains, their networking infrastructure becomes a critical asset as more and more money and services are made on top of them. However, they are largely undocumented and may be prone to performance issues and severe attacks so that the question of the resiliency of their overlay network arises. With regard to the state of the art on P2P networks security, the fact that a service infrastructure is distributed is not sufficient to assess its reliability, as many bias (for instance, if nodes are concentrated in a given geographical location) and attacks (eclipse, Sybil or partition attacks) are still possible and may severely disturb the network.

Overall, according to the scientific literature, the security provided by the proof of work consensus and the huge size of the main public blockchains seem to protect them well from large scale attacks (51% attack, selfish mining attack, etc.) whose cost to be successful becomes prohibitive and often exceeds the expected gain. However, rather than only focusing on the application level, an attacker could rather try to disturb the underlying P2P network to weaken the consensus in some specific parts of the blockchain network to gain advantage. Our current work uses a third-party crawler to get an accurate view of the Bitcoin overlay network. We are currently analyzing the data with graph theory metrics to identify possible anomalies or flaws that could be exploited by attackers.

7.1.4. Quality of Experience Monitoring

Participants: Isabelle Chrisment [contact], Antoine Chemardin, Frédéric Beck, Lakhdar Meftah [University of Lille], Romain Rouvoy [University of Lille].

We carried on our collaboration with the SPIRALS team (Inria/Université de Lille). Even though mobile crowdsourcing allows industrial and research communities to build realistic datasets, it can also be used to track participants' activity and to collect insightful reports from the environment (e.g., air quality, network quality). While data anonymization for mobile crowdsourcing is commonly achieved *a posteriori* on the server side, we have proposed a decentralized approach, named Fougere [19], which introduces an *a priori* data anonymization process. In order to validate our privacy preserving proposal, two testing frameworks (ANDROFLEET and PEERFLEET [20]) have been designed and implemented. They allows developers to automate reproducible testing of nearby peer-to-peer (P2P) communications.

In the context of both ANR BottleNet (section 9.2.1.1) and IPL BetterNet (section 9.2.5.1) projects, we continued to work on our open measurement platform for the quality of mobile Internet access (i.e., setup and manage the backend infrastructure for data collection and analysis). This platform is hosted by the High Security Laboratory [0] located at Inria Nancy Grand-Est. A collect campaign has been performed with a small set of volunteer users selected by the INSEAD-Sorbonne Université Behavioural Lab [0].

7.2. Experimentation

This section covers our work on experimentation on testbeds (mainly Grid'5000), on emulation (mainly around the Distem emulator), and on Reproducible Research.

7.2.1. Grid'5000 Design and Evolutions

Participants: Benjamin Berard [SED], Luke Bertot, Alexandre Merlin, Lucas Nussbaum [contact], Nicolas Perrin, Patrice Ringot [SISR LORIA], Teddy Valette [SED].

The team was again heavily involved in the evolutions and the governance of the Grid'5000 testbed.

Technical team management. Since the beginning of 2017, Lucas Nussbaum serves as the *directeur technique* (CTO) of Grid'5000 in charge of managing the global technical team (10 FTE). He is also a member of the *Bureau* of the GIS Grid'5000.

[0] https://lhs.loria.fr

[0] https://www.insead.edu/centres/insead-sorbonne-universite-lab-en

SILECS project. We are also heavily involved in the ongoing SILECS project, that aims to create a new infrastructure on top of the foundations of Grid'5000 and FIT in order to meet the experimental research needs of the distributed computing and networking communities.

SLICES ESFRI proposal. At the European level, we are involved in a ESFRI proposal submission. We submitted a *Design Study* project in November 2019, and are in the final stages of submitting the ESFRI proposal itself in early 2020.

TILECS workshop. We participated in the organization of the TILECS workshop. TILECS (*Towards an Infrastructure for Large-Scale Experimental Computer Science*, https://www.silecs.net/tilecs-2019/) gathered about 80 members (mostly faculty) of the testbeds designers and users community in France, to discuss the future plans for research infrastructures in the networking and distributed computing fields. During that workshop, Lucas Nussbaum presented Grid'5000 [32].

Group storage. A technical contribution from the team is the addition of a *group storage* service that allows groups of users to share data, with improved security and performance compared to what was previously available.

Support for Debian 10. Another notable technical contribution from the team is the work of Teddy Valette on supporting Debian 10 in the set of Grid'5000 system environments made available to users.

New clusters available in Nancy: graffiti, gros, grue. Finally, the team was also heavily involved in the purchase and installation of several new clusters in the Nancy site, gathering funding from CPER LCHN, CPER Entreprises, MULTISPEECH team, LARSEN team. This greatly increases the resources available locally, both for GPUs (graffiti and grue), and for large-scale experiments (gros).

7.2.2. Involvement in the Fed4FIRE Testbeds Federation

Participants: Luke Bertot, Lucas Nussbaum [contact].

In the context of the Fed4FIRE+ project (section 9.3.1.1), Grid'5000 was officially added to the Fed4FIRE federation at the beginning of 2019. In 2019, we implemented on-demand *stitching* between Grid'5000 experiments and other testbeds of the federation (through VLANs provided by GEANT and RENATER), allowing experiments that combine resources from Grid'5000 and other testbeds [27]. We are also improving our implementation of an SFA Aggregate Manager in order to allow the use of Grid'5000 through Fed4FIRE tools, such as the jFed GUI.

We also worked on the issue of classifying and presenting the set of testbeds available in the federation. This was the subject of a presentation at the GEFI collaboration workshop [31].

7.2.3. I/O Emulation Support in Distem

Participants: Alexandre Merlin, Abdulqawi Saif, Lucas Nussbaum [contact].

We finished the work on adding I/O emulation support in Distem, in order to experiment how Big Data solution can handle degraded situations [22].

7.2.4. Distributing Connectivity Management in Cloud-Edge infrastructures

Participant: Lucas Nussbaum [contact].

In the context of David Espinel's PhD (CIFRE Orange, co-supervised with Adrien Lebre and Abdelhadi Chari), we worked on distributing connectivity management in Cloud-Edge infrastructures [38]. The classic approach of deploying large data centers to provide Cloud services is being challenged by the emerging needs of Internet of Things applications, Network Function Virtualization services or Mobile edge computing. A massively distributed Cloud-Edge architecture could better fit the requirements and constraints of these new trends by deploying on-demand Infrastructure as a Service in different locations of the Internet backbone (i.e, network point of presences). A key requirement in this context is the establishment of connectivity among several virtual infrastructure managers in charge of operating each site. In this work, we analyzed the requirements and challenges raised by the inter-site connectivity management in a Cloud-Edge infrastructure.

7.2.5. NDN Experimentation

Participants: Thibault Cholez [contact], Xavier Marchal, Olivier Festor.

While ICN is a promising technology, we currently lack experiments carrying real user traffic. This also highlights the difficulty of making the link between the new NDN world and the current IP world. To address this issue, we designed and implemented an HTTP/NDN gateway (composed of ingress and egress gateways) that can seamlessly transport the traffic of regular web users over an NDN island, making them benefit from the good properties of the protocol to deliver content (request mutualization, caching, etc.). The gateway itself is part of a wider architecture that aims to use NFV to deploy NDN and benefit from its orchestration capability to address performance and security issues inherent to new network architectures.

To validate the whole architecture, a testbed involving real users was made. The gateway was used by dozens of users for a few weeks to prove that running a NDN network over NFV is a viable solution to address the transition between both worlds. Users accessed many websites through the NDN network in a very satisfying way. The results have been published in IEEE Communications Magazine [5].

7.3. Analytics

7.3.1. CPS Security Analytics

Participants: Abdelkader Lahmadi [contact], Mingxiao Ma, Isabelle Chrisment.

During 2019, we evaluated a novel type of attack, named Measurement as Reference attack (MaR), on the cooperative control and communication layers in microgrids, where the attacker targets the communication links between distributed generators (DGs) and manipulates the reference voltage data exchanged by their controllers. We assessed its impact on reference voltage synchronization at the different control layers of a microgrid. Results and the development of an experimental platform are presented in [18] to demonstrate this attack, in particular the maximum voltage deviation and inaccurate reference voltage synchronization it causes in a microgrid. ML algorithms are also applied on the collected datasets from this platform for the detection of this attack.

7.3.2. Optimal and Verifiable Packet Filtering in Software-Defined Networks

Participants: Abdelkader Lahmadi [contact], Ahmad Abboud, Michael Rusinowitch [Pesto team], Miguel Couceiro [Orpailleur team], Adel Bouhoula [Numeryx].

Packet filtering is widely used in multiple networking appliances and applications, in particular, to block malicious traffic (protection of network infrastructures through firewalls and intrusion detection systems). It is also widely deployed on routers, switches and load balancers for packet classification. This mechanism relies on the packet's header fields to filter such traffic by using range rules of IP addresses or ports. However, the set of packet filters has to handle a growing number of connected nodes and many of them are compromised and used as sources of attacks. For instance, IP filter sets available in blacklists may reach several millions of entries, and may require large memory space for their storage in filtering appliances. In [40], [39], we proposed a new method based on a double mask IP prefix representation together with a linear transformation algorithm to build a minimized set of range rules. We have formally defined the double mask representation over range rules and proved that the number of required masks for any range is at most $2w-4$, where w is the length of a field. This representation makes the network more secure, reliable and easier to maintain and configure. We show empirically that the proposed method achieves an average compression ratio of 11% on real-life blacklists and up to 74% on synthetic range rule sets. Finally, we add support of double mask into a real SDN network.

7.3.3. Port Scans Analysis

Participants: Jérôme François [contact], Frederic Beck, Sofiane Lagraa [University of Luxembourg], Yutian Chen [Telecom Nancy], Laurent Evrard [University of Namur], Jean-Noël Colin [University of Namur].

TCP/UDP port scanning or sweeping is one of the most common technique used by attackers to discover accessible and potentially vulnerable hosts and applications. Although extracting and distinguishing different port scanning strategies is a challenging task, the identification of dependencies among probed ports is primordial for profiling attacker behaviors, with as a final goal to better mitigate them. In [6], we proposed an approach that allows us to track port scanning behavior patterns among multiple probed ports and identify intrinsic properties of observed group of ports. Our method is fully automated and based on graph modeling and data mining techniques including text mining. It provides to security analysts and operators relevant information about services that are jointly targeted by attackers. This is helpful to assess the strategy of the attacker, such that understanding the types of applications or environment she targets. We applied our method to data collected through a large Internet telescope (or Darknet).

In addition, we decided to leverage this knowledge for improving data analysis techniques applied to network traffic monitoring. Network traffic monitoring is primordial for network operations and management for many purposes such as Quality-of-Service or security. However, one major difficulty when dealing with network traffic data (packets, flows...) is the poor semantic of individual attributes (number of bytes, packets, IP addresses, protocol, TCP/UDP port number...). Many attributes can be represented as numerical values but cannot be mapped to a meaningful metric space. Most notably are application port numbers. They are numerical but comparing them as integers is meaningless. In [13], [12], we propose a fine grained attacker behavior-based network port similarity metric allowing traffic analysis to take into account semantic relations between port numbers. The behavior of attackers is derived from passive observation of a Darknet or telescope, aggregated in a graph model, from which a semantic dissimilarity function is defined. We demonstrated the veracity of this function with real world network data in order to pro-actively block 99% of TCP scans.

7.4. Orchestration

7.4.1. Mutualization of Monitoring Functions in Edge Computing

Participants: Jérôme François [contact], Mohamed Abderrahim [Orange Labs], Meryem Ouzzif [Orange Labs], Karine Guillouard [Orange Labs], Adrien Lebre [STACK Inria team, IMT Atlantique], Charles Prud'Homme [IMT Atlantique], Xavier Lorca [IMT Mines Albi, France].

By relying on small sized and massively distributed infrastructures, the edge computing paradigm aims at supporting the low latency and high bandwidth requirements of the next generation services that will leverage IoT devices (e.g., video cameras, sensors). To favor the advent of this paradigm, management services, similar to the ones that made the success of cloud computing platforms, should be proposed. However, they should be designed in order to cope with the limited capabilities of the resources that are located at the edge. In that sense, they should mitigate as much as possible their footprint. Among the different management services that need to be revisited, we investigated in [10] the monitoring one. Monitoring functions tend to become compute-, storage-and network-intensive, in particular because they will be used by a large part of applications that rely on real-time data. To reduce as much as possible the footprint of the whole monitoring service, we proposed to mutualize identical processing functions among different tenants while ensuring their quality-of-service (QoS) expectations. We formalized our approach as a constraint satisfaction problem and show through micro-benchmarks its relevance to mitigate compute and network footprints.

This work has been achieved in the context of the Inria-Orange joint lab (section 9.2.2.1).

7.4.2. Software-Defined Security for Clouds

Participants: Rémi Badonnel [contact], Olivier Festor, Maxime Compastié.

Cloud infrastructures provide new facilities to build elaborated added-value services by composing and configuring a large variety of computing resources, from virtualized hardware devices to software products. They are however further exposed to security attacks than traditional environments. We have pursued our efforts on a software-defined security strategy based on the TOSCA language, in order to support the protection of cloud resources using unikernel techniques [11]. This language enables the specification of cloud services and their orchestration. We have extended it to drive the integration and configuration of security mechanisms within cloud resources, at the design and operation phases, according to different security levels. We rely on unikernel techniques to elaborate cloud resources using a minimal set of libraries, in order to reduce the attack surface. We have designed a framework to interpret this extended language and to generate and configure protected unikernel virtual machines, in accordance with contextual changes. The adaptation is typically performed through the regeneration of protected unikernel virtual machines in a dynamic manner. We have quantified the benefits and limits of this approach through extensive series of experiments. As future work, we are interested in investigating security issues specifically related to cloud resource migrations, and evaluating to what extent our hardening techniques can be complemented by security chains.

This word has been achieved in the context of the Inria-Orange joint lab (section 9.2.2.1).

7.4.3. Chaining of Security Functions

Participants: Rémi Badonnel [contact], Abdelkader Lahmadi, Stephan Merz, Nicolas Schnepf.

Software-defined networking offers new opportunities for protecting end users and their applications. It enables the elaboration of security chains that combines different security functions, such as firewalls, intrusion detection systems, and services for preventing data leakage. In that context, we have continued our efforts on the orchestration and verification of security chains, in collaboration with Stephan Merz from the VeriDis project-team at Inria Nancy, and concretized with the PhD defense of Nicolas Schnepf in September 2019 [3]. In particular, we have proposed this year an approach for automating the merging of security chains in software-defined networks [24]. This method complements the inference-based generation techniques that we proposed in [9]. The merging algorithms are designed to compose several security chains into a single one, in order to minimize the number of security functions and rules, while preserving the semantics of the initial chains. The algorithms have been implemented in Python and have been integrated into a proof-of-concept prototype that also contains the learning and inference components [23]. The performance of this implementation has been evaluated through extensive experiments. In particular, we have compared different approaches to merging security chains in terms of the complexity of the resulting chains, their accuracy, and the overhead incurred in computing the combined chains. The proposed solution is able to minimize the number of security functions and rules. It also facilitates the building of security chains at runtime, through a decoupling from the generation of individual chains.

7.4.4. Software-Defined Traffic Engineering to Absorb Influx of Network Traffic

Participants: Jérôme François [contact], Abdelkader Lahmadi, Romain Azais [MOSAIC team], Benoit Henry [IMT Lille Douai], Shihabur Chowdhury [University of Waterloo], Raouf Boutaba [University of Waterloo].

Existing shortest path-based routing in wide area networks or equal cost multi-path routing in data center networks do not consider the load on the links while taking routing decisions. As a consequence, an influx of network traffic stemming from events such as distributed link flooding attacks and data shuffle during large scale analytics can congest network links despite the network having sufficient capacity on alternate paths to absorb the traffic. This can have several negative consequences, service unavailability, delayed flow completion, packet losses, among others. In this regard and under the context of NetMSS associate team (section 9.4.1.1), we proposed SPONGE [15], a traffic engineering mechanism for handling sudden influx of network traffic. SPONGE models the network as a stochastic process, takes the switch queue occupancy and traffic rate as inputs, and leverages the multiple available paths in the network to route traffic in a way that minimizes the overall packet loss in the network. We demonstrated the practicality of SPONGE through an OpenFlow based implementation, where we periodically and pro-actively reroute network traffic to the routes computed by SPONGE. Mininet emulations using real network topologies show that SPONGE is capable of reducing packet drops by 20% on average even when the network is highly loaded because of an ongoing link flooding attack.

SOCRATE Project-Team

5. New Results

5.1. Flexible Radio Front-End

Activities in this axis could globally be divided in three main topics: wake-up radio and wireless power transfer, RFID systems and combination of spatial modulation and full-duplex.

5.1.1. Wake-Up radio and wireless power transfer

The ubiquity of wireless sensor networks (WSN), as well as the rapid development of the Internet of Things (IoT), impel new approaches to reduce the energy consumption of the connected devices. The wake-up radio receivers (WuRx) were born in this context to reduce as much as possible the energy consumption of the radio communication part. We aim at proposing a low-cost, high-efficiency rectifier to improve a quasi-passive WuRx performance in terms of communication range. By optimizing the wideband matching circuit and the proposed rectifier's load impedance, the sensitivity was increased by 5 dB, corresponding to an increase of the communication range (13 meters in free space) [10].

We also studied an original solution to maximize the DC power collected in the case of a wireless power transfer (WPT) scenario. Using state-space model representation, the WPT System is considered as a feedback approach in order to maximize the amount of harvested energy. To do this, a global simulation is performed to show the importance of taking into account the propagation channel and the rectifier circuit aspects in the case of optimizing the waveform to increase the harvested energy. By using an optimized multi-sine signal with zero phase as the excitation, taking into account the characteristics of the channel and the physical contributions of the rectifier, we managed to obtain better output DC values compared to a single tone source or a multi-sine signal without optimization, with the same average power input [14].

We plan now to apply this optimized WPT technique to feed Wireless sensors in the particular case of ventilation ducts (HVAC) [24].

5.1.2. RFID

The ARA (Auvergne Rhone Alpes) RAFTING project mainly deals with the design and analysis of wire antennas for RFID tags in the context of wearable electronics. More specifically, an helical dipole antenna dedicated to the smart textile yarn applications has been designed. Moreover, the performance was analysized with respect to mechanical constraints, together with the extraction of accurate electrical models. This work was done in collaboration with Primo 1D company. In perspective, the integration of the NFC protocol together with RFID UHF and the integration of sensing capabilities is envisaged [6], [19], [7], [12], [21].

The Spie ICS- INSA Lyon chair on IoT has granted us for a PhD thesis on Scatter Radio and RFID tag-to-tag communications. Some seminal results have shown that it is actually possible to create a communication between two RFID tags, just using ambient radiowaves or a dedicated distant radio source, without the need of generating a signal from the tag itself. Theoretical and simulated performance have been studied.

5.1.3. Combination of spatial modulation and full-duplex

Spatial modulation (SM) as a new MIMO technique is based on transmitting part of the information by activating different emitting antennas. SM increases spectral efficiency and uses only one radio frequency chain. Moreover, for full-duplex (FD) communication systems, self-interference (SI) is always a central problem. Therefore, combining FD and SM can dramatically reduce the difficulty of SIC (Self-interference Cancellation) because of the single SI chain. A Full Duplex Spatial Modulation (FDSM) system is proposed and an active analog SIC is designed in this work. Moreover, the impact of SIC accuracy on the system performance is studied. The results demonstrate that the accuracy requirement will increase as the INR (Self-interference-to-noise Ratio) increases. The FDSM system is less sensitive than the FD system, which can get a better BER (Bit Error Rate) performance as errors increase. Furthermore, an SI detector is proposed to resolve the influence of the number of detected symbols.

5.2. Software Radio Programming Model

5.2.1. *Transiently powered systems and Non-Volatile Memory*

Socrate is studying the new NVRAM (Non-Volatile Radom Access Memory) technology and its use in ultra-low power context. Non-Volatile memory has been extisting for a while (Nand Flash for instance) but was not sufficiently fast to be used as main memory. Many emerging technologies are forseen for Non-Volatile RAM to replace current RAM [32].

Socrate has started a work on the applicability of NVRAM for *transiently powered systems*, i.e. systems which may undergo power outage at any time. This study resulted in the Sytare software published in IEEE Transaction on Computer [3] and is also studied in an Inria Project Lab ZEP (https://project.inria.fr/iplzep/teams/).

The Sytare software introduces a checkpointing system that takes into account peripherals (ADC, leds, timer, radio communication, etc.) present on all embedded systems. Checkpointing is the natural solution to power outage: regularly save the state of the system in NVRAM so as to restore it when power is on again. However, no work on checkpointing took into account the restoration of the states of peripherals, Sytare provides this possibility.

Another acheivement in this domain is the PhD of Tristan Delizy that concerns memory heterogeneity that results from new NVM technologies. While emerging memory technologies may offer power reduction and high integration density, they come with major drawbacks such as high latency or limited endurance. As a result, system designers tend to juxtapose several memory technologies on the same chip. We aim to provide the embedded application programmer with a transparent software mechanism to leverage this memory heterogeneity. The work of Tristan Delizy studies the interaction between dynamic memory allocation and memory heterogeneity. He provides cycle accurate simulation of embedded platforms with various memory technologies and shows that different dynamic allocation strategies have a major impact on performance. He demonstrates that interesting performance gains can be achieved even for a low fraction of memory using low latency technology, but only with a clever placement strategy between memory banks. This work will soon be proposed to publication.

5.2.2. *Sytare integration in Riot*

The ADT SytaRiot has been granted to provide transient power management in the Riot operating system [27]. This integration was realized by Gero muller, here is a summary of the technical tasks and correponding pull request on Riot GitHub:

5.2.2.1. Port RIOT to MSP430+FRAM micro-controllers

- Bring-up the chip against the newer msp430-elf compiler and integrate the toolchain into the RIOT CI infrastructure, cf https://github.com/RIOT-OS/riotdocker/pull/67 , https://github.com/RIOT-OS/riotdocker/pull/82 , https://github.com/RIOT-OS/riotdocker/pull/91
- Implement initial support for the MSP430FR59xx in RIOT, including device drivers for key on-chip peripherals (UART, Timers, GPIO, etc). cf https://github.com/RIOT-OS/RIOT/pull/11012
- Implement a board support package for the MSP-EXP430FR5969 Launchpad Development Kit and the Boost-IR daughter-board (Infrared transceiver + keypad), cf https://github.com/geromueller/RIOT/commit/f13d33
- Participate in IETF hackathon 104 (Prague, March 23–29, 2019) to work on SUIT IoT Firmware Update, cf https://trac.ietf.org/trac/ietf/meeting/wiki/104hackathon

5.2.2.2. Explicit checkpointing in RIOT

- Implement the required low-level code (e.g. DMA driver) for saving/restoring the state of the application to FRAM. cf https://github.com/geromueller/RIOT/commits/checkpoint
- Implement save/restore methods in all relevant device drivers (DMA, GPIO, UART, Timers) and design an API to expose checkpointing as a general system service in RIOT. cf https://github.com/geromueller/RIOT/commit/8b301e

- Participate in the RIOT Summit (Helsinki, September 5–6, 2019) to give a talk about checkpointing and power measurement. cf https://summit.riot-os.org/2019/blog/speakers/gero-muller/ Power measurement

5.2.3. A high-performance ammeter for embedded systems

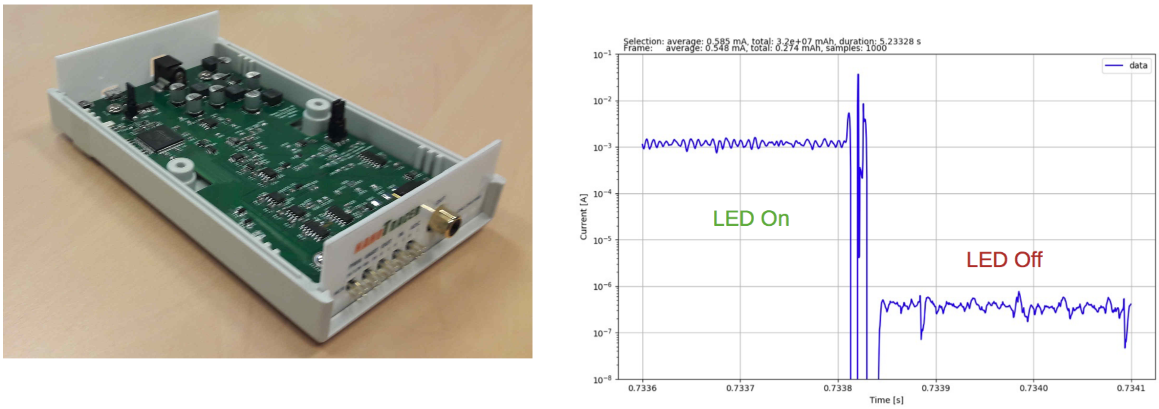

Figure 4. Photo (left) of first packaged nanoTracer prototype and snapshot (right) of a measurement provided by nanoTracer

In embedded low power processing, precise power consumption is a key issue. The Socrate team realized that existing tools could not fullfill the requirements needed for harvesting devices monitoring (measuring from nano-Amperes to milli-Ampere current values at a high sampling rate and continuously).

With the skills of Gero Müller hired on the SytaRiot ADT, the socrate team designed and built a high performance ammeter dedicated to power measurements for small devices. Our prototype measures currents between 100nA and 100mA (gain is audo-adjusted dynamically) with a sampling frequency of 2Msps. Data is streamed to a PC over USB which enables long-running experiments, or just real-time visualization of data (cf screenshot in Fig. 4).

The device, named *nanoTracer*, is referenced in the software section, it is an open project on gitlab (https://gitlab.inria.fr/nanotracer/). A first version is currently tested at Inria (Alexandre Abadie from the IoT SED team) and should soon be available for free for Inria and Academic researcher. We are working on solutions to provide a commercial circuit if requests come from other actors.

5.2.4. Ultra-low latency audio on FPGA

Recently the Socrate team started a collaboration with the researchers of the GRAME group. GRAME is a "Centre National de Création Musicale" (CNCM) organized in three departments: music production, transmission/mediation, and computer music research. Four GRAME researchers have expertise in computer science (compilation), audio DSP, digital lutherie, and human-computer interaction in general. GRAME has been leading the development of the FAUST[0] programming language since its creation in 2004. The GRAME researchers have been associated to CITI as external members in September 2019.

[0] FAUST is a domain specific language for real-time audio signal processing primarily developed at GRAME-CNCM and by a worldwide community. FAUST is based on a compiler "translating" DSP specifications written in FAUST into a wide range of lower-level languages (e.g., C, C++, Rust, Java, WASM, LLVM bitcode, etc.). Thanks to its "architecture" system, generated DSP objects can be embedded into template programs (wrappers) used to turn a FAUST program into a specific ready-to-use object (e.g., standalone, plug-in, smartphone app, webpage, etc.).

Socrate and GRAME have started a collaboration through the Syfala *(synthèse audio faible latence)* project funded by the Fédération Informatique de Lyon. The goal of Syfala is to design an FPGA-based platform for multichannel ultra-low-latency audio Digital Signal Processing (DSP), programmable at high-level with FAUST and using Socrate's software FloPoCo (http://flopoco.gforge.inria.fr). This platform is intended to be usable for various applications ranging from sound synthesis and processing to active sound control and artificial sound field/room acoustics.

Two internships have been working on this project. A first result was a presentation by Florent de Dinechin and Tanguy Risset, introducing the use of HLS and FPGA for audio, at the second *Programmable Audio Workshop* (https://faust.grame.fr/paw/) organized by GRAME.

5.2.5. Evaluation of the posit number system

The posit number system is a very elegant way to represent real numbers in a computers. Its proponents promote it as a better replacement for floating-point arithmetic: posits do indeed improve the application-level accuracy of some applications. However, this also comes with accuracy regressions in other cases. Socrate members, along with members of the AriC project-team, first studied some numerical aspects of posits [18]. Socrate then performed a thorough evaluation of the implementation of the main posit operators, improving the state of the art in hardware posit in the process. Posit operators were then compared to IEEE 754-compliant floating-point operators, and were found to be about twice as slow and twice as expensive [20], [15].

5.2.6. Evaluation of the Unum number system

CEA researcher, in collaboration with Socrate members, designed a complete accelerator for the UNUM number system, including hardware [8] and compiler support [11]. A novelty of this work is the use of a variable-length, self-describing, and memory-oriented floating-point number format [23].

5.2.7. General computer arithmetic

The 10th anniversary of the FloPoCo open-source arithmetic core generator project was the occasion to reflect on the evolutions of the field in a special session about arithmetic generator challenges organized at the ARITH conference [16].

A marked evolution over this period has been the deployment of very good High-Level Synthesis tools, thanks to which hardware is described using a software programming language (usually C++). This comes with many new arithmetic optimization opportunities, some of which have been reviewed in collaboration with Steven Derrien, from Inria Cairn [25]

An issue was the lack in this context of a portable, unified, and hardware-oriented library of arbitrary precision integers. In collaboration with David Thomas from Imperial College, London, we worked on such a library, and demonstrated that it enables a safe description of complex small-grain architectures (such as floating-point or posit operators) with a performance matching traditional hardware description languages [9].

Meanwhile, we keep studying the most basic operators. There has always existed two main methods of implementing mulitplication by a constant in hardware: Table-Based, and Shift-And-Add. This deserved a qualitative and quantitative comparison [17]. This work (with Martin Kumm, from Fulda Technical University, and Silviu Filip, from Inria Cairn) also includes a refined ILP-based algorithm for the problem of multiplying a fixed-point input number by a real constant.

TRIBE Project-Team

6. New Results

6.1. Human Mobility completion of Sparse Call Detail Records

Participants: Guangshuo Chen [Inria], Aline Carneiro Viana, Marco Fiore [CNR], Carlos Sarraute [Gran-Data].

Mobile phone data are a popular source of positioning information in many recent studies that have largely improved our understanding of human mobility. These data consist of time-stamped and geo-referenced communication events recorded by network operators, on a per-subscriber basis. They allow for unprecedented tracking of populations of millions of individuals over long time periods that span months. Nevertheless, due to the uneven processes that govern mobile communications, the sampling of user locations provided by mobile phone data tends to be sparse and irregular in time, leading to substantial gaps in the resulting trajectory information. In this work, we illustrate the severity of the problem through an empirical study of a large-scale Call Detail Records (CDR) dataset. We then propose two novel and effective techniques to reduce temporal sparsity in CDR that outperform existing ones. the fist technique performs completion (1) at nightime by identifying temporal home boundary and (2) at daytime by inferring temporal boundaries of users, i.e., the time span of the cell position associated with each communication activity. The second technique, named Context-enhanced Trajectory Reconstruction, complete individual CDR-based trajectories that hinges on tensor factorization as a core method by leveraging regularity in human movement patterns.

Our approach lets us revisit seminal works in the light of complete mobility data, unveiling potential biases that incomplete trajectories obtained from legacy CDR induce on key results about human mobility laws, trajectory uniqueness, and movement predictability. In addition, the CTR solution infers missing locations with a median displacement within two network cells from the actual position of the user, on a hourly basis and even when as little as 1% of her original mobility is known.

These works have been published at two journals: EPJ Data Science in 2019 and at Computer Communication Elsevier in 2018.

6.2. Adaptive sampling frequency of human mobility

Participants: Panagiota Katsikouli [AGORA], Aline Carneiro Viana, Marco Fiore [CNR], Diego Madariaga.

In recent years, mobile device tracking technologies based on various positioning systems have made location data collection a ubiquitous practice. Applications running on smartphones record location samples at different frequencies for varied purposes.The frequency at which location samples are recorded is usually pre-defined and fixed but can differ across applications; this naturally results in big location datasets of various resolutions. What is more, continuous recording of locations results usually in redundant information, as humans tend to spend significant amount of their time either static or in routine trips, and drains the battery of the recording device.

In this work, we aim at answering the question *"at what frequency should one sample individual human movements so that they can be reconstructed from the collected samples with minimum loss of information?"*. Our first analyses on fine-grained GPS trajectories from users around the world unveil *(i)* seemingly universal spectral properties of human mobility, and *(ii)* a linear scaling law of the localization error with respect to the sampling interval. Such results were published at a paper at IEEE Globecom 2017.

Building on these results, we challenge the idea of a fixed sampling frequency and present a lightweight mobility aware adaptive location sampling mechanism. This is an on-going work with Panagiota Katsikouli, who spent 5 months in our team working as an internship in 2017, and Diego Madariaga who spent 3 months in 2018 in our team working as an internship and has started a PhD in co-tutelle with Aline C. Viana and Javier Bustos (NIC/Univ. of Chile).

Our mechanism can serve as a standalone application for *adaptive location sampling*, or as complimentary tool alongside auxiliary sensors (such as accelerometer and gyroscope). In this work, we implemented our mechanism as an application for mobile devices and tested it on mobile users worldwide. Our experiments show that our method adjusts the sampling frequency to the mobility habits of the tracked users, it reliably tracks a mobile user incurring acceptable approximation errors and significantly reduces the energy consumption of the mobile device.

A journal paper is being prepared for submission.

6.3. Inference of human personality from mobile phones datasets

Participants: Adriano Di Luzio [Sapienza U. di Rome], Aline Carneiro Viana, Julinda Stefa [Sapienza U. di Rome], Katia Jaffres-Runser [U. of Toulouse], Alessandro Mei [Sapienza U. di Rome].

Related to human behavioral studies, personality prediction research has enjoyed a strong resurgence over the past decade. Due to the recognition that personality is predictive of a wide range of behavioral and social outcomes, the human migration to the digital environment renders also possible to base prediction of individual personality traits on digital records (i.e., datasets) mirroring human behaviors. In psychology, one of the most commonly used personality model is the Big5, based on five crucial traits and commonly abbreviated as OCEAN: Openness (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N). They are relatively stable over time, differ across individuals, and, most importantly, guide our emotions and our reactions to life circumstances. It is so for social and work situations, and even for things as simple as the way we use our smartphone. For instance, a person that is curious and open to new experiences will tend to look continuously for new places to visit and thrills to experience.

This work brings the deepest investigation in the literature on the prediction of human personality (*i.e.,* captured by the Big5 traits) from smartphone data describing daily routines and habits of individuals. This work shows that human personality can be accurately predicted by looking at the data generated by our smartphones. GPS location, calls, battery usage and charging, networking context like bluetooth devices and WiFi access points in proximity, and more give enough information about individual habits, reactions, and idiosyncrasies to make it possible to infer the psychological traits of the user. We demonstrate this by using machine learning techniques on a dataset of 55 volunteers who took a psychological test and allowed continuous collection of data from their smartphones for a time span of up to three years. Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (the so called Big5 personality traits) can be predicted with good accuracy even by using just a handful of features. The possible applications of our findings go from network optimization, to personal advertising, and to the detection of mental instability and social hardship in cities and neighborhoods. We also discuss the ethical concerns of our work, its privacy implications, and ways to tradeoff privacy and benefits.

A paper describing this work is under submission at ACM Transactions on Data Science (TDS) , but a technical report is also registered under the name hal-01954733.

6.4. Data offloading decision via mobile crowdsensing

Participants: Emanuel Lima [U. of Porto], Aline Carneiro Viana, Ana Aguiar [U. of Porto], Paulo Carvalho [Univ. Do Minho].

According to Cisco forecasts, mobile data traffic will grow at a compound annual growth rate of 47 % from 2016 to 2021 with smartphones surpassing four-fifths of mobile data traffic. It is known that mobile network operators are struggling to keep up with such traffic demand, and part of the solution is to offload communications to WiFi networks. Mobile data offloading systems can assist mobile devices in the decision making of when and what to offload to WiFi networks. However, due to the limited coverage of a WiFi AP, the expected offloading performance of such a system is linked with the users mobility. Unveiling and understanding human mobility patterns is a crucial issue in supporting decisions and prediction activities for mobile data offloading.

Several studies on the analysis of human mobility patterns have been carried out focusing on the identification and characterization of important locations in users' life in general. We extended these works by studying human mobility from the perspective of mobile data offloading. In our study, offloading zones are identified and characterized from individual GPS trajectories when small offloading time windows are considered. The characterization is performed in terms availability, sojourn, transition time; type and spatial characteristics. We then evaluate the offloading opportunities provided to users while they are travelling in terms of availability, time window to offload and offloading delay. We also study the mobility predictability in an offloading scenario through the theoretical and practical evaluation of several mobility predictors. The results show that (i) attending to users mobility, ten seconds is the minimum offloading time window that can be considered; (ii) offloading predictive methods can have variable performance according to the period of the day; and (iii) per-user opportunistic decision models can determine offloading system design and performance.

This work was published at ACM CHANTS 2018 and its extension will be submitted to WoWMON 2020. This is an on-going work with the the PhD Emanuel Lima (one of my co-supervision), who spent 4 months as an intern in our team in 2018, and his advisors.

6.5. Identifying how places impact each other by means of user mobility

Participants: Lucas Santos de Oliveira [EMBRACE], Pedro Olmo Stancioli [Federal U. of Minas Gerais], Aline Carneiro Viana.

The way in which city neighborhoods become popular and how people trajectory impacts the number of visitation is a fundamental area of study in traditional urban studies literature. Many works address this problem by means of user mobility prediction and POI recommendation. In a different approach, other works address the human mobility in terms of social influence which refers to the case when individuals change their behaviors persuaded by others. Nevertheless, fewer works measure influence of POI based on human mobility data.

Different from previous literature, in this work, we are interested in understanding how the neighborhood POI affect each other by means of human mobility using location-based social networks (LBSNs) data source. Key location identification in cities is a central in human mobility investigation as well as for societal problem comprehension. In this context, we propose a methodology to quantify the power of point-of-interests (POIs) in their vicinity, in terms of impact and independence – the first work in the literature (to the best of our knowledge). Different from literature, we consider the flow of people in our analysis, instead of the number of neighbor POIs or their structural locations in the city. Thus, we first modeled POI's visits using the multiflow graph model where each POI is a node and the transitions of users among POIs are a weighted direct edge. Using this multiflow graph model, we compute the attract, support and independence powers. The attract power and support power measure how many visits a POI gather from and disseminate over its neighborhood, respectively. Moreover, the independence power captures the capacity of POI to receive visitors independently from other POIs. Using a dataset describing the mobility of individuals in the Dartmouth College campus, we identify a slight dependence among buildings as well as the tendency of people to be mostly stationary in few buildings with short transit periods among them.

This work was published in ACM MobiWac 2019 [14] and an extended version is being prepared. Lucas is doing an internship in our team from Nov. 2019 to Jan. 2020.

6.6. Infering friends in the crowd in Device-to-Device communication

Participants: Rafael Lima Da Costa [CAPES], Aline Carneiro Viana, Leobino Sampaio [Federal U. of Bahia], Artur Ziviani [LNCC].

The next generation of mobile phone networks (5G) will have to deal with spectrum bottleneck and other major challenges to serve more users with high-demanding requirements. Among those are higher scalability and data rates, lower latencies and energy consumption plus reliable ubiquitous connectivity. Thus, there is a need for a better spectrum reuse and data offloading in cellular networks while meeting user expectations. According to literature, one of the 10 key enabling technologies for 5G is device-to-device (D2D) communications, an approach based on direct user involvement. Nowadays, mobile devices are attached to human daily life activities, and therefore communication architectures using context and human behavior information are promising for the future. User-centric communication arose as an alternative to increase capillarity and to offload data traffic in cellular networks through opportunistic connections among users. Although having the user as main concern, solutions in the user-centric communication/networking area still do not see the user as an individual, but as a network active element. Hence, these solutions tend to only consider user features that can be measured from the network point of view, ignoring the ones that are intrinsic from human activity (e.g., daily routines, personality traits, etc).

In this work, we first introduce the Tactful Networking paradigm, whose goal is to add perceptive senses to the network, by assigning it with human-like capabilities of observation, interpretation, and reaction to daily-life features and involved entities. To achieve this, knowledge extracted from human inherent behavior (routines, personality, interactions, preferences, among others) is leveraged, empowering user-needs learning and prediction to improve QoE while respecting privacy. We survey the area, propose a framework for enhancing human raw data to assist networking solutions and discuss the tactful networking impact through representative examples. Finally, we outline challenges and opportunities for future research. This tutorial paper is under submittion to ACM Computing and Surveys and a technical report is registered as hal-01675445.

Besides, we investigate how human-aspects and behavior can be useful to leverage future device-to-device communication. We have designed a strategy to select next-hops in a D2D communication that will be human-aware: i.e., that will consider not only available physical resources at the mobile device of a wireless neighbor, her mobility features and restrictions but also any information allowing to infer how much sharing willing she is. Such forwarders nodes will be then used at the offloading of content data through Device-to-Device (D2D) communication, from devices to the closest Mobile Edge Computing infrastructure, transforming mobile phone neighbors in service providers. The selection of next hops based on mobility behavior, resource capability as well as collaboration constitute the novelty we plan to exploit. A conference paper is under preparation and a Brazilian paper under submission to SBRC 2020.

6.7. Deciphering Predictability Limits in Human Mobility

Participants: Douglas Do Couto Teixiera, Aline Carneiro Viana, Jussara Almeida [Federal U. of Minas Gerais], Mario S. Alvim [Federal U. of Minas Gerais].

Human mobility has been studied from different perspectives. One approach addresses predictability, deriving theoretical limits on the accuracy that any prediction model can achieve in a given dataset. Measuring the predictability of any phenomenon is a very useful, but hard task, and especially so in the case of human behavior. Such complexity is due to the uncertain and heterogeneous behavior of humans, as well as to the variability of parameters influencing such behavior. Predictability is concerned with the maximum theoretical accuracy that an ideal prediction model could achieve in a scenario expressed by a given dataset. As such, unlike particular comparisons of alternative prediction models on different datasets, it does not depend on a specific prediction strategy but rather on human behavior, as captured by the available data. Besides, it does not rely on the tuning of a multitude of sensible parameters, providing instead a parameter-free view of how predictable human mobility can be (as expressed in the data).

This approach focuses on the inherent nature and fundamental patterns of human behavior captured in the dataset, filtering out factors that depend on the specificities of the prediction method adopted. In this work, we revisit the state-of-the-art method for estimating the predictability of a person's mobility, which, despite being widely adopted, suffers from low interpretability and disregards external factors that have been suggested to improve predictability estimation, notably the use of contextual information (e.g., weather, day of the week, and time of the day). We propose a new measure, *regularity*, which together with *stationarity*, helps us understand what makes a person's mobility trajectory more or less predictable, as captured by Song et al.'s technique. We show that these two simple measures are complementary and jointly are able to explain most of the variation in Song et al.'s predictability. As such, we here use them as proxies of that technique to analyze how one's mobility predictability varies.

Additionally, we investigate strategies to incorporate different types of contextual information into predictability estimates. In particular, we were the first to quantify the impact of different types of contextual information on predictability in human mobility, for different prediction tasks and datasets. Our results show that, for the next place prediction problem, the use of contextual information plays a larger role than one's history of visited locations in estimating their predictability. Finally, we propose and evaluate alternative estimates of predictability which, while being much easier to interpret, provide comparable results to the state-of-the-art. We show that these estimators, while being more interpretable, provide comparable results in terms of predictability.

This paper was published at ACM SIGSPATIAL 2019, a A+-ranked conference in our domain, and was indicated as a top-six best paper candidate. An extended version is being prepared for submission to a journal.

6.8. Identifying and profiling novelty-seeking behavior in human mobility

Participants: Licia Amichi, Aline Carneiro Viana, Mark Corvella [Boston Univ.], Antonio F. Loureiro [Federal U. of Minas Gerais].

The prediction of individuals' dynamics has attracted significant community attention and has implication for many fields: e.g. epidemic spreading, urban planning, recommendation systems. Current prediction models, however, are unable to capture uncertainties in the mobility behavior of individuals, and consequently, suffer from *the inability to predict visits to new places*. This is due to the fact that current models are oblivious to the exploration aspect of human behavior.

Many prediction models have been proposed to forecast individuals trajectories. However, they all show limited bounded predictive performance. Regardless of the applied methods (e.g., Markov chains, Naive Bayes, neural networks), the type of prediction (i.e., next-cell or next place) or the used data sets (e.g., GPS, CDR, surveys), accuracy of prediction never reaches the coveted 100%. The reasons for such limitations in the accuracy are manyfold: the lack of ground truth data, human beings' complex nature and behavior, as well the exploration phenomenon (i.e., visits to never seen before places). In this work, we focus on the exploration problem, which has rarely been tackled in the literature but indeed, represents a real issue. By construction, most prediction models attempt to forecast future locations from the set of known places, which hinders predicting new unseen places and by consequence, reduces the predictive performance.

Thus, when considering the exploration problem, previous studies either did not provide any consideration of the exploration factors of individuals, or divided the population based on properties that are not always consistent, or assumed that all individuals have the same propensity to explore. Our main goal in this work is to understand the exploration phenomenon and answer the following question: *What type of visits characterize the mobility of individuals?* Using newly designed metrics capturing spatiotemporal properties of human mobility – i.e., known/new and recurrent/intermittent visits – our strategy identifies three groups of individuals according to their degree of exploration: scouters, routineers, and regulars. In the future, we plan to deeply investigate the mobility behavior of individuals in each profile and to assign to each individual an *exploration factor* describing her susceptibility to explore.

This work was published at the Student workshop of ACM CONEXT 2019 [9]. An extended version is being prepared for submission to an int. conference.

6.9. How Geo-indistinguishability Affects Utility in Mobility-based Geographic Datasets

Participants: Adriano Di Luzio [Inria], Aline Carneiro Viana, Catuscia Palamidessi [Comete – Inria], Konstantinos Chatzikokolakis [Comete – Inria], Georgi Dikov [Comete – Inria], Julinda Stefa [Sapienza University].

Many of the scientific challenges that we face today deal with improving the quality of our everyday lives. They aim at making the cities around us smarter, more efficient, and more sustainable (e.g., how to schedule public transport during peak hours or what is the most efficient path for waste disposal). All these challenges share a common ground. They rely on datasets gathered from the real world that depict the mobility of hundreds of thousands individuals and picture, with great detail, the whereabouts of their lives—where they live, work, shop for groceries, and hangout with friends. At the same time, however, the collection of personal data also endangers the privacy of the users that to whom these data belong. To protect the privacy of the users, it is necessary to sanitize these datasets before releasing them to the public.

When we sanitize the datasets we trade the accuracy of the information they contain to protect the privacy of their users. The task of this work is to shed light on the effects of the trade-off between privacy and utility in mobility-based geographic datasets. We aim at finding out whether it is possible to protect the privacy of the users in a dataset while, at the same time, maintaining intact the utility of the information that it contains. In particular, we focus on geo-indistinguishability as a privacy-preserving sanitization methodology, and we evaluate its effects on the utility of the Geolife dataset. We test the sanitized dataset in two real world scenarios: (1) Deploying an infrastructure of WiFi hotspots to offload the mobile traffic of users living, working, or commuting in a wide geographic area; (2) Simulating the spreading of a gossip-based epidemic as the outcome of a device-to-device communication protocol. We show the extent to which the current geo-indistinguishability techniques trade privacy for utility in real world applications and we focus on their effects at the levels of the population as a whole and of single individuals.

This paper was published at the LocalRec 2019 workshop, jointly with ACM SIGSPATIAL [12].

6.10. General-purpose Low-power Secure Firmware Updates for Constrained IoT Devices

Participants: Koen Zandberg [Inria / Freie Universität Berlin], Kaspar Schleiser [Inria / Freie Universität Berlin], Francisco Acosta [Inria], Hannes Tschofenig [Arm Ltd., Cambridge, U.K], Emmanuel Baccelli.

While the IoT deployments multiply in a wide variety of verticals, the most IoT devices lack a built-in secure firmware update mechanism. Without such a mechanism, however, critical security vulnerabilities cannot be fixed, and the IoT devices can become a permanent liability, as demonstrated by recent large-scale attacks. In this paper, we survey open standards and open source libraries that provide useful building blocks for secure firmware updates for the constrained IoT devices–by which we mean low-power, microcontroller-based devices such as networked sensors/actuators with a small amount of memory,among other constraints. We design and implement a prototype that leverages these building blocks and assess the security properties of this prototype. We present experimental results including first experiments with SUIT, a new IETF standard for secure IoT firmware updates. We evaluate the performance of our implementation on a variety of commercial off-the-shelf constrained IoT devices. We conclude that it is possible to create a secure, standards-compliant firmware update solution that uses the state-of-the-art security for the IoT devices with less than 32 kB of RAM and 128 kB of flash memory. Moreover, our prototype is general-purpose, in that it works out-of-the-box or with minimal adaptation on 80% of the hardware supported by RIOT (i.e. approximately 100 different types of IoT devices). As such, this work paves the way towards generic and secure low-power IoT firmware updates.

This paper was published in the IEEE journal IEEE Access [8].

6.11. LoRa-MAB: A Flexible Simulator for Decentralized Learning Resource Allocation in IoT Networks

Participants: Duc-Tuyen Ta [LRI and Inria], Kinda Khawam [UVSQ], Samer Lahoud [ESIB], Cédric Adjih, Steven Martin [LRI, Université Paris-Saclay].

LoRaWAN is a media access control (MAC) protocol for wide area networks. It is designed to allow low-powered devices to communicate with Internet-connected applications over long-range wireless connections. The targeted dense deployment will inevitably cause a shortage of radio resources. Hence, autonomous and lightweight radio resource management is crucial to offer ultra-long battery lifetime for LoRa devices. One of the most promising solutions to such a challenge is the use of artificial intelligence. This will enable LoRa devices to use innovative and inherently distributed learning techniques, thus freeing them from draining their limited energy by constantly communicating with a centralized controller.Before proceeding with the deployment of self-managing solutions on top of a LoRaWAN application, it is sensible to conduct simulation-based studies to optimize the design of learning-based algorithms as well as the application under consideration. Unfortunately, a network simulator for such a context is not fully considered or lacks real deployment parameters. In order to address this shortcoming, we have developed an event-based simulator for resource allocation in LoRaWAN. To demonstrate the usefulness of our simulator, extensive simulations were run in a realistic environment taking into account physical phenomenon in LoRaWAN such as the capture effect and inter-spreading factor interference. The simulation results show that the proposed simulator provides a flexible and efficient environment to evaluate various network design parameters and self-management solutions as well as verify the effectiveness of distributed reinforcement-based learning algorithms for resource allocation problems in LoRaWAN.

This paper was published at the conference WCNC 2019 [15].

6.12. A Survey of Recent Extended Variants of the Traveling Salesman and Vehicle Routing Problems for Unmanned Aerial Vehicles

Participants: Ines Khoufi [Telecom SudParis], Anis Laouiti [Telecom SudParis], Cédric Adjih.

The use of Unmanned Aerial Vehicles (UAVs) is rapidly growing in popularity. Initially introduced for military purposes, over the past few years, UAVs and related technologies have successfully transitioned to a whole new range of civilian applications such as delivery, logistics, surveillance, entertainment, and so forth. They have opened new possibilities such as allowing operation in otherwise difficult or hazardous areas, for instance. For all applications, one foremost concern is the selection of the paths and trajectories of UAVs, and at the same time, UAVs control comes with many challenges, as they have limited energy, limited load capacity and are vulnerable to difficult weather conditions. Generally, efficiently operating a drone can be mathematically formalized as a path optimization problem under some constraints. This shares some commonalities with similar problems that have been extensively studied in the context of urban vehicles and it is only natural that the recent literature has extended the latter to fit aerial vehicle constraints. The knowledge of such problems, their formulation, the resolution methods proposed—through the variants induced specifically by UAVs features—are of interest for practitioners for any UAV application. Hence, in this study, we propose a review of existing literature devoted to such UAV path optimization problems, focusing specifically on the sub-class of problems that consider the mobility on a macroscopic scale. These are related to the two existing general classic ones—the Traveling Salesman Problem and the Vehicle Routing Problem. We analyze the recent literature that adapted the problems to the UAV context, provide an extensive classification and taxonomy of their problems and their formulation and also give a synthetic overview of the resolution techniques, performance metrics and obtained numerical results.

This paper was published in the journal "Drones" 2019, 3(3), 66 [5].

6.13. LoRa-MAB: Toward an Intelligent Resources Allocation Approach for LoRaWAN Networks

Participants: Duc-Tuyen Ta [LRI and Inria], Kinda Khawam [UVSQ], Samer Lahoud [ESIB], Cédric Adjih, Steven Martin [LRI, Université Paris-Saclay].

For a seamless deployment of the Internet of Things (IoT), self-managing solutions are needed to overcome the challenges of IoT, including massively dense networks and careful management of constrained resources in terms of calculation, memory, and battery. Leveraging on artificial intelligence will enable IoT devices to operate autonomously by using inherently distributed learning techniques. Fully distributed resource management will free devices from draining their limited energy by constantly communicating with a centralized controller. The present work is devoted to a specific IoT context, that of LoRaWAN, where devices communicate with the access network via ALOHA-type access and spread spectrum technology. Concurrent transmissions on different spreading factors increase the network capacity. However, the bottleneck is inevitable with the expected massive deployment of LoRa devices. To address this issue, we resort to the popular EXP3 (Exponential Weights for Exploration and Exploitation) algorithm to steer autonomously the decision of LoRa devices towards the least solicited spreading factors. Furthermore, the spreading factor selection is cast as a proportional fair optimization problem used as a benchmark for the learning-based algorithm. Extensive simulations were run in a realistic environment taking into account physical phenomena in LoRaWAN such as the capture effect and inter-spreading factor collision, as well as non-uniform device distribution. In such a realistic setting, we evaluate the performances of the EXP3.S algorithm, an efficient variant of the EXP3 algorithm, and show its relevance against the fair centralized solution and basic heuristics.

This paper was published at the conference GLOBECOM 2019 [16].

6.14. An IoT-Blockchain Architecture Based on Hyperledger Framework for Healthcare Monitoring Application

Participants: Oumaima Attia, Ines Khoufi [Telecom SudParis], Anis Laouiti [Telecom SudParis], Cédric Adjih.

Blockchains are one of the most promising technologies in the domain of the Internet of Things (IoT). At the same time, healthcare monitoring is one of IoT applications where many devices are connected, and collect data that need to be stored in a highly secure way. In this context, we focus on IoT Blockchain architectures for healthcare monitoring applications. We start our study by exploring both IoT and blockchain technologies and identify how Fabric Hyperledger is a blockchain framework that fits our application needs. In this paper, we propose a security architecture based on this framework. We validate our approach first at a design level through concrete examples, then by showing some implemented functionalities.

This paper was published at the conference NTMS 2019 [10].