RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2018

Section New Results

Edition: 2019-03-07

Coast Project-Team

6. New Results

6.1. Design and Analysis of Collaborative Editing Approaches

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

Since the Web 2.0 era, the Internet is a huge content editing place on which users collaborate. Thousand of people can edit this shared document. However, current consistency maintenance algorithms are not adapted to massive collaborative updating involving large number of contributors and a high velocity of changes. This year we studied collaborative editing user behaviour and started to work on an optimised solution for sequence CRDTs. Version control systems such as Git became very widespread in the open-source community. In these collaborative systems, conflict resolution that arise during synchronisation of parallel changes might become a burden for the user. We analysed concurrency and conflicts in Git repository of four projects: Rails, IkiWiki, Samba and Linux Kernel. We analysed the collaboration process of these projects at specific periods revealing how change integration and conflict rates vary during the project development life-cycle. Our study suggests that developers should use more intensively awareness mechanisms close to release dates where changes integration rate is higher. We also discussed the mechanism adopted by Git to consider concurrent changes made on two adjacent lines as conflicting. Based on the high rate of false positives of this mechanism, our study suggests that Git should reconsider signalling adjacent line conflicts inside the source code files [4]. Sequence Conflict-free Replicated Data Types (CRDTs) allow one to replicate and edit, without any kind of coordination, sequences in distributed systems. To ensure convergence, existing works from the literature add metadata to each element but they do not bind its footprint, which impedes their adoption. Several approaches were proposed to address this issue but they do not fit a fully distributed setting. We started to work on the design and validation of a fully distributed renaming mechanism, setting a bound to the metadata's footprint [14]. Addressing this issue opens new perspectives of adoption of these CRDTs in distributed applications.

6.2. Trustworthy Collaboration

Participants: Claudia-Lavinia Ignat, Victorien Elvinger, François Charoy, Olivier Perrin, Gérald Oster, Hoang Long Nguyen.

Trust between users is an important factor for the success of a collaboration. Users might want to collaborate only with those users they trust. We are interested in assessing users trust according to their behaviour during collaboration in a large scale environment. We studied the trust assessment problem and designed a computational trust model for collaborative systems [1]. We also studied how to predict the trust relation between users that did not interact in the past. Given a network in which the links represent the trust/distrust relations between users, we aimed to predict future relations. We proposed a link-sign prediction algorithm [6] that does not require full graph information, is suitable for dynamic networks and takes into account the creation time of the links in the network. Our solution combines state-of-the-art techniques in natural language processing (Doc2Vec [25]) and deep learning (Recurrent Neural Networks [31] with Long-Short Term Memory [24]) with the random walk graph sampling [26]. Our algorithm outperforms state-of-the-art approaches on real world signed directed social network datasets. In distributed collaborative systems, participants maintain a replicated copy of shared documents. They edit their own copy and then share their modifications without any coordination. Copies follow successions of divergence and convergence. Convergence is a liveness property of collaborative systems. Some malicious participants may find an advantage to make the collaboration fail. To that end, they can preclude convergence of the copies. To protect convergence of copies, participants can exploit an authenticated log of modifications. New participants have to retrieve the entire log in order to contribute. Unfortunately, the cost of joining a collaboration increases with the size of this log. Causal Stability allows to prune authenticated logs in a static collaborative group without

any malicious participants. We tailored Causal Stability to dynamic groups in the presence of malicious participants. We also proposed a mechanism to verify the consistency of a pruned log and a mechanism to authenticate a snapshot from a pruned log [7]. Public key server is a simple yet effective way of key management in secure end-to-end communication. To ensure the trustworthiness of a public key server, CONIKS [27] employs a tamper-evident data structure on the server and a gossiping protocol among clients in order to detect compromised servers. However, due to lack of incentive and vulnerability to malicious clients, a gossiping protocol is hard to implement in practice. Meanwhile, alternative solutions such as EthIKS [21] are too costly. We proposed Trusternity [13], [12], an auditing scheme relying on Ethereum blockchain that is easy to implement, inexpensive to operate and resilient to malicious clients. We also conducted an empirical study of system behaviour in face of attacks and proposed a lightweight anomaly detection algorithm to protect clients against such attacks.

6.3. Trust and data sharing in crisis management

Participants: François Charoy, Béatrice Linot, Valerie Shalin.

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 organisations 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 [17], [10] and conducted extensive analysis of the use of existing systems (e.g. the French platform CRISORSEC) through interviews, observation and data analysis. [11]

6.4. Cloud Provisioning for Elastic BPM

Participants: François Charoy, Samir Youcef, Guillaume Rosinosky.

Cloud computing providers do not help consumers to use optimally the available resources. Several approaches have been proposed [33] that take benefit from the elasticity of the Cloud, starting and stopping virtual machines on demand. They suffer from several shortcomings. Often they consider only one objective, the reduction of the cost, or a level of quality of service. We proposed to optimise two conflicting objectives, the number of migrations of tenants that is helpful to reach the optimal cost and the cost incurred considering a set of resources. Our approach allows to take into account the multi-tenancy property and the Cloud computing elasticity, and is efficient as shown by an extensive experimentation based on real data from Bonita BPM customers. In the continuation of our previous work we proposed and validated a more efficient algorithm for elastic execution of processes in the cloud [16]. To ensure a realistic validation, we collaborated with colleagues from the University of Lugano to set up a benchmarking platform in order to evaluate the impact of migration in a multi-tenant setting. This allowed us to execute reproducible experiments and to validate our hypothesis regarding the effect of migration and the parameters that affect them [15]. This platform is now an asset that can be used for all kinds of live migration experiments of software architectures.

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

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

The lack of trust in cloud organisations 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 categorises 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 formalises the relations and the responsibilities between the different actors of the cloud. This allows to identify the different information required to assess and quantify security risks in cloud environments. The framework is a comprehensive approach that decomposes a business process into fragments that can automatically be deployed on multiple clouds. The framework also integrates a selection algorithm that combines the security information of cloud offers and of the process with other quality of service criteria to generate an optimised configuration. It is implemented in a tool to assess cloud providers and decompose processes. Rooted in past years work, we are contributing this year at the methodological and framework levels in two directions:

- At the methodological level, while our risk computing model rested previously only on data provided by cloud providers (provider-side risk model), we are developing a risk model integrating client-side knowledge (client-side risk model).
- Additionally are developing a simulation tool for supporting designer decision with the ability to balance risk with cost when selecting the best cloud configuration [2].

6.6. Scheduling and Resource Allocation in Business Processes

Participants: Khalid Benali, Abir Ismaili-Alaoui.

Business Process Management (BPM) is concerned with continuously enhancing business processes by adapting a systematic approach that enables companies to increase the performance of their existing business processes and achieve their business goals. Business processes are generally considered as blind, stateless and reactive. This means that in each business process execution we do not take into consideration either the results from last process instances nor the context (for most cases). The rise of new technologies such as big and fast data, cloud computing, Internet of Things (IoT), etc, implies new business process scheduling problems. They are linked to limited resources (human and/or machine) or the need to use resources in an optimal and exible way. In order to avoid either under-provisioning (when there is an underestimation for the needed resources, business processes may not be executed) or over-provisioning (the resources planned in advance to cover peak times demands were not used in non-peak time) and also to take into consideration the priority level of each business process instances.

CTRL-A Project-Team

6. New Results

6.1. Programming support for Autonomic Computing

6.1.1. Reactive languages

Participants: Gwenaël Delaval, Eric Rutten.

Our work in reactive programming for autonomic computing systems is focused on the specification and compilation of declarative control objectives, under the form of contracts, enforced upon classical mode automata as defined in synchronous languages. The compilation involves a phase of Discrete Controller Synthesis, 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 [6].

Recent work concerns exploring new possibilities offered by logico-numeric control. We consider Symbolic Limited Lookahead Control for Best-effort Dynamic Computing Resource Management. We put forward a new modeling technique for Dynamic Resource Management (DRM) based on discrete events control for symbolic logico-numerical systems, especially Discrete Controller Synthesis (DCS). The resulting models involve state and input variables defined on an infinite domain (Integers), thereby no exact DCS algorithm exists for safety control. We thus formally define the notion of limited lookahead, and associated best-effort control objectives targeting safety and optimization on a sliding window for a number of steps ahead. We give symbolic algorithms, illustrate our approach on an example model for DRM, and report on performance results based on an implementation in the tool ReaX. This work is in cooperation with the Sumo team at Inria Rennes (Hervé Marchand) and University of Liverpool (Nicolas Berthier), and is published in the WODES 2018 conference [14].

We also have ongoing activities 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) and on compilation and diagnosis for discrete controller synthesis (which is made special by the declarative nature of the compilation, where it is not easy to precisely diagnose cases where no solution can be found).

On the applicative side, we also consider such modular and logico-numeric approaches for the control of different targets in self-adaptive and reconfigurable systems (see below in Section 6.2.2.1 and 6.2.1.3 [20], [15].

6.1.2. 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 6.1.1 [5].

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 are working on a proposal for a DSL called Ctrl-DPR, allowing designers to easily generate Autonomic Managers for DPR FPGA systems (see Section 6.2.1.3). Users can describe their system and their management strategies, in terms of the entities composing the system : tasks, versions, applications, ressources, policies. The DSL relies on a behavioral modeling of these entities, targeted at the design of autonomic managers to control the reconfigurations in such a way as to enforce given policies and strategies. These model-based control techniques are embedded in a compiler, connected to the reactive language and discrete controller synthesis tool of Section 6.1.1 , which enables to generate a C implementation of the controller enforcing the management strategies. We apply our DSL for the management of a video application on a UAV. This work is in cooperation with LabSticc in Lorient (Jean-Philippe Diguet), and is published in the ICAC 2018 conference [16].

Ongoing work involves a generalization from our experiences in software components, DPR FPGA, as well as Rule-based autonomic manager as in Section 6.1.3 . As we observed a similarity in objects and structures (e.g., tasks, implmentation 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 6.1.1 . This direction will also lead us to study the definition of architectural patterns for multiple loop Autonomic Managers, particularly hierarchical, with lower layers autonomy alleviating management burden from the upper layers.

6.1.3. Rule-based systems

Participants: Adja Sylla, Gwenaël Delaval, Eric Rutten.

This work concerns a high-level language for safe rule-based programming in the LINC transactional rule-based platform developed at CEA [10]. Rule based middlewares such as LINC enable high level programming of distributed adaptive systems behaviours. LINC also provides the systems with transactional guarantees and hence ensures their reliability at runtime. However, the set of rules may contain design errors (e.g. conflicts, violations of constraints) that can bring the system in unsafe safe or undesirables states, despite the guarantees provided by LINC. On the other hand, automata based languages such as Heptagon/BZR enable formal verification and especially synthesis of discrete controllers to deal with design errors. Our work studies these two languages and combines their execution mechanisms, from a technical perspective. We target applications to the domain of Internet of Things and more particularly smart building, office or home (see Section 6.2.2.1).

This work is in cooperation with CEA LETI/DACLE (Maxime Louvel), it was the topic of the PhD of Adja Sylla at CEA, co-advised with M. Louvel, and aspects on applications of logico-numeric control are published in the CCTA 2018 conference [20].

6.2. 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 [23] [9].

6.2.1. High-Performance Computing

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

6.2.1.1. Automated regulation and software transactional memory

A parallel program needs to manage the trade-off between the time spent in synchronisation and computation. This trade-off is significantly affected by its parallelism degree. A high parallelism degree may decrease computing time while increasing synchronisation cost. We performed work on dynamic control of thread parallelism and mapping. We address concurrency issues via Software Transactional Memory (STM). We implement feedback control loops to automate management of threads and diminish program execution time.

This work was performed in the framework on the PhD of Naweiluo Zhou, and published in the journal on Concurrency and Computation: Practice and Experience [13].

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

HPC systems are facing more and more variability in their behavior, related to e.g., performance and power consumption, and the fact that they are less predictable requires more runtime management. 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. This work resulted in first results addressing the problem of automated resource management in an HPC infrastructure, using techniques from Control Theory to design a controller that maximizes cluster utilization while avoiding overload. We put in place a mechanism for feedback (Proportional Integral, PI) control system software, through a maximum number of jobs to be sent to the cluster, in response to system information about the current number of jobs processed. Additionally, we developed a Model-Predictive Controller to improve the performance of the system.

This work is done in cooperation with the Datamove team of Inria/LIG, and Gipsa-lab. It was the topic of the Master's thesis of Agustin Yabo [25]. Preliminary results were published in the AIScience workshop (Autonomous Infrastructure for Science) of the HPDC conference [19].

6.2.1.3. Reconfiguration control in DPR FPGA

6.2.1.3.1. DPR FPGA and discrete control for reconfiguration

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

This work is in the framework of the ANR project HPeC (see Section 8.2.1), and is published in the International Workshop on High Performance and Dynamic Reconfigurable Systems and Networks (DRSN 2018), part of the HPCS 2018 conference [17] ; for the evaluation of the apllication of logico-numeric control, in the CCTA 18 conference [15] ; for the proposal of a Domain Specific Language, in the ICAC 2018 conference [16].

6.2.1.3.2. 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 on the PhD of Chabha Hireche, and published in the journal on Sensors [24], [12].

6.2.2. IoT

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

6.2.2.1. Control of smart buildings

A smart environment is equipped with numerous devices (i.e., sensors, actuators) that are possibly distributed over different locations (e.g., rooms of a smart building). These devices are automatically controlled to achieve different objectives related, for instance, to comfort, security and energy savings. Our work proposes a design framework based on the combination of the rule based middleware LINC and the automata based language Heptagon/BZR (H/BZR). It consists of: an abstraction layer for the heterogeneity of devices, a transactional execution mechanism to avoid inconsistencies and a controller that, based on a generic model of the environment, makes appropriate decisions and avoids conflicts. A case study with concrete devices, in the field of building automation, is presented to illustrate the framework.

This work is in the framework of the cooperation with CEA (see Section 7.1), and is published in the CCTA 2018 conference [20].

6.2.2.2. Device management

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

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

6.2.2.3. Security in SCADA industrial systems

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

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

This work is in the framework of the ANR project Sacade on cybersecurity of industrial systems (see Section 8.2.2) [18] [22] [21].

The work is also supported by Grenoble Alpes Cybersecurity Institute (see Section 8.1.1).

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.

DELYS 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 [correspondent], Pierre Sens, Julien Sopena.

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

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

In 2018, we obtained the following results.

5.1.1. Scheduling in uncertain environments

In [19], we consider scheduling with faults/errors and we introduce a new non-probabilistic model with explorable (query-able) uncertainty. Each unit-time error is characterized by an uncertainty area during which the error will occur, and it is possible to learn the exact slot at which it will appear by issuing a query operation of unit cost. We study two problems: (i) the error-query scheduling problem, whose aim is to reveal enough error-free slots with the minimum number of queries, and (ii) the lexicographic error-query scheduling problem where we seek the earliest error-free slots with the minimum number of queries. We consider both the off-line and the on-line versions of the above problems. In the former, the whole instance and its characteristics are known in advance and we give a polynomial-time algorithm for the error-query scheduling problem. In the latter, the adversary has the power to decide, in an on-line way, the time-slot of appearance for each error. We propose then both lower bounds and algorithms whose competitive ratios asymptotically match these lower bounds.

5.1.2. Failure detectors in dynamic systems

The failure detector abstraction was introduced as a way to circumvent the impossibility of solving consensus in asynchronous systems prone to crash failures. A failure detector is a local oracle that provides processes in the system with unreliable information on process failures. But a failure detector that is sufficient to solve a given problem in a static system is not necessarily sufficient to solve the same problem in a dynamic system. In [37], we adapt an existing failure detector for mutual exclusion and prove that it is the weakest failure detector to solve mutual exclusion in dynamic systems, which means that it is weaker than any other failure detector capable of solving mutual exclusion.

We also propose in [15] a new failure detector, called the Impact failure detector (FD), that expresses the confidence with regard to the system as a whole. Similarly to a reputation approach, it is possible to indicate the relative importance of each process of the system, while a threshold offers a degree of flexibility for failures and false suspicions. Performance evaluation results, based on real PlanetLab traces, confirm the degree of flexible of the failure detector.

5.1.3. Causal information dissemination

A causal broadcast ensures that messages are delivered to all nodes (processes) preserving causal relation of the messages. In [33], we propose a new causal broadcast protocol for distributed systems whose nodes are logically organized in a virtual hypercube-like topology called VCube. Messages are broadcast by dynamically building spanning trees rooted in the message's source node. By using multiple trees, the contention bottleneck problem of a single root spanning tree approach is avoided. Furthermore , different trees can intersect at some node. Hence, by taking advantage of both the out-of-order reception of causally related messages at a node and these paths intersections, a node can delay to one or more of its children in the tree. Experimental evaluation conducted on top of PeerSim simulator confirms the communication effectiveness of our causal broadcast protocol in terms of latency and message traffic reduction

5.1.4. Graceful Degradation

Gracefully degrading algorithms was introduced by Biely *et al.*i. Such algorithms offer the desirable properties to circumvent impossibility results in dynamic systems by adapting themselves to the dynamics. Indeed, such a algorithms solve a given problem under some dynamics and, moreover, guarantees that a weaker (but related) problem is solved under a higher dynamics under which the original problem is impossible to solve. The underlying intuition is to solve the problem whenever possible but to provide some kind of quality of service if the dynamics become (unpredictably) higher.

In [36], we apply for the first time this approach to robot networks. We focus on the fundamental problem of gathering a squad of autonomous robots on an unknown location of a dynamic ring. In this goal, we introduce a set of weaker variants of this problem. Motivated by a set of impossibility results related to the dynamics of the ring, we propose a gracefully degrading gathering algorithm.

5.1.5. Unreliable Hints

In [23], we address the question of a mobile agent deterministically searching for a target in the Euclidean plane. We assume that the mobile agent is equipped with a compass and a measure of length has to find an inert treasure in the Euclidean plane. Both the agent and the treasure are modeled as points. In the beginning, the agent is at a distance at most $D > 0$ from the treasure, but knows neither the distance nor any bound on it. Finding the treasure means getting at distance at most 1 from it. The agent makes a series of moves. Each of them consists in moving straight in a chosen direction at a chosen distance. In the beginning and after each move the agent gets a hint consisting of a positive angle smaller than 2π whose vertex is at the current position of the agent and within which the treasure is contained. We investigate the problem of how these hints permit the agent to lower the cost of finding the treasure, using a deterministic algorithm, where the cost is the worst-case total length of the agent's trajectory. It is well known that without any hint the optimal (worst case) cost is $\Theta(D^2)$. We show that if all angles given as hints are at most π, then the cost can be lowered to $O(D)$, which is optimal. If all angles are at most β, where $\beta < 2\pi$ is a constant unknown to the agent, then the cost is at most $O(D^2 - \epsilon)$, for some $\epsilon > 0$. For both these positive results we present deterministic algorithms achieving the above costs. Finally, if angles given as hints can be arbitrary, smaller than 2π, then we show that cost $\Theta(D^2)$ cannot be beaten.

5.1.6. 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] and [22], we consider these tasks from a deterministic point of view in networks modeled as undirected and anonymous graphs. An adversary chooses the initial nodes of the agents (the number of agents may be larger than the number of nodes) and assigns a different positive integer (called label) to each of them. Initially, each agent knows its label as well as some global knowledge shared by all the agents. The agents can communicate with each other only when located at the same node.

This task has been considered in the literature under two alternative scenarios: weak and strong. Under the weak scenario, agents may meet either at a node or inside an edge. Under the strong scenario, they have to meet at a node, and they do not even notice meetings inside an edge. Gathering and rendezvous algorithms under the strong scenario are known for synchronous agents. For asynchronous agents, gathering and rendezvous under the strong scenario are impossible even in the two-node graph, and hence only algorithms under the weak scenario were constructed.

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.

Gathering mobile agents can be made drastically more difficult to achieve when some agents are subject to faults, especially the Byzantine ones that are known as being the worst faults to handle. Byzantine means that the agent is subject to unpredictable and arbitrary faults. For instance, such an agent may choose to never stop or to never move. In [22] we study the task of Byzantine gathering among synchronous agents under the strong scenario: despite the presence of f Byzantine agents, all the other (correct) agents have to meet at the same node. In this respect, assuming that the agents are in a *strong team* i.e., a team in which the number of correct agents is at least some prescribed value that is quadratic in f, we show an algorithm that solves Byzantine gathering with all strong teams in all graphs of size at most n, for any integers n and f, in a time polynomial in n and the length $|l_{min}|$ of the binary representation of the smallest label of a good agent. The algorithm works using a global knowledge of size $\mathcal{O}(\log\log\log n)$, which we prove to be of optimal order of magnitude in our context to reach a time complexity that is polynomial in n and $|l_{min}|$.

5.1.7. Self-Stabilizing Minimum Diameter Spanning Tree

In [13], we present a self-stabilizing algorithm for the minimum diameter spanning tree construction problem in the state model. Our protocol has the following attractive features. It is the first algorithm for this problem that operates under the *unfair and distributed* adversary (or *daemon*). In other words, no restriction is made on the asynchronous behavior of the system. Second, our algorithm needs only $O(\log n)$ bits of memory per process (where n is the number of processes), that improves the previous result by a factor n. These features are not achieved to the detriment of the convergence time, which stays polynomial.

5.2. Large-scale data distribution

Participants: Saalik Hatia, Mesaac Makpangou, Sébastien Monnet, Sreeja Nair, Jonathan Sid-Otmane, Pierre Sens, Marc Shapiro, Alejandro Tomsic, Ilyas Toumlilt, Dimitrios Vasilas, Paolo Viotti.

5.2.1. Impossibility results for distributed transactional reads

We study the costs and trade-offs of providing transactional consistent reads in a distributed storage system. We identify the following dimensions: read consistency, read delay (latency), and data freshness. We show that there is a three-way trade-off between them, which can be summarised as follows: (i) it is not possible to ensure at the same time order-preserving (e.g., causally-consistent) or atomic reads, Minimal Delay, and maximal freshness; thus, reading data that is the most fresh without delay is possible only in a weakly-isolated mode; (ii) to ensure atomic or order-preserving reads at Minimal Delay imposes to read data from the past (not fresh); (iii) however, order-preserving minimal-delay reads can be fresher than atomic; (iv) reading atomic or order-preserving data at maximal freshness may block reads or writes indefinitely. Our impossibility results hold independently of other features of the database, such as update semantics (totally ordered or not) or data model (structured or unstructured). Guided by these results, we modify an existing protocol to ensure minimal-delay reads (at the cost of freshness) under atomic-visibility and causally-consistent semantics. Our experimental evaluation supports the theoretical results.

This work was published at Middleware 2018 [31].

5.2.2. Co-design and verification of an available file system

Distributed file systems play a vital role in large-scale enterprise services. However, the designer of a distributed file system faces a vexing choice between strong consistency and asynchronous replication. The former supports a standard sequential model by synchronising operations, but is slow and fragile. The latter is highly available and responsive, but exposes users to concurrency anomalies. We describe a rigorous and general approach to navigating this trade-off by leveraging static verification tools that allow to verify different file system designs. We show that common file system operations can run concurrently without synchronisation, while still retaining a semantics reasonably similar to Posix hierarchical structure. The one exception is the "move" operation, for which we prove that, unless synchronised, it will have an anomalous behaviour.

This work was published at VMCAI 2018 [28].

5.3. Resources management in system software

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

5.3.1. Multicore schedulers

In collaboration with WHISPER team, we have contributed to an analysis of the impact on application performance of the design and implementation choices made in two widely used open-source schedulers: ULE, the default FreeBSD scheduler, and CFS, the default Linux scheduler. In a paper published at USENIX ATC'18 [24], we compare ULE and CFS in otherwise identical circumstances. This work involves porting ULE to Linux, and using it to schedule all threads that are normally scheduled by CFS. We compare the performance of a large suite of applications on the modified kernel running ULE and on the standard Linux kernel running CFS. The observed performance differences are solely the result of scheduling decisions, and do not reflect differences in other subsystems between FreeBSD and Linux. We found that there is no overall winner. On many workloads the two schedulers perform similarly, but for some workloads there are significant and even surprising differences. ULE may cause starvation, even when executing a single application with identical threads, but this starvation may actually lead to better application performance for some workloads. The more complex load balancing mechanism of CFS reacts more quickly to workload changes, but ULE achieves better load balance in the long run.

MIMOVE Project-Team

7. New Results

7.1. Ontology categorization for IoT semantics

Participants: Rachit Agarwal, Nikolaos Georgantas, Valérie Issarny.

IoT systems are now being deployed worldwide to sense phenomena of interest. The existing IoT systems are often independent which limits the use of sensor data to only one application. Semantic solutions have been proposed to support reuse of sensor data across IoT systems and applications. This allows integration of IoT systems for increased productivity by solving challenges associated with their interoperability and heterogeneity. Several ontologies have been proposed to handle different aspects of sensor data collection in IoT systems, ranging from sensor discovery to applying reasoning on collected sensor data for drawing inferences. In this work, we study and categorise the existing ontologies based on the fundamental ontological concepts (e.g., sensors, context, location, and more) required for annotating different aspects of data collection and data access in an IoT application. We identify these fundamental concepts by answering the 4Ws (What, When, Who, Where) and 1H (How) identified using the 4W1H methodology.

7.2. Massively-Parallel Feature Selection for Big Data

Participant: Vassilis Christophides.

We present the Parallel, Forward-Backward with Pruning (PFBP) algorithm for feature selection (FS) in Big Data settings (high dimensionality and/or sample size). To tackle the challenges of Big Data FS, PFBP partitions the data matrix both in terms of rows (samples, training examples) as well as columns (features). By employing the concepts of p-values of conditional independence tests and meta-analysis techniques, PFBP manages to rely only on computations local to a partition while minimizing communication costs. Then, it employs powerful and safe (asymptotically sound) heuristics to make early, approximate decisions, such as Early Dropping of features from consideration in subsequent iterations, Early Stopping of consideration of features within the same iteration, or Early Return of the winner in each iteration. PFBP provides asymptotic guarantees of optimality for data distributions faithfully representable by a causal network (Bayesian network or maximal ancestral graph). Our empirical analysis confirms a superlinear speedup of the algorithm with increasing sample size, linear scalability with respect to the number of features and processing cores, while dominating other competitive algorithms in its class.

7.3. Universal Social Network Bus

Participants: Ehsan Ahvar, Shohreh Ahvar, Rafael Angarita, Nikolaos Georgantas, Valérie Issarny, Bruno Lefèvre.

Online social network services (OSNSs) are changing the fabric of our society, impacting almost every aspect of it. Over the last decades, the 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 USNB -*Universal Social Network Bus*, 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–, –profile- and non-profile-based– OSNSs through a *reference social interaction service*.

USNB enables users to reach out to their social peers independently of the communication service (and especially underlying platform) each one uses in the virtual world. The success and massive adoption of OSNSs -as magnified by the success of Facebook- shows that online social communication is an essential tool for people. This further paves the way for collective and collaborative actions at the Internet scale. However, existing online collaborative tools come along with their communication platform, which is either a proprietary solution or a third-party OSNS. We argue that USNB contributes to enabling participatory systems at a larger inclusive scale by overcoming the technical boundaries set by existing online communication platforms. In that direction, we investigate the customization of USNB for specific applications and more specifically: participatory systems and massive open online courses.

7.4. Middleware for Mobile Crowdsensing

Participants: Yifan Du, Valérie Issarny, Bruno Lefèvre, Françoise Sailhan.

Mobile Phone Sensing (MPS) offers a great opportunity toward the large scale monitoring of urban phenomena, such as the exposition of the population to environmental pollution. Indeed, mobile crowdsensing empowers ordinary citizens to contribute (whether pro-actively or passively) data sensed or generated from their mobile devices. It allows acquiring hyperlocal knowledge at scale, thanks to the proliferation of mobile devices and the ubiquity of wireless broadband connection. On-demand mobile crowdsensing is in particular a cost-effective service model for smart cities. Numerous sensor types embedded in today's smartphones contribute valuable quantitative observations about the urban environment (e.g., noise, temperature, atmospheric pressure, humidity, light, magnetism). The observations further come along with the related spatial and temporal data, which allows for the analysis of hyper-local environmental knowledge. However, mobile crowdsensing brings valuable knowledge only if a sufficiently large crowd contributes and if we overcome the relatively low accuracy of the gathered data. This is the focus of our research.

We have in particular studied how to reduce the gap between the need for the massive collection of relevant data, and the quantity and accuracy of the measurements that are actually gathered. We specifically carried out an iterative research process to tackle this challenge, which combines technological innovation and social design. We have been developing a number of social tools to study the motivations and usages of MPS-based smart city apps, with the Ambiciti app serving as our use case. Our study has been taking into account the cultural and societal contexts that the usages of Ambiciti could feed, spanning health, environment, education, and urban policies. We carried out an online survey together with interviews with users and local actors in Europe, i.e., France, Belgium, and Finland. The research results contribute to a better understanding of why and how people use mobile phone sensing applications; the results also inform how to best leverage mobile crowd-sensing in the development of smart cities and how it may serve addressing urban challenges related to, e.g., public health or urban planning.

The quality of the contributed measurements challenges the aggregation of relevant knowledge from crowdsensed observations. The measurements quality depends on the *accuracy* of the contributing sensors and the adequacy of the *sensing context*. Addressing the former relies on the sensor calibration for which we study both micro- and macro-level solutions. Addressing the latter requires a supporting inference mechanism, for which we introduce a *personalized hierarchical inference* of all the context elements that are relevant to the phenomenon that is monitored through crowdsensing, and under which the crowdsensor operates. This enables accounting for the specific behavior of the contributing end-user across time, as well as for all the features -and only those- that are relevant and locally available, while reducing the feedback required from the user for the personalization.

7.5. QoS-Aware Resource Allocation for Mobile IoT Pub/Sub Systems

Participants: Georgios Bouloukakis, Nikolaos Georgantas.

IoT applications are usually characterized by large-scale demand and the widespread use of mobile devices. Similarly, performing interaction among application and system components in a decoupled and elastic way, and enforcing Quality of Service (QoS) usually also become issues. Hence, paradigms such as pub/sub on top of cloud resources represent a suitable strategy for application development. However, management of QoS-aware resource allocation for pub/sub systems remains challenging, especially when system peers connect in an intermittent way. In this work, we propose a new approach for resource allocation focusing on end-to-end performance in face of peers' disconnections. We evaluate and demonstrate the benefits of our approach using simulations. QoS enforcement was achieved in almost all scenarios, and we have shown that our approach can help reasoning about efficient resource allocation.

7.6. Queueing Network Modeling Patterns for Reliable & Unreliable Pub/Sub Protocols

Participants: Georgios Bouloukakis, Nikolaos Georgantas, Patient Ntumba, Valérie Issarny.

Mobile Internet of Things (IoT) applications are typically deployed on resource-constrained devices with intermittent network connectivity. To support the deployment of such applications, the Publish/Subscribe (pub/sub) interaction paradigm is often employed, as it decouples mobile peers in time and space. Furthermore, pub/sub middleware protocols and APIs consider the Things' hardware limitations and support the development of effective applications by providing Quality of Service (QoS) features. These features aim to enable developers to tune an application by switching different levels of response times and delivery success rates. However, the profusion of pub/sub middleware protocols coupled with intermittent network connectivity result in non-trivial application tuning. In this work, we model the performance of middleware protocols found in IoT, which are classified within the pub/sub interaction paradigm – both reliable and unreliable underlying network layers are considered. We model reliable and unreliable protocols, by considering QoS semantics for data validity, buffer capacities, as well as the intermittent availability of peers. To this end, we rely on queueing network models, which offer a simple modeling environment that can be used to represent IoT interactions by combining multiple queueing model types. Based on these models, we perform statistical analysis by varying the QoS semantics, demonstrating their significant effect on response times and on the rate of successful interactions. We showcase the application of our analysis in concrete scenarios relating to Traffic Information Management systems, that integrate both reliable and unreliable participants. The consequent PerfMP performance modeling pattern may be tailored for a variety of deployments, in order to control fine-grained QoS policies.

7.7. Lightweight, General Inference of Streaming Video Quality from Encrypted Traffic

Participants: Francesco Bronzino, Sara Ayoubi, Renata Teixeira, Sarah Wasserman.

Accurately monitoring application performance is becoming more important for Internet Service Providers (ISPs), as users increasingly expect their networks to consistently deliver acceptable application quality. At the same time, the rise of end-to-end encryption makes it difficult for network operators to determine video stream quality–including metrics such as startup delay, resolution, rebuffering, and resolution changes–directly from the traffic stream. This work develops general methods to infer streaming video quality metrics from encrypted traffic using lightweight features. Our evaluation shows that our models are not only as accurate as previous approaches, but they also generalize across multiple popular video services, including Netflix, YouTube, Amazon Instant Video, and Twitch. The ability of our models to rely on lightweight features points to promising future possibilities for implementing such models at a variety of network locations along the end-to-end network path, from the edge to the core.

7.8. Service traceroute: Tracing Paths of Application Flows

Participants: Ivan Morandi, Francesco Bronzino, Renata Teixeira.

Traceroute is often used to help diagnose when users experience issues with Internet applications or services. Unfortunately, probes issued by classic traceroute tools differ from application traffic and hence can be treated differently by middleboxes within the network. This work proposes a new traceroute tool, called Service traceroute. Service traceroute leverages the idea from paratrace, which passively listens to application traffic to then issue traceroute probes that pretend to be part of the application flow. We extend this idea to work for modern Internet services with support for identifying the flows to probe automatically, for tracing of multiple concurrent flows, and for UDP flows. We implement command-line and library versions of Service traceroute, which we release as open source. This paper also presents an evaluation of Service traceroute when tracing paths traversed by Web downloads from the top-1000 Alexa websites and by video sessions from Twitch and Youtube. Our evaluation shows that Service traceroute has no negative effect on application flows. Our comparison with Paris traceroute shows that a typical traceroute tool that launches a new flow to the same destination discovers different paths than when embedding probes in the application flow in a significant fraction of experiments (from 40% to 50% of our experiments in PlanetLab Europe).

MYRIADS Project-Team

7. New Results

7.1. Scaling Clouds

7.1.1. Fog Computing

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

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

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

We investigated the challenges of efficiently deploying Docker containers in fog platforms composed of tiny single-board computers such as Raspberry PIs, and demonstrated that major performance gains can be obtained with relatively simple modifications in the way Docker imports container images [12]. This work is currently being extended in a variety of ways: exploiting distributed storage services to share image among fog nodes, reorganizing the Docker images to allow them to be booted before the image has been fully downloaded, exploiting checkpoint/restart mechanisms to efficiently deploy application that have a long startup time. We expect a few publications on these topics in the coming year.

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

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

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

7.1.2. Community Clouds

Participants: Jean-Louis Pazat, Bruno Stevant.

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

7.1.3. Stream Processing

Participants: Cédric Tedeschi, Mehdi Belkhiria.

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

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

Participants: Yasmina Bouizem, Christine Morin, Nikos Parlavantzas.

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

7.1.5. Cost-effective Reconfiguration for Multi-cloud Applications

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

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

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

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

In the context of our collaboration with Thales Research and Technology and Baptiste Goupille-Lescar's PhD work, we are applying cloud resource management techniques to high-performance, multi-sensor, embedded systems with real-time constraints. The objective is to increase the flexibility and efficiency of resource allocation in such systems, enabling the execution of dynamic sets of applications with strict QoS requirements. In 2018, we proposed an online scheduling approach for executing real-time applications on heavily-constrained embedded architectures. The approach enables dynamically allocating resources to fulfill

requests coming from several sensors, making the most of the computing platform, while providing guaranties on quality of service levels. The approach was tested in an industrial use case concerning the operation of a multi-function surface active electronically scanned array (AESA) radar. We showed that the approach allows us to obtain lower execution latencies than current mapping solutions while maintaining high predictability and allowing gradual performance degradation in overload scenarios [22].

7.2. Greening Clouds

7.2.1. Energy Models

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

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

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

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

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

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

Participant: Anne-Cécile Orgerie.

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

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

We investigate the end-to-end energy consumption of IoT platforms. Our aim is to evaluate, on concrete use-cases, the benefits of edge computing platforms for IoT regarding energy consumption. We aim at proposing end-to-end energy models for estimating the consumption when offloading computation from the objects to the edge or to the core Cloud, depending on the number of devices and the desired application QoS, in particular trading-off between performance (response time) and reliability (service accuracy). This work has been published in [10].

7.2.3. Exploiting Renewable Energy in Distributed Clouds

Participants: Benjamin Camus, Anne-Cécile Orgerie.

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

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

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

7.2.4. Smart Grids

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

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

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

7.2.5. Involving Users in Energy Saving

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

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

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

We explore possible collaborations between virtual machine suppliers, service providers and users of Clouds in order to provide users with ways of participating in the reduction of the Clouds energy consumption. This work will follow two directions: 1) to investigate compromises between power and performance/service quality that cloud providers can offer to their users and to propose them a variety of options adapted to their workload; and 2) to develop mechanisms for each layer of the Cloud software stack to provide users with a quantification of the energy consumed by each of their options as an incentive to become greener. This work was explored in the context of David Guyon's PhD thesis (defended on December 7, 2018). For 2018, it resulted in one publication in the International Journal of Grid and Utility Computing [8] and two publications in conferences: IC2E [23] and SBAC-PAD [24].

7.3. Securing Clouds

7.3.1. Security Monitoring in Clouds

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

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

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

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

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

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

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

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

7.4. Experimenting with Clouds

7.4.1. Simulating Distributed IT Systems

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

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

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

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

7.4.2. Formal Methods for IT Systems

Participants: The Anh Pham, Martin Quinson.

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

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

7.4.3. Executing Epidemic Simulation Applications in the Cloud

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

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

7.4.4. Implicit locality awareness of Remote Procedure Calls evaluation

Participants: Javier Rojas Balderrama, Matthieu Simonin.

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

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

7.4.5. Tools for the experimentation

Participant: Matthieu Simonin.

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

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

SPIRALS Project-Team

7. New Results

7.1. Software Product Lines for Setup and Adaptation of Multi-Cloud Computing Systems

In 2018, in the domain of cloud computing, we proposed a new software product line-based approach for managing the variability in order to automate the setup and adaptation of multi-cloud environments. Building such systems is still very challenging and time consuming due to the heterogeneity across cloud providers' offerings and the high-variability in the configuration of cloud providers. This variability is expressed by the large number of available services and the many different ways in which they can be combined and configured. In order to ensure correct setup of a multi-cloud environment, developers must be aware of service offerings and configuration options from multiple cloud providers. Our results enable to automatically generate a configuration or reconfiguration plan for a multi-cloud environment from a description of its requirements. The conducted experiments aim to assess the impact of the approach on the automated analysis of feature models and the feasibility of the approach to automate the setup and adaptation of multi-cloud environments. These results have been obtained in the context of the PhD thesis of Gustavo Sousa [12] defended in June 2018.

7.2. Automated Software Repair with Patch Generation in Production

In 2018, in the domain of automated software repair, we proposed new patch generation techniques. Patch creation is one of the most important actions in the life cycle of an application. Creating patches is a time-consuming task. Not only because it is difficult to create a sound and valid patch, but also because it requires the intervention of humans. Our work proposes new patch generation techniques that remove the human intervention. Our idea is to put as close as possible the patch generation in the production environment. We adopt this approach because the production environment contains all the data and human interactions that lead to the bug. We show how to exploit this data to detect bugs, generate and validate patches. We evaluate this approach on seven different benchmarks of real bugs collected from open-source projects. During the evaluation, we are particularly attentive to the number of generated patches, to their correctness, readability and to the time required for generating them. Our evaluation shows the applicability and feasibility of our approach to generate patches in the production environment without the intervention of a developer. These results have been obtained in the context of the PhD thesis of Thomas Durieux [11] defended in September 2018.

7.3. Flexible Framework for Elasticity in Cloud Computing

In 2018, in the domain of cloud computing, we proposed a new framework for managing elasticity. The main factor motivating the use of cloud is its ability to provide resources according to the customer needs or what is referred to as elasticity. Adapting cloud applications during their execution according to demand variation is nevertheless a challenging task. In addition, cloud elasticity is diverse and heterogeneous because it encompasses different approaches, policies, purposes, etc. In this work, three contributions are proposed: (1) an up-to-date state-of-the-art of the cloud elasticity for both virtual machines and containers, (2) ELASTIC-DOCKER, an approach to manage container elasticity including vertical elasticity, live migration, and elasticity combination between different virtualization techniques, and (3) MODEMO, a new unified standard-based, model-driven, highly extensible and reconfigurable framework that supports multiple elasticity policies, vertical and horizontal elasticity, different virtualization techniques and multiple cloud providers. These results have been obtained in the context of the PhD thesis of Yahya Al-Dhuraibi defended in December 2018.

7.4. Semantic Interoperability in Multi-Cloud Computing Systems

In 2018, in the domain of cloud computing, we proposed two major results related to semantic interoperability. First, an approach based on reverse-engineering to extract knowledge from the ambiguous textual documentation of cloud APIs and to enhance its representation using MDE techniques has been proposed. This approach is applied to Google Cloud Platform (GCP), where we provide GCP Model, a precise model-driven specification for GCP. GCP Model is automatically inferred from GCP textual documentation, conforms to the OCCIware Metamodel and is implemented within OCCIware Studio. It allows one to perform qualitative and quantitative analysis of the GCP documentation. Second, we have proposed the FCLOUDS framework to achieve semantic interoperability in multi-clouds, i.e., to identify the common concepts between cloud APIs and to reason over them. The FCLOUDS language is a formalization of OCCI concepts and operational semantics in Alloy formal specification language. To demonstrate the effectiveness of the FCLOUDS language, we formally specify thirteen case studies and verify their properties. Then, thanks to formal transformation rules and equivalence properties, we draw a precise alignment between my case studies, which promotes semantic interoperability in multi-clouds. These results have been obtained in the context of the PhD thesis of Stéphanie Challita defended in December 2018.

WHISPER Project-Team

7. New Results

7.1. Software engineering for infrastructure software

The most visible tool developed in the Whisper team is Coccinelle, which this year marked the 10th anniversary of its release in open source. The paper "Coccinelle: 10 Years of Automated Evolution in the Linux Kernel," published at USENIX ATC'18 [14], traced the history of Coccinelle, its underlying design decisions and impact. The Coccinelle C-code matching and transformation tool was first released in 2008 to facilitate specification and automation in the evolution of Linux kernel code. The novel contribution of Coccinelle was to allow software developers to write code manipulation rules in terms of the code structure itself, via a generalization of the patch syntax. Over the years, Coccinelle has been extensively used in Linux kernel development, resulting in over 6000 commits to the Linux kernel, and has found its place as part of the Linux kernel development process. The USENIX ATC paper studies the impact of Coccinelle on Linux kernel development and the features of Coccinelle that have made it possible. It provides guidance on how other research-based tools can achieve practical impact in the open-source development community. This work was also presented to Linux kernel developers at Kernel Recipes and Open Source Summit Europe, and at the 8th Inria/Technicolor Workshop On Systems.

In a modern OS, kernel modules often use spinlocks and interrupt handlers to monopolize a CPU core to execute concurrent code in atomic context. In this situation, if the kernel module performs an operation that can sleep at runtime, a system hang may occur. We refer to this kind of concurrency bug as a sleep-in-atomic-context (SAC) bug. In practice, SAC bugs have received insufficient attention and are hard to find, as they do not always cause problems in real executions. In a paper published at USENIX ATC'18 [12], we propose a practical static approach named DSAC, to effectively detect SAC bugs and automatically recommend patches to help fix them. DSAC uses four key techniques: (1) a hybrid of flow-sensitive and -insensitive analysis to perform accurate and efficient code analysis; (2) a heuristics-based method to accurately extract kernel interfaces that can sleep at runtime; (3) a path-check method to effectively filter out repeated reports and false bugs; (4) a pattern-based method to automatically generate recommended patches to help fix the bugs. We evaluate DSAC on kernel modules (drivers, file systems, and network modules) of the Linux kernel, and on the FreeBSD and NetBSD kernels, and in total find 401 new real bugs. 272 of these bugs have been confirmed by the relevant kernel maintainers, and 43 patches generated by DSAC have been applied by kernel maintainers.

7.2. Trustworthy domain-specific compilers

To achieve safety and composability, we believe that an holistic approach is called for, involving not only the design of a domain-specific *syntax* but also of a domain-specific *semantics*. Concretely, we are exploring the design of *certified domain-specific compilers* that integrate, from the ground up, a denotational and domain-specific semantics as part of the design of a domain-specific language. This vision is illustrated by our work on the safe compilation of Coq programs into secure OCaml code [10]. It combines ideas from gradual typing – through which types are compiled into run-time assertions – and the theory of ornaments [31] – through which Coq datatypes can be related to OCaml datatypes. Within this formal framework, we enable a secure interaction, termed *dependent interoperability*, between correct-by-construction software and untrusted programs, be it system calls or legacy libraries. To do so, we trade static guarantees for runtime checks, thus allowing OCaml values to be safely coerced to dependently-typed Coq values and, conversely, to expose dependently-typed Coq programs defensively as OCaml programs. Our framework is developed in Coq: it is constructive and verified in the strictest sense of the terms. It thus becomes possible to internalize and hand-tune the extraction of dependently-typed programs to interoperable OCaml programs within Coq itself. This work is the result of a collaboration with Eric Tanter, from the University of Chile, and Nicolas Tabareau, from the Gallinette Inria project-team.

7.3. High-performance domain-specific compilers

As part of Darius Mercadier's PhD project, we are developing a synchronous dataflow language targeting high-performance (and, eventually, verified) implementations of bitsliced algorithms, with application to cryptographical algorithms [33]. Using our Usuba language, cryptographers can specify a block cipher at a very high level as a set of dataflow equations. From such a description, our `usubac` compiler is able to generate efficient, vectorized code exploiting the SIMD instruction sets of the underlying architecture. We have demonstrated that our generated code performs on par with hand-tuned assembly programs while, at the same time, being able to target multiple CPU architectures as well as multiple generations of SIMD instruction sets on each architecture. This project illustrates perfectly our methodology: the design of Usuba is driven by semantic considerations (bitslicing is only meaningful for bit parallel operations) that are then structured using types and subsequently reifed into syntactic artefacts. Our preliminary results [15], published in an international workshop, are encouraging.

7.4. Multicore schedulers

As a side-effect of our work on verification of schedulers [48], we have contributed to an analysis of the impact on application performance of the design and implementation choices made in two widely used open-source schedulers: ULE, the default FreeBSD scheduler, and CFS, the default Linux scheduler. In a paper published at USENIX ATC'18 [13], we compare ULE and CFS in otherwise identical circumstances. This work involves porting ULE to Linux, and using it to schedule all threads that are normally scheduled by CFS. We compare the performance of a large suite of applications on the modified kernel running ULE and on the standard Linux kernel running CFS. The observed performance differences are solely the result of scheduling decisions, and do not reflect differences in other subsystems between FreeBSD and Linux. We found that there is no overall winner. On many workloads the two schedulers perform similarly, but for some workloads there are significant and even surprising differences. ULE may cause starvation, even when executing a single application with identical threads, but this starvation may actually lead to better application performance for some workloads. The more complex load balancing mechanism of CFS reacts more quickly to workload changes, but ULE achieves better load balance in the long run.

WIDE Project-Team

6. New Results

6.1. Scalable Systems

6.1.1. Nobody cares if you liked Star Wars: KNN graph construction on the cheap.

Participants: Olivier Ruas, François Taïani.

K-Nearest-Neighbors (KNN) graphs play a key role in a large range of applications. A KNN graph typically connects entities characterized by a set of features so that each entity becomes linked to its k most similar counterparts according to some similarity function. As datasets grow, KNN graphs are unfortunately becoming increasingly costly to construct, and the general approach, which consists in reducing the number of comparisons between entities, seems to have reached its full potential. In work [27] we propose to overcome this limit with a simple yet powerful strategy that samples the set of features of each entity and only keeps the least popular features. We show that this strategy outperforms other more straightforward policies on a range of four representative datasets: for instance, keeping the 25 least popular items reduces computational time by up to 63%, while producing a KNN graph close to the ideal one.

This work was done in collaboration with Anne-Marie Kermarrec (Mediego/EPFL).

6.1.2. Pleiades: Distributed structural invariants at scale

Participants: Simon Bouget, David Bromberg, Adrien Luxey, François Taïani.

Modern large scale distributed systems increasingly espouse sophisticated distributed architectures characterized by complex distributed structural invariants. Unfortunately, maintaining these structural invariants at scale is time consuming and error prone, as developers must take into account asynchronous failures, loosely coordinated sub-systems and network delays. To address this problem, we propose Pleiades [31], a new framework to construct and enforce large-scale distributed structural invariants under aggressive conditions. Pleiades combines the resilience of self-organizing overlays, with the expressiveness of an assembly-based design strategy. The result is a highly survivable framework that is able to dynamically maintain arbitrary complex distributed structures under aggressive crash failures. Our evaluation shows in particular that Pleiades is able to restore the overall structure of a 25,600 node system in less than 11 asynchronous rounds after half of the nodes have crashed.

6.1.3. CASCADE: Reliable distributed session handoff for continuous interaction across devices

Participants: David Bromberg, Adrien Luxey, François Taïani.

Allowing users to navigate seamlessly between their personal devices while protecting their privacy remains today an ongoing challenge. Existing solutions rely on peer-to-peer designs, and blindly flood the network with session messages. It is particularly hard to come up with proposals that are both cost-efficient and dependable while relying on poorly connected mobile appliances. In [24] we propose Cascade, a distributed protocol to share applicative sessions among one's devices. Our proactive session handoff algorithm takes inspiration from the BitTorrent P2P file sharing protocol, but adapts it to the specific characteristics of our problem. It eschews in particular trackers, and limits the seeders of each session to the devices most likely to be used next, as computed by a decentralized aggregation protocol. A key aspect of our approach is to trade off network costs for reliability, while providing a faster session handoff than centralized solutions in the vast majority of the cases.

6.1.4. Sprinkler: A probabilistic dissemination protocol to provide fluid user interaction in multi-device ecosystems

Participants: David Bromberg, Adrien Luxey, François Taïani.

Offering fluid multi-device interactions to users while protecting their privacy largely remains an ongoing challenge. Existing approaches typically use a peer-to-peer design and flood session information over the network, resulting in costly and often unpractical solutions. In [29], we propose Sprinkler, a decentralized probabilistic dissemination protocol that uses a gossip-based learning algorithm to intelligently propagate session information to devices a user is most likely to use next. Our solution allows designers to efficiently trade off network costs for fluidity, and is for instance able to reduce network costs by up to 80% against a flooding strategy while maintaining a fluid user experience.

This work was done in collaboration with Fabio Costa, Ricardo Da Rocha and Vinicius Lima from the Universidade Federal de Goias (UFG).

6.2. Personalization and Privacy

6.2.1. GoldFinger

Participants: Olivier Ruas, François Taïani.

In work [37] we propose fingerprinting, a new technique that consists in constructing compact, fast-to-compute and privacy-preserving representation of datasets. We illustrate the effectiveness of our approach on the emblematic big data problem of K-Nearest-Neighbor (KNN) graph construction and show that fingerprinting can drastically accelerate a large range of existing KNN algorithms, while efficiently obfuscating the original data, with little to no overhead. Our extensive evaluation of the resulting approach (dubbed GoldFinger) on several realistic datasets shows that our approach delivers speed-ups up to 78.9% compared to the use of raw data while only incurring a negligible to moderate loss in terms of KNN quality. To convey the practical value of such a scheme, we apply it to item recommendation, and show that the loss in recommendation quality is negligible.

This work was done in collaboration with Rachid Guerraoui (EPFL) and Anne-Marie Kermarrec (Mediego/EPFL).

6.2.2. Collaborative filtering under a Sybil attack: Similarity metrics do matter!

Participant: Davide Frey.

Recommendation systems help users identify interesting content, but they also open new privacy threats. For this reason, in [22] we deeply analyzed the effect of a Sybil attack that tries to infer information on users from a user-based collaborative-filtering recommendation systems. We evaluated the impact of different similarity metrics used to identity users with similar tastes in the trade-off between recommendation quality and privacy. Based on our results, we proposed and evaluated a novel similarity metric that combines the best of both worlds: a high recommendation quality with a low prediction accuracy for the attacker. Our experiments, on a state-of-the-art recommendation framework and on real datasets showed that existing similarity metrics exhibit a wide range of behaviors in the presence of Sybil attacks, while our new similarity metric consistently achieves the best trade-off while outperforming state-of-the-art solutions.

This work was carried out in collaboration with Antoine Boutet from INSA Lyon, former-intern Florestan De Moor, Rachid Guerraoui and Antoine Rault from EPFL, and Anne-Marie Kermarrec from Mediego.

6.3. Network and Graph Algorithms

6.3.1. Rumor spreading and conductance

Participant: George Giakkoupis.

In [16], we study the completion time of the PUSH-PULL variant of rumor spreading, also known as randomized broadcast. We show that if a network has n nodes and conductance ϕ then, with high probability, PUSH-PULL will deliver the message to all nodes in the graph within $O(\log n/\phi)$ many communication rounds. This bound is best possible. We also give an alternative proof that the completion time of PUSH-PULL is bounded by a polynomial in $\log n/\phi$, based on graph sparsification. Although the resulting asymptotic bound is not optimal, this proof shows an interesting and, at the outset, unexpected connection between rumor spreading and graph sparsification. Finally, we show that if the degrees of the two endpoints of each edge in the network differ by at most a constant factor, then both PUSH and PULL alone attain the optimal completion time of $O(\log n/\phi)$, with high probability.

This work was done in collaboration with Flavio Chierichetti (Sapienza University of Rome), Silvio Lattanzi (Google Research), and Alessandro Panconesi (Sapienza University of Rome).

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

Participant: George Giakkoupis.

A Coalescing-Branching Random Walk (CoBra) is a natural extension to the standard random walk on a graph. The process starts with one pebble at an arbitrary node. In each round of the process every pebble splits into k pebbles, which are sent to k random neighbors. At the end of the round all pebbles at the same node coalesce into a single pebble. The process is also similar to randomized rumor spreading, with each informed node pushing the rumor to k random neighbors each time it receives a copy of the rumor. Besides its mathematical interest, this process is relevant as an information dissemination primitive and a basic model for the spread of epidemics. In [21], we study the cover time of CoBra walks, which is the time until each node has seen at least one pebble. Our main result is a bound of $O(\log n/\phi)$ rounds with high probability on the cover time of a CoBra walk with $k = 2$ on any regular graph with n nodes and conductance ϕ. This bound improves upon all previous bounds in terms of graph expansion parameters. Moreover, we show that for any connected regular graph the cover time is $O(n \log n)$ with high probability, independently of the expansion. Both bounds are asymptotically tight. Since our bounds coincide with the worst-case time bounds for Push rumor spreading on regular graphs until all nodes are informed, this raises the question whether CoBra walks and Push rumor spreading perform similarly in general. We answer this negatively by separating the cover time of CoBra walks and the rumor spreading time of Push by a super-polylogarithmic factor on a family of tree-like regular graphs.

This work was done in collaboration with Petra Berenbrink and Peter Kling from the University of Hamburg.

6.3.3. The quadratic shortest path problem: Complexity, approximability, and solution methods

Participant: Davide Frey.

In work [20] we considered the problem of finding a shortest path in a directed graph with a quadratic objective function (the QSPP). We show that the QSPP cannot be approximated unless $\mathsf{P} = \mathsf{NP}$. For the case of a convex objective function, we presented an n-approximation algorithm, where n is the number of nodes in the graph, and we proved APX-hardness. Furthermore, we proved that even if only adjacent arcs play a part in the quadratic objective function, the problem still cannot be approximated unless $\mathsf{P} = \mathsf{NP}$. In order to solve the general problem we first proposed a mixed integer programming formulation, and then devised an efficient exact Branch-and-Bound algorithm for the general QSPP. This algorithm computes lower bounds by considering a reformulation scheme that is solvable through a number of minimum-cost-flow problems. We carried out computational experiments solving to optimality different classes of instances with up to 1000 nodes.

This work was carried out in collaboration with Borzou Rostami from Polytechnique Montréal, Adreé Chasssein and Michael Hopf from TU Kaiserslautern, Christoph Buchheim from TU Dortmund, Federico Malucelli from Politecnico di Milano, and Marc Goerigk from Lancaster University.

6.3.4. Weighting past on the geo-aware state deployment problem

Participant: François Taïani.

The geographical barrier between mobile devices and mobile application servers (typically hosted in the Cloud) imposes an unavoidable latency and jitter that negatively impacts the performance of modern mobile systems. Fog Computing architectures can mitigate this impact if there is a middleware service able to correctly partition and deploy the state of an application at optimal locations. Geo-aware state deployment is challenging as it must consider the mobility of the devices and the dependencies arising when multiple devices concurrently manipulate the same application state. In [28], we propose a range of new object-graph-based strategies for geo-aware state deployment. In particular, our investigation focuses on understanding the role of preserving previously observed associations between state items on application performance.

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

6.3.5. *Mind the gap: Autonomous detection of partitioned MANET systems using opportunistic aggregation*

Participants: Simon Bouget, David Bromberg, François Taïani.

Mobile Ad-hoc Networks (MANETs) use limited-range wireless communications and are thus exposed to partitions when nodes fail or move out of reach of each other. Detecting partitions in MANETs is unfortunately a nontrivial task due to their inherently decentralized design and limited resources such as power or bandwidth. In [32], we propose a novel and fully decentralized approach to detect partitions (and other *large* membership changes) in MANETs that is both accurate and resource efficient. We monitor the current composition of a MANET using the lightweight aggregation of compact membership-encoding filters. Changes in these filters allow us to infer the likelihood of a partition with a quantifiable level of confidence. We first present an analysis of our approach, and show that it can detect close to 100% of partitions under realistic settings, while at the same time being robust to false positives due to churn or dropped packets. We perform a series of simulations that compare against alternative approaches and confirm our theoretical results, including above 90% accurate detection even under a 40% message loss rate.

This work was performed in collaboration with Etienne Rivière from UC Louvain (Belgium) and Hugues Mercier from University of Neuchâtel (Switzerland).

6.4. Theory of Distributed Systems

6.4.1. *An improved bound for random binary search trees with concurrent insertions*

Participant: George Giakkoupis.

Recently, Aspnes and Ruppert (DISC 2016) defined the following simple random experiment to determine the impact of concurrency on the performance of binary search trees: n randomly permuted keys arrive one at a time. When a new key arrives, it is first placed into a buffer of size c. Whenever the buffer is full, or when all keys have arrived, an adversary chooses one key from the buffer and inserts it into the binary search tree. The ability of the adversary to choose the next key to insert among c buffered keys, models a distributed system, where up to c processes try to insert keys concurrently. Aspnes and Ruppert showed that the expected average depth of nodes in the resulting tree is $O(\log n + c)$ for a comparison-based adversary, which can only take the relative order of arrived keys into account. In work [25], we generalize and strengthen this result. In particular, we allow an adversary that knows the actual values of all keys that have arrived, and show that the resulting expected average node depth is $D_{avg}(n) + O(c)$, where $D_{avg}(n) = 2ln(n) - Theta(1)$ is the expected average node depth of a random tree obtained in the standard unbuffered version of this experiment. Extending the bound by Aspnes and Ruppert to this stronger adversary model answers one of their open questions.

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

6.4.2. *Acyclic strategy for silent self-stabilization in spanning forests*

Participant: Anaïs Durand.

Self-stabilization is a general paradigm to enable the design of distributed systems tolerating any finite number of transient faults. Many self-stabilizing algorithms are designed using the same patterns. In [30] we formalize some of those design patterns to obtain general statements regarding both correctness and time complexity. Precisely, we study a class of algorithms devoted to networks endowed with a sense of direction describing a spanning forest whose characterization is a simple (i.e., quasi-syntactic) condition. We show that any algorithm of this class is (1) silent and self-stabilizing under the distributed unfair daemon (the weakest scheduling assumption in the considered model), and (2) has a stabilization time polynomial in moves and asymptotically optimal in rounds. Our condition mainly uses the concept of acyclic strategy, which is based on the notions of top-down and bottom-up actions. We have combined this formalization together with a notion of acyclic causality between actions and a last criteria called correct-alone (n.b., only this criteria is not syntactic) to obtain the notion of acyclic strategy. We show that any algorithm following an acyclic strategy reaches a terminal configuration in a polynomial number of moves, assuming a distributed unfair daemon. Hence, if its terminal configurations satisfy the specification, the algorithm is both silent and self-stabilizing. Unfortunately, we show that this condition is not sufficient to obtain an asymptotically optimal stabilization time in rounds. So, we enforce the acyclic strategy with the property of local mutual exclusivity to have an asymptotically optimal round complexity. We also propose a simple method to make any algorithm, that follows an acyclic strategy, locally mutually exclusive. This method has no overhead in moves. Finally, to show the versatility of our approach, we review works where our results apply.

This work was done in collaboration with Karine Altisen and Stéphane Devismes (VERIMAG, Université Grenoble Alpes).

6.4.3. *Set agreement and renaming in the presence of contention-related crash failures*

Participants: Anaïs Durand, Michel Raynal.

Given a predefined contention threshold λ, consider all executions in which process crashes are restricted to occur only when process contention is smaller than or equal to λ. If crashes occur after contention bypassed λ, there are no correctness guarantees (e.g., termination is not guaranteed). It is known that, when $\lambda = n-1$, consensus can be solved in an n-process asynchronous read/write system despite the crash of one process, thereby circumventing the well-known FLP impossibility result. Furthermore, it was shown that when $\lambda = n-k$ and $k \geq 2$, k-set agreement can be solved despite the crash of $2k-2$ processes.

In work [33] we consider two types of process crash failures: λ-constrained crash failures (as previously defined), and classical crash failures (that we call *any time* failures). We present two algorithms suited to these types of failures. The first algorithm solves k-set agreement, where $k = m + f$, in the presence of $t = 2m + f-1$ crash failures, $2m$ of them being $(n-k)$-constrained failures, and $(f-1)$ being any time failures. The second algorithm solves $(n + f)$-renaming in the presence of $t = m + f$ crash failures, m of them being $(n-t-1)$-constrained failures, and f being any time failures. It follows that the differentiation between λ-constrained crash failures and any time crash failures enlarges the space of executions in which the impossibility of k-set agreement and renaming in the presence of asynchrony and process crashes can be circumvented. In addition to its behavioral properties, both algorithms have a noteworthy first class property, namely, their simplicity.

This work was done in collaboration with Gadi Taubenfeld (IDC Herzliya).

6.4.4. *Anonymous obstruction-free (n, k)-set agreement with $n-k+1$ atomic read/write registers*

Participant: Michel Raynal.

The k-set agreement problem is a generalization of the consensus problem. Namely, assuming that each process proposes a value, every non-faulty process must decide one of the proposed values, under the constraint that at most k different values are decided. This is a hard problem in the sense that it cannot be solved in a pure read/write asynchronous system, in which k or more processes may crash. One way to sidestep this impossibility result consists in weakening the termination property, requiring only that a process decides if it executes alone during a long enough period of time. This is the well-known obstruction-freedom progress

condition. Consider a system of n anonymous asynchronous processes that communicate through atomic read/write registers, and such that any number of them may crash. In work [14] we address and solve the challenging open problem of designing an obstruction-free k-set agreement algorithm with only $(n-k+1)$ atomic registers. From a shared memory cost point of view, our algorithm is the best algorithm known to date, thereby establishing a new upper bound on the number of registers needed to solve this problem. For the consensus case ($k = 1$), the proposed algorithm is up to an additive factor of 1 close to the best known lower bound. Further, the paper extends this algorithm to obtain an x-obstruction-free solution to the k-set agreement problem that employs $(n-k+x)$ atomic registers, as well as a space-optimal solution for the repeated version of k-set agreement. Using this last extension, we prove that n registers are enough for every colorless task that is obstruction-free solvable with identifiers and any number of registers.

This work was done in collaboration with Zohir Bouzid and Pierre Sutra (CNRS).

ALPINES Project-Team

7. New Results

7.1. First kind Galerkin boundary element method for the Hodge-Laplacian in three dimensions

Boundary value problems for the Euclidean Hodge-Laplacian in three dimension $-\Delta_{HL} = \mathbf{curlcurl} - \mathbf{grad}\mathrm{div}$ lead to variational formulations set in subspaces of $\mathbf{H}(\mathbf{curl},\Omega) \cap \mathbf{H}(\mathrm{div},\Omega)$, $\Omega \subset \mathbb{R}^3$ a bounded Lipschitz domain. Via a representation formula and Calderón identities we derive corresponding first-kind boundary integral equations set in trace spaces of $H^1(\Omega)$, $\mathbf{H}(\mathbf{curl},\Omega)$, and $\mathbf{H}(\mathrm{div},\Omega)$. They give rise to saddle-point variational formulations and feature kernels whose dimensions are linked to fundamental topological invariants of Ω.

Kernels of the same dimensions also arise for the linear systems generated by low-order conforming Galerkin boundary element (BE) discretization. On their complements, we can prove stability of the discretized problems, nevertheless. We prove that discretization does not affect the dimensions of the kernels and also illustrate this fact by numerical tests.

7.2. Boundary integral multi-trace formulations and Optimised Schwarz Methods

In the present contribution, we consider Helmholtz equation with material coefficients being constant in each subdomain of a geometric partition of the propagation medium (discarding the presence of junctions), and we are interested in the numerical solution of such a problem by means of local multi-trace boundary integral formulations (local-MTF). For a one dimensional problem and configurations with two subdomains, it has been recently established that applying a Jacobi iterative solver to local-MTF is exactly equivalent to an Optimised Schwarz Method (OSM) with a non-local impendance. In the present contribution, we show that this correspondance still holds in the case where the subdomain partition involves an arbitrary number of subdomains. From this, we deduce that the depth of the adjacency graph of the subdomain partition plays a critical role in the convergence of linear solvers applied to local-MTF: we prove it for the case of homogeneous propagation medium and show, through numerical evidences, that this conclusion still holds for heterogeneous media. Our study also shows that, considering variants of local-MTF involving a relaxation parameter, there is a fixed value of this relaxation parameter that systematically leads to optimal speed of convergence for linear solvers.

7.3. Poroelasticity

In [38], we design and study a fully coupled numerical scheme for the poroelasticity problem modelled through Biot's equations. The classical way to numerically solve this system is to use a finite element method for the mechanical equilibrium equation and a finite volume method for the fluid mass conservation equation. However, to capture specific properties of underground media such as heterogeneities, discontinuities and faults, meshing procedures commonly lead to badly shaped cells for finite element based modelling. Consequently, we investigate the use of the recent virtual element method which appears as a potential discretization method for the mechanical part and could therefore allow the use of a unique mesh for the both mechanical and fluid flow modelling. Starting from a first insight into virtual element method applied to the elastic problem in the context of geomechanical simulations, we apply in addition a finite volume method to take care of the fluid conservation equation. We focus on the first order virtual element method and the two point flux approximation for the finite volume part. A mathematical analysis of this original coupled scheme is provided, including existence and uniqueness results and a priori estimates. The method is then illustrated by some computations on two or three dimensional grids inspired by realistic application cases.

7.4. Hybrid discontinuous Galerkin discretisation and domain decomposition preconditioners for the Stokes problem

Solving the Stokes equation by an optimal domain decomposition method derived algebraically involves the use of nonstandard interface conditions whose discretisation is not trivial. For this reason the use of approximation methods such as hybrid discontinuous Galerkin appears as an appropriate strategy: on the one hand they provide the best compromise in terms of the number of degrees of freedom in between standard continuous and discontinuous Galerkin methods, and on the other hand the degrees of freedom used in the nonstandard interface conditions are naturally defined at the boundary between elements. In this paper, we introduce the coupling between a well chosen discretisation method (hybrid discontinuous Galerkin) and a novel and efficient domain decomposition method to solve the Stokes system. We present the detailed analysis of the hybrid discontinuous Galerkin method for the Stokes problem with non standard boundary conditions. This analysis is supported by numerical evidence. In addition, the advantage of the new preconditioners over more classical choices is also supported by numerical experiments. The full paper [18] is available at https://hal.archives-ouvertes.fr/hal-01967577

7.5. A class of efficient locally constructed preconditioners based on coarse spaces

In [14] 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. This splitting leads to construct algebraically and locally a class of efficient coarse spaces which bound the spectral condition number of the preconditioned matrix by a number defined a priori. We also introduce the notion of filtering subspace. This concept helps compare the dimension minimality of coarse spaces. Some PDEs-dependant preconditioners correspond to a special case. The examples of the algebraic coarse spaces in this paper are not practical due to expensive construction. We propose a heuristic approximation that is not costly. Numerical experiments illustrate the efficiency of the proposed method.

7.6. Enlarged Krylov methods for reducing communication

Krylov methods are widely used for solving large sparse linear systems of equations. On distributed architectures, their performance is limited by the communication needed at each iteration of the algorithm. In [34], we study the use of so-called enlarged Krylov subspaces for reducing the number of iterations, and therefore the overall communication, of Krylov methods. In particular, we consider a reformulation of the Conjugate Gradient method using these enlarged Krylov subspaces: the enlarged Conjugate Gradient method. We present the parallel design of two variants of the enlarged Conjugate Gradient method as well as their corresponding dynamic versions where the number of search directions is dynamically reduced during the iterations. For a linear elasticity problem with heterogeneous coefficients using a block Jacobi preconditioner, we show that this implementation scales up to 16,384 cores, and is up to 6,9 times faster than the PETSc implementation of PCG.

In [15] we propose a variant of the GMRES method for solving linear systems of equations with one or multiple right-hand sides. Our method is based on the idea of the enlarged Krylov subspace to reduce communication. It can be interpreted as a block GMRES method. Hence, we are interested in detecting inexact breakdowns. We introduce a strategy to perform the test of detection. Furthermore, we propose an eigenvalues deflation technique aiming to have two benefits. The first advantage is to avoid the plateau of convergence after the end of a cycle in the restarted version. The second is to have a very fast convergence when solving the same system with different right-hand sides, each given at a different time (useful in the context of CPR preconditioner). With the same memory cost, we obtain a saving of up to 50% in the number of iterations to reach convergence with respect to the original method.

7.7. Recycling Krylov subspaces and reducing deflation subspaces for solving a sequence of linear systems

In [32] we present deflation strategies related to recycling Krylov subspace methods for solving one or a sequence of linear systems of equations. Besides well-known strategies of deflation, Ritz and harmonic Ritz based deflation, we introduce an SVD-based deflation technique. We consider the recycling in two contexts, recycling the Krylov subspace between the cycles of restarts and recycling a deflation subspace when the matrix changes in a sequence of linear systems. Numerical experiments on real-life reservoir simulations demonstrate the impact of our proposed strategy.

7.8. Solving linear equations with messenger-field and conjugate gradient techniques: an application to CMB data analysis

In [26] we discuss linear system solvers invoking a messenger-field and compare them with (preconditioned) conjugate gradients approaches. We show that the messenger-field techniques correspond to fixed point iterations of an appropriately preconditioned initial system of linear equations. We then argue that a conjugate gradient solver applied to the same preconditioned system, or equivalently a preconditioned conjugate gradient solver using the same preconditioner and applied to the original system, will in general ensure at least a comparable and typically better performance in terms of the number of iterations to convergence and time-to-solution. We illustrate our conclusions on two common examples drawn from the Cosmic Microwave Background data analysis: Wiener filtering and map-making. In addition, and contrary to the standard lore in the CMB field, we show that the performance of the preconditioned conjugate gradient solver can depend importantly on the starting vector. This observation seems of particular importance in the cases of map-making of high signal-to-noise sky maps and therefore should be of relevance for the next generation of CMB experiments.

7.9. Low rank approximation of a sparse matrix based on LU factorization with column and row tournament pivoting

In [23] we present an algorithm for computing a low rank approximation of a sparse matrix based on a truncated LU factorization with column and row permutations. We present various approaches for determining the column and row permutations that show a trade-off between speed versus deterministic/probabilistic accuracy. We show that if the permutations are chosen by using tournament pivoting based on QR factorization, then the obtained truncated LU factorization with column/row tournament pivoting, LU_CRTP, satisfies bounds on the singular values which have similarities with the ones obtained by a communication avoiding rank revealing QR factorization. Experiments on challenging matrices show that LU_CRTP provides a good low rank approximation of the input matrix and it is less expensive than the rank revealing QR factorization in terms of computational and memory usage costs, while also minimizing the communication cost. We also compare the computational complexity of our algorithm with randomized algorithms and show that for sparse matrices and high enough but still modest accuracies, our approach is faster.

7.10. ALORA: affine low-rank approximations

In [17] we introduce the concept of affine low-rank approximation for an $m \times n$ matrix, consisting in fitting its columns into an affine subspace of dimension at most $k << \min(m, n)$. We show that the optimal affine approximation can be obtained by applying an orthogonal projection to the matrix before constructing its best approximation. Moreover, we present the algorithm ALORA that constructs an affine approximation by slightly modifying the application of any low-rank approximation method. We focus on approximations created with the classical QRCP and subspace iteration algorithms. For the former, we present a detailed analysis of the existing pivoting techniques and furthermore, we provide a bound for the error when an arbitrary pivoting technique is used. For the case of subspace iteration, we prove a result on the convergence of singular vectors, showing a bound that is in agreement with the one for convergence of singular values proved recently. Finally, we present numerical experiences using challenging matrices taken from different fields, showing good performance and validating the theoretical framework.

7.11. Linear-time CUR approximation of BEM matrices

In [33] we propose linear-time CUR approximation algorithms for admissible matrices obtained from the hierarchical form of Boundary Element matrices. We propose a new approach called geometric sampling to obtain indices of most significant rows and columns using information from the domains where the problem is posed. Our strategy is tailored to Boundary Element Methods (BEM) since it uses directly and explicitly the cluster tree containing information from the problem geometry. Our CUR algorithm has precision comparable with low-rank approximations created with the truncated QR factorization with column pivoting (QRCP) and the Adaptive Cross Approximation (ACA) with full pivoting, which are quadratic-cost methods. When compared to the well-known linear-time algorithm ACA with partial pivoting, we show that our algorithm improves, in general, the convergence error and overcomes some cases where ACA fails. We provide a general relative error bound for CUR approximations created with geometrical sampling. Finally, we evaluate the performance of our algorithms on traditional BEM problems defined over different geometries.

7.12. Fractional decomposition of matrices and parallel computing

In [40] we are interested in the design of parallel numerical schemes for linear systems. We give an effective solution to this problem in the following case: the matrix A of the linear system is the product of p nonsingular matrices A_i^m with specific shape: $A_i = I - h_i X$ for a fixed matrix X and real numbers h_i. Although having the special form, these matrices A_i arise frequently in the discretization of evolutionary Partial Differential Equations. The idea is to express A^{-1} as a linear combination of elementary matrices A_i^{-k}. Hence the solution of the linear system with matrix A is a linear combination of the solutions of linear systems with matrices A_i^k. These systems are solved simultaneously on different processors.

AVALON Project-Team

7. New Results

7.1. Energy Efficiency in HPC and Large Scale Distributed Systems

Participants: Laurent Lefèvre, Dorra Boughzala, Christian Perez, Issam Raïs, Mathilde Boutigny.

7.1.1. 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 [22], [25].

7.1.2. Automatic Energy Efficient HPC Programming: A Case Study

Energy consumption is one of the major challenges of modern datacenters and supercomputers. By applying Green Programming techniques, developers have to iteratively implement and test new versions of their software, thus evaluating the impact of each code version on their energy, power and performance objectives. This approach is manual and can be long, challenging and complicated, especially for High Performance Computing applications. In [24], we formally introduces the definition of the Code Version Variability (CVV) leverage and present a first approach to automate Green Programming (*i.e.,* CVV usage) by studying the specific use-case of an HPC stencil-based numerical code, used in production. This approach is based on the automatic generation of code versions thanks to a Domain Specific Language (DSL), and on the automatic choice of code version through a set of actors. Moreover, a real case study is introduced and evaluated though a set of benchmarks to show that several trade-offs are introduced by CVV 1. Finally, different kinds of production scenarios are evaluated through simulation to illustrate possible benefits of applying various actors on top of the CVV automation.

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

In the article [9], 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.

7.1.4. Energy Simulation of GPU based Infrastructures

Through the IPL Hac-Specis and the PhD of Dorra Boughzala we begin to explore the modeling and calibrating of energy consumption of GPU architectures. We use the SimGrid simulation framework for the integration and validation on large scale systems.

7.2. HPC Component Models and Runtimes

Participants: Thierry Gautier, Christian Perez, Jérôme Richard.

7.2.1. On the Impact of OpenMP Task Granularity

Tasks are a good support for composition. During the development of a high-level component model for HPC, we have experimented to manage parallelism from components using OpenMP tasks. Since version 4-0, the standard proposes a model with dependent tasks that seems very attractive because it enables the description of dependencies between tasks generated by different components without breaking maintainability constraints such as separation of concerns. In [20], we present our feedback on using OpenMP in our context. We discover that our main issues are a too coarse task granularity for our expected performance on classical OpenMP runtimes, and a harmful task throttling heuristic counter-productive for our applications. We present a completion time breakdown of task management in the Intel OpenMP runtime and propose extensions evaluated on a testbed application coming from the Gysela application in plasma physics.

7.2.2. Building and Auto-Tuning Computing Kernels: Experimenting with BOAST and StarPU in the GYSELA Code

Modeling turbulent transport is a major goal in order to predict confinement performance in a tokamak plasma. The gyrokinetic framework considers a computational domain in five dimensions to look at kinetic issues in a plasma; this leads to huge computational needs. Therefore, optimization of the code is an especially important aspect, especially since coprocessors and complex manycore architectures are foreseen as building blocks for Exascale systems. This project [6] aims to evaluate the applicability of two auto-tuning approaches with the BOAST and StarPU tools on the gysela code in order to circumvent performance portability issues. A specific computation intensive kernel is considered in order to evaluate the benefit of these methods. StarPU enables to match the performance and even sometimes outperform the hand-optimized version of the code while leaving scheduling choices to an automated process. BOAST on the other hand reveals to be well suited to get a gain in terms of execution time on four architectures. Speedups in-between 1.9 and 5.7 are obtained on a cornerstone computation intensive kernel.

7.3. Modeling and Simulation of Parallel Applications and Distributed Infrastructures

Participants: Eddy Caron, Zeina Houmani, Frédéric Suter.

7.3.1. SMPI Courseware: Teaching Distributed-Memory Computing with MPI in Simulation

It is typical in High Performance Computing (HPC) courses to give students access to HPC platforms so that they can benefit from hands-on learning opportunities. Using such platforms, however, comes with logistical and pedagogical challenges. For instance, a logistical challenge is that access to representative platforms must be granted to students, which can be difficult for some institutions or course modalities; and a pedagogical challenge is that hands-on learning opportunities are constrained by the configurations of these platforms. A way to address these challenges is to instead simulate program executions on arbitrary HPC platform configurations. In [15] we focus on simulation in the specific context of distributed-memory computing and MPI programming education. While using simulation in this context has been explored in previous works, our approach offers two crucial advantages. First, students write standard MPI programs and can both debug and analyze the performance of their programs in simulation mode. Second, large-scale executions can be simulated in short amounts of time on a single standard laptop computer. This is possible thanks to SMPI, an MPI simulator provided as part of SimGrid. After detailing the challenges involved when using HPC platforms for HPC education and providing background information about SMPI, we present SMPI Courseware. SMPI Courseware is a set of in-simulation assignments that can be incorporated into HPC courses to provide students with hands-on experience for distributed-memory computing and MPI programming learning objectives. We describe some these assignments, highlighting how simulation with SMPI enhances the student learning experience.

7.3.2. Evaluation through Realistic Simulations of File Replication Strategies for Large Heterogeneous Distributed Systems

File replication is widely used to reduce file transfer times and improve data availability in large distributed systems. Replication techniques are often evaluated through simulations, however, most simulation platform models are oversimplified, which questions the applicability of the findings to real systems. In [17], we investigate how platform models influence the performance of file replication strategies on large heterogeneous distributed systems, based on common existing techniques such as prestaging and dynamic replication. The novelty of our study resides in our evaluation using a realistic simulator. We consider two platform models: a simple hierarchical model and a detailed model built from execution traces. Our results show that conclusions depend on the modeling of the platform and its capacity to capture the characteristics of the targeted production infrastructure. We also derive recommendations for the implementation of an optimized data management strategy in a scientific gateway for medical image analysis.

7.3.3. WRENCH: Workflow Management System Simulation Workbench

Scientific workflows are used routinely in numerous scientific domains, and Workflow Management Systems (WMSs) have been developed to orchestrate and optimize workflow executions on distributed platforms. WMSs are complex software systems that interact with complex software infrastructures. Most WMS research and development activities rely on empirical experiments conducted with full-fledged software stacks on actual hardware platforms. Such experiments, however, are limited to hardware and software infrastructures at hand and can be labor- and/or time-intensive. As a result, relying solely on real-world experiments impedes WMS research and development. An alternative is to conduct experiments in simulation.

In [16] we presented WRENCH, a WMS simulation framework, whose objectives are (i) accurate and scalable simulations; and (ii) easy simulation software development. WRENCH achieves its first objective by building on the SimGrid framework. While SimGrid is recognized for the accuracy and scalability of its simulation models, it only provides low-level simulation abstractions and thus large software development efforts are required when implementing simulators of complex systems. WRENCH thus achieves its second objective by providing high- level and directly re-usable simulation abstractions on top of SimGrid. After describing and giving rationales for WRENCH's software architecture and APIs, we present a case study in which we apply WRENCH to simulate the Pegasus production WMS. We report on ease of implementation, simulation accuracy, and simulation scalability so as to determine to which extent WRENCH achieves its two above objectives. We also draw both qualitative and quantitative comparisons with a previously proposed workflow simulator.

7.3.4. A Microservices Architectures for Data-Driven Service Discovery

Usual microservices discovery mechanisms are normally based on a specific user need *(Goal-based approaches)*. However, in today's evolving architectures, users need to discover what features they can take advantage of before looking for the available microservices. In collaboration with RDI2 (Rutgers University) we developped a data-driven microservices architecture that allows users to discover, from specific objects, the features that can be exerted on these objects as well as all the microservices dedicated to them [28]. This architecture, based on the main components of the usual microservices architectures, adopts a particular communication strategy between clients and registry to achieve the goal. This article contains a representation of a microservice data model and a P2P model that transforms our architecture into a robust and scalable system. Also, we designed a prototype to validate our approach using Istio library.

7.4. Cloud Resource Management

Participants: Eddy Caron, Hadrien Croubois, Jad Darrous, Christian Perez.

7.4.1. Nitro: Network-Aware Virtual Machine Image Management in Geo-Distributed Clouds

Recently, most large cloud providers, like Amazon and Microsoft, replicate their Virtual Machine Images (VMIs) on multiple geographically distributed data centers to offer fast service provisioning. Provisioning a service may require to transfer a VMI over the wide-area network (WAN) and therefore is dictated by the distribution of VMIs and the network bandwidth in-between sites. Nevertheless, existing methods to facilitate VMI management (*i.e.,* retrieving VMIs) overlook network heterogeneity in geo-distributed clouds. In [19], we design, implement and evaluate Nitro, a novel VMI management system that helps to minimize the transfer time of VMIs over a heterogeneous WAN. To achieve this goal, Nitro incorporates two complementary features. First, it makes use of deduplication to reduce the amount of data which will be transferred due to the high similarities within an image and in-between images. Second, Nitro is equipped with a network-aware data transfer strategy to effectively exploit links with high bandwidth when acquiring data and thus expedites the provisioning time. Experimental results show that our network-aware data transfer strategy offers the optimal solution when acquiring VMIs while introducing minimal overhead. Moreover, Nitro outperforms state-of-the-art VMI storage systems (*e.g.,* OpenStack Swift) by up to 77%.

7.4.2. Toward an Autonomic Engine for Scientific Workflows and Elastic Cloud Infrastructure

The constant development of scientific and industrial computation infrastructures requires the concurrent development of scheduling and deployment mechanisms to manage such infrastructures. Throughout the last decade, the emergence of the Cloud paradigm raised many hopes, but achieving full platform autonomicity is still an ongoing challenge. We built a workflow engine that integrated the logic needed to manage workflow execution and Cloud deployment on its own. More precisely, we focus on Cloud solutions with a dedicated Data as a Service (DaaS) data management component. Our objective was to automate the execution of workflows submitted by many users on elastic Cloud resources. This contribution proposes a modular middleware infrastructure and details the implementation of the underlying modules:

- A workflow clustering algorithm that optimises data locality in the context of DaaS-centered communications;
- A dynamic scheduler that executes clustered workflows on Cloud resources;
- A deployment manager that handles the allocation and deallocation of Cloud resources according to the workload characteristics and users' requirements.

All these modules have been implemented in a simulator to analyse their behaviour and measure their effectiveness when running both synthetic and real scientific workflows. We also implemented these modules in the Diet middleware to give it new features and prove the versatility of this approach. Simulation running the WASABI workflow (waves analysis based inference, a framework for the reconstruction of gene regulatory networks) showed that our approach can decrease the deployment cost by up to 44% while meeting the required deadlines [13].

7.4.3. Madeus: A Formal Deployment Model

Distributed software architecture is composed of multiple interacting modules, or components. Deploying such software consists in installing them on a given infrastructure and leading them to a functional state. However, since each module has its own life cycle and might have various dependencies with other modules, deploying such software is a very tedious task, particularly on massively distributed and heterogeneous infrastructures. To address this problem, many solutions have been designed to automate the deployment process. In [18], we introduce Madeus, a component-based deployment model for complex distributed software. Madeus accurately describes the life cycle of each component by a Petri net structure, and is able to finely express the dependencies between components. The overall dependency graph it produces is then used to reduce deployment time by parallelizing deployment actions. While this increases the precision and performance of the model, it also increases its complexity. For this reason, the operational semantics need to be clearly defined to prove results such as the termination of a deployment. In this paper, we formally describe the operational semantics of Madeus, and show how it can be used in a use-case: the deployment of a real and large distributed software (*i.e.,* , OpenStack).

In [18], we have proposed an extension based on component behavioral interfaces to the Aeolus component model to better separate the concerns of component users (*e.g.,* application architect) from component developers.

7.5. Data Stream Processing on Edge Computing

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

7.5.1. Latency-Aware Placement of Data Stream Analytics on Edge Computing

The interest in processing data events under stringent time constraints as they arrive has led to the emergence of architecture and engines for data stream processing. Edge computing, initially designed to minimize the latency of content delivered to mobile devices, can be used for executing certain stream processing operations. Moving opera- tors from cloud to edge, however, is challenging as operator-placement decisions must consider the application requirements and the network capabilities. We introduce strategies to create placement configurations for data stream processing applications whose operator topologies follow series parallel graphs[35]. We consider the operator characteristics and requirements to improve the response time of such appli- cations. Results show that our strategies can improve the response time in up to 50% for application graphs comprising multiple forks and joins while transferring less data and better using the resources.

7.5.2. Estimating Throughput of Stream Processing Applications in FoG Computing

Recent trends exploit decentralized infrastructures (e.g.. Fog computing) to deploy DSP (Data Stream Processing) applications and leverage the computational power. Fog computing overlaps some features of Cloud computing and includes others, for instance, location awareness. The operator placement problem consists of determining, within a set of distributed computing resources, the computing resources that should host and execute each operator of the DSP application, with the goal of optimizing QoS requirements of the application. The QoS requirements of the application refer to processing time, costs, throughput, etc. We propose a model to estimate the application throughput at each layer of Fog computing (Devices, Edge and Cloud) by considering a given placement solution. The estimated throughput provides a useful insight to determine the amount of physical resources to meet the QoS requirements. The model allows to identify the application bottleneck, when facing data rate variations, and provides information to self-scale in or out the DSP application.

DATAMOVE Project-Team

6. New Results

6.1. Integration of High Performance Computing and Data Analytics

6.1.1. I/O Survey

First contribution is a comprehensive survey on parallel I/O in the HPC context [14]. As the available processing power and amount of data increase, I/O remains a central issue for the scientific community. This survey focuses on a traditional I/O stack, with a POSIX parallel file system. Through the comprehensive study of publications from the most important conferences and journals in a five-year time window, we discuss the state of the art of I/O optimization approaches, access pattern extraction techniques, and performance modeling, in addition to general aspects of parallel I/O research. This survey enables us to identify the general characteristics of the field and the main current and future research topics.

6.1.2. Task Based In Situ Processing

One approach to bypass the I/O bottleneck is *in situ* processing, an important research topic at DataMove. The *in situ* paradigm proposes to reduce data movement and to analyze data while still resident in the memory of the compute node by co-locating simulation and analytics on the same compute node. The simplest approach consists in modifying the simulation timeloop to directly call analytics routines. However, several works have shown that an *asynchronous* approach where analytics and simulation run concurrently can lead to a significantly better performance. Today, the most efficient approach consists in running the analytics processes on a set of dedicated cores, called helper cores, to isolate them from the simulation processes. Simulation and analytics thus run concurrently on different cores but this static isolation can lead to underused resources if the simulation or the analytics do not fully use all the assigned cores.

In this work performed in collaboration with CEA, we developed TINS, a task-based in situ framework that implements a novel *dynamic helper core* strategy. TINS relies on a work stealing scheduler and on task-based programming. Simulation and analytics tasks are created concurrently and scheduled on a set of worker threads created by a single instance of the work stealing scheduler. Helper cores are assigned dynamically: some worker threads are dedicated to analytics when analytics tasks are available while they join the other threads for processing simulation tasks otherwise, leading to a better resource usage. We leverage the good compositionality properties of task-based programming to seamlessly keep the analytics and simulation codes well separated and a plugin system enables to develop parallel analytics codes outside of the simulation code.

TINS is implemented with the Intel Threading Building Blocks (TBB) library that provides a task-based programming model and a work stealing scheduler. The experiments are conducted with the hybrid MPI+TBB ExaStamp molecular dynamics code that we associate with a set of analytics representative of computational physics algorithms. We show up to 40% performance improvement over various other approaches, including the standard helper core, on experiments on up to 14,336 Broadwell cores.

6.1.3. Stream Processing

Stream processing is the Big Data equivalent of in situ processing. It consists in analyzing on-line incoming streams of data, often produced from sensors or social networks like Twitter. We investigated the convergence between both paradigms through different directions: how the programming environment developed specifically for stream processing can applied to the data produced by large parallel simulations [18]; Proposing a dynamics data structure to keep sorted data streams [12]; Evaluating the performance of the FlameMR framework on data produced from a parallel simulation[13]. We summarize here the 2 first contributions.

6.1.3.1. Packed Memory QuadTree.

Over the past years, several in-memory big-data management systems have appeared in academia and industry. In-memory databases systems avoid the overheads related to traditional I/O disk-based systems and have made possible to perform interactive data-analysis over large amounts of data. A vast literature of systems and research strategies deals with different aspects, such as the limited storage size and a multi-level memory-hierarchy of caches. Maintaining the right data layout that favors locality of accesses is a determinant factor for the performance of in-memory processing systems. Stream processing engines like Spark or Flink support the concept of *window*, which collects the latest events without a specific data organization. It is possible to trigger the analysis upon the occurrence of a given criterion (time, volume, specific event occurrence). After a window is updated, the system shifts the processing to the next batch of events. There is a need to go one step further to keep a live window continuously updated while having a fine grain data replacement policy to control the memory footprint. The challenge is the design of dynamic data structures to absorb high rate data streams, stash away the oldest data to stay in the allowed memory budget while enabling fast queries executions to update visual representations. A possible solution is the extension of database structures like R-trees used in SpatiaLite or PostGis, or to develop dedicated frameworks like Kit based on a pyramid structure.

We developed a novel self-organized cache-oblivious data structure, called PMQ, for in-memory storage and indexing of fixed length records tagged with a spatiotemporal index. We store the data in an array with a controlled density of gaps (*i.e.*, empty slots) that benefits from the properties of the *Packed Memory Arrays*. The empty slots guarantee that insertions can be performed with a low amortized number of data movements ($O(\log^2(N))$) while enabling efficient spatiotemporal queries. During insertions, we rebalance parts of the array when required to respect density constraints, and the oldest data is stashed away when reaching the memory budget. To spatially subdivide the data, we sort the records according to their Morton index, thus ensuring spatial locality in the array while defining an implicit, recursive quadtree, which leads to efficient spatiotemporal queries. We validate PMQ for consuming a stream of tweets to answer visual and range queries. PMQ significantly outperforms the widely adopted spatial indexing data structure R-tree, typically used by relational databases, as well as the conjunction of Geohash and B^+-tree, typically used by NoSQL databases.

6.1.3.2. Flink based in situ Processing.

We proposed to leverage Apache Flink, a scalable stream processing engine from the Big Data domain, in this HPC context. Flink enables to program analyses within a simple window based map/reduce model, while the runtime takes care of the deployment, load balancing and fault tolerance. We build a complete in transit analytics workflow, connecting an MD simulation to Apache Flink and to a distributed database, Apache HBase, to persist all the desired data. To demonstrate the expressivity of this programming model and its suitability for HPC scientific environments, two common analytics in the Molecular Dynamics field have been implemented. We assessed the performance of this framework, concluding that it can handle simulations of sizes used in the literature while providing an effective and versatile tool for scientists to easily incorporate on-line parallel analytics in their current workflows.

6.2. Data Aware Batch Scheduling

6.2.1. Batch Scheduling for Energy

The project COSMIC [24], [22], [16], [17], in collaboration with Myriads team in Inria Rennes-Atlantique, targets the optimization of green energy usage in Clouds. The project considers a geographically distributed cloud, with each data center associated with a local photovoltaic (PV) farm. The objective is to maximize the photovoltaic energy by allocation the computing workload to the data centers according to its energy production. The production forecasting is modeled with a truncated normal law, permitting to consider the uncertainty of the forecast.

Chapter [24] considers a simple model with homogeneous Virtual Machines submitted at unpredictable rate. This study has resulted in a scheduling algorithm for task allocation. The chapter demonstrates the optimality of this algorithm at current time slot according to production forecast parameters.

Paper [22] extends these results to heterogeneous VM. Each VM is defined by its arrival date, its execution time, its memory requirement and its CPU usage. In this model, due to execution time durations, the possibility to migrate running VM was considered. An algorithm is detailed in the paper that is compared to standard algorithm through simulations.

A third study [16], [17] has carefully modeled the interactions between the Cloud and the energy supplier. Due to variability of PV production and workload submission, each data center will alternatively inject energy into the electricity grid or purchase energy. The energy model considers a virtual energy pool mitigating the surplus and deficit of the different data center, with reduced costs regarding the difference between electricity cost and electricity injection tariff. The algorithm detailed in this paper outperforms well-known round-robin approaches, as shown by simulations.

6.2.2. Learning Methods for Batch Scheduling

Most of Job Scheduling algorithms apply greedy tasks ordering, as First Come First Served (FCFS) or Shortest Processing time First (SPF). They give simple methods, highly practical with certain guarantees. They are however far from optimal. Mixed methods, combining many of this basic methods permit to improve their performance. DataMove has developed [27] a learning method permitting to adapt the Mixed method to benchmarks. An extensive experimental campaign has permitted to determine the possibilities of basic and mixed methods according to the benchmarks characteristics, enhancing the efficiency of mixed methods.

6.2.3. Reproducibility

Related to batch scheduling experimentation, DataMove has led investigations on reproducibility [23]. Existing approaches focus on repeatability, but this is only the first step to reproducibility: Continuing a scientific work from a previous experiment requires to be able to modify it. This ability is called reproducibility with Variation. We show that capturing the environment of execution is necessary but not sufficient ; we also need the environment of development. The variation also implies that those environments are subject to evolution, so the whole software development lifecycle needs to be considered. To take into account these evolutions, software environments need to be clearly defined, reconstructible with variation, and easy to share. In this context, we propose new way of seeing reproducibility through the scientific software development lifecycle. Each step in this lifecycle requires a software environment. We define a software environment by a set of applications and libraries, with all their dependencies, and their configurations, required to achieve a step in a scientific workflow.

6.2.4. Online Algorithms

Rob van Stee wrote a review of 2018 online algorithms including our recent contributions on resource augmentation [0] We quote him here:

Progress was also made on scheduling to minimize weighted flow time on unrelated machines. In ESA 2016, Giorgio Lucarelli et al. [1] had considered a version where the online algorithm can reject some $\varepsilon_r > 0$ fraction (by weight) of the jobs and have machines that are $1 + \varepsilon_s$ as fast as the offline machines, for some $\varepsilon_s > 0$. They showed that this is already enough to achieve a competitive ratio of $O(1/(\varepsilon_s\varepsilon_r))$.

In SPAA 2018, Giorgio Lucarelli et al.[20] (a superset of the previous authors) showed that it is in fact sufficient to reject a 2ε fraction of the total number of jobs to achieve a competitive ratio of $2(\frac{1+\varepsilon}{\varepsilon})$ for minimizing the total flow time. This algorithm sometimes rejects a job other than the one that has just arrived. The authors show that this is necessary, as otherwise there is a lower bound of $\Omega(\Delta)$ even on a single machine. Here Δ is the size ratio (the ratio of largest to smallest job size). (Obviously this lower bound also holds if you cannot reject jobs at all.)

They also consider the speed scaling model, in which machines can be sped up if additional energy is invested, and the goal is to minimize the total weighted flow time plus energy usage. If the power function of machine i is given by $P(s_i(t)) = s_i(t)^\alpha$, where $s_i(t)$ is the current speed of machine i, there is an algorithm which is $O((1 + 1/\varepsilon)^{\alpha/(\alpha-1)})$-competitive that rejects jobs of total weight at most a fraction ε of the total weight of all

[0] Rob van Stee. 2018. SIGACT News Online Algorithms Column 34: 2018 in review. SIGACT News 49, 4 (December 2018), 36-45.

the jobs. They also give a positive result for jobs with hard deadlines, where the goal is to minimize the total energy usage and no job may be rejected.

In ESA 2018, the same set of authors [11] improved/generalized these results by showing that rejection alone is sufficient for an algorithm to be competitive even for weighted flow time. They presented an $O(1/\varepsilon^3)$-competitive algorithm that rejects at most $O(\varepsilon)$ of the total weight of the jobs. In this algorithm, jobs are assigned (approximately) greedily to machines, and each machine runs the jobs assigned to it using Highest Density First. A job may be rejected if it is running while much heavier jobs arrive or if it is in the queue while very many jobs arrive. The second rule simulates the resource augmentation on the speed.

HIEPACS Project-Team

7. New Results

7.1. High-performance computing on next generation architectures

7.1.1. Evaluation of dataflow programming models for electronic structure theory

Dataflow programming models have been growing in popularity as a means to de- liver a good balance between performance and portability in the post-petascale era. In this paper we evaluate different dataflow programming models for electronic struc- ture methods and compare them in terms of programmability, resource utilization, and scalability. In particular, we evaluate two programming paradigms for expressing scientific applications in a dataflow form: (1) explicit dataflow, where the dataflow is specified explicitly by the developer, and (2) implicit dataflow, where a task schedul- ing runtime derives the dataflow using per-task data-access information embedded in a serial program. We discuss our findings and present a thorough experimental analysis using methods from the NWChem quantum chemistry application as our case study, and OpenMP, StarpU and ParSEC as the task-based runtimes that enable the different forms of dataflow execution. Furthermore, we derive an abstract model to explore the limits of the different dataflow programming paradigms.

More information on these results can be found in [8]

7.1.2. On soft errors in the Conjugate Gradient method: sensitivity and robust numerical detection

The conjugate gradient (CG) method is the most widely used iterative scheme for the solution of large sparse systems of linear equations when the matrix is symmetric positive definite. Although more than sixty year old, it is still a serious candidate for extreme-scale computation on large computing platforms. On the technological side, the continuous shrinking of transistor geometry and the increasing complexity of these devices affect dramatically their sensitivity to natural radiation, and thus diminish their reliability. One of the most common effects produced by natural radiation is the single event upset which consists in a bit-flip in a memory cell producing unexpected results at application level. Consequently, the future computing facilities at extreme scale might be more prone to errors of any kind including bit-flip during calculation. These numerical and technological observations are the main motivations for this work, where we first investigate through extensive numerical experiments the sensitivity of CG to bit-flips in its main computationally intensive kernels, namely the matrix-vector product and the preconditioner application. We further propose numerical criteria to detect the occurrence of such faults; we assess their robustness through extensive numerical experiments.

More information on these results can be found in [16].

7.1.3. Energy analysis of a solver stack for frequency-domain electromagnetics

High-performance computing (HPC) 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 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.

More information on these results can be found in [18].

7.1.4. A compiler front-end for OpenMP's variants

OpenMP 5.0 introduced the concept of *variant*: a directive which can be used to indicate that a function is a variant of another existing *base function*, in a specific context (eg: `foo_gpu_nvidia` could be declared as a variant of `foo`, but only when executing on specific NVIDIA hardware).

In the context of PRACE-5IP, in collaboration with the Inria STORM team, we want to leverage this construct to be able to take advantage of the StarPU heterogeneous scheduler through the interoperability layer between OpenMP and StarPU. We started this work by implementing the necessary changes in the Clang front-end to support OpenMP's *variant*. We have assessed this support in the `Chameleon` dense linear algebra package. Indeed, `Chameleon` relies on sequential task-based algorithms where sub-tasks of the overall algorithms are submitted to a runtime system. Additionally to the `quark`, `PaRSEC` and `StarPU` support, we have implemented an OpenMP support in `Chameleon`. The originality of the proposed approach is that this OpenMP support can either rely on a native OpenMP runtime system or indirectly use the above mentioned OpenMP-StarPU back-end. We are currently assessing the approach on multicore homogeneous machines, the next step being heterogeneous architectures.

7.2. High performance solvers for large linear algebra problems

7.2.1. Partitioning and communication strategies for sparse non-negative matrix factorization

Non-negative matrix factorization (NMF), the problem of finding two non-negative low-rank factors whose product approximates an input matrix, is a useful tool for many data mining and scientific applications such as topic modeling in text mining and blind source separation in microscopy. In this paper, we focus on scaling algorithms for NMF to very large sparse datasets and massively parallel machines by employing effective algorithms, communication patterns, and partitioning schemes that leverage the sparsity of the input matrix. In the case of machine learning workflow, the computations after SpMM must deal with dense matrices, as Sparse-Dense matrix multiplication will result in a dense matrix. Hence, the partitioning strategy considering only SpMM will result in a huge imbalance in the overall workflow especially on computations after SpMM and in this specific case of NMF on non-negative least squares computations. Towards this, we consider two previous works developed for related problems, one that uses a fine-grained partitioning strategy using a point-to-point communication pattern and on that uses a checkerboard partitioning strategy using a collective-based communication pattern. We show that a combination of the previous approaches balances the demands of the various computations within NMF algorithms and achieves high efficiency and scalability. From the experiments, we could see that our proposed algorithm communicates atleast 4x less than the collective and achieves upto 100x speed up over the baseline FAUN on real world datasets. Our algorithm was experimented in two different super computing platforms and we could scale up to 32000 processors on Bluegene/Q.

More information on these results can be found in [21].

7.2.2. Low-rank factorizations in data sparse hierarchical algorithms for preconditioning Symmetric positive definite matrices

We consider the problem of choosing low-rank factorizations in data sparse ma- trix approximations for preconditioning large scale symmetric positive definite matrices. These approximations are memory efficient schemes that rely on hierarchical matrix partitioning and compression of certain sub-blocks of the matrix. Typically, these matrix approximations can be constructed very fast, and their matrix product can be applied rapidly as well. The common prac- tice is to express the compressed sub-blocks by low-rank factorizations, and the main contribution of this work is the numerical and spectral analysis of SPD preconditioning schemes represented by $2\times$ 2 block matrices, whose off-diagonal sub-blocks are low-rank approximations of the original matrix off-diagonal sub-blocks. We propose an optimal choice of low-rank approximations which minimizes the condition number of the preconditioned system, and demonstrate that the analysis can be applied to the class of hierarchically off-diagonal low-rank matrix approximations. Spec- tral estimates that take into account the error propagation through levels of the hierarchy which quantify the impact of the choice of low-rank compression on the global condition number are provided. The numerical results indicate that

the properties of the preconditioning scheme using proper low-rank compression are superior to employing standard choices for low-rank compression. A major goal of this work is to provide an insight into how proper reweighted prior to low-rank compression influences the condition number for a simple case, which would lead to an extended analysis for more general and more efficient hierarchical matrix approximation techniques.

More information on these results can be found in [5].

7.2.3. Analyzing the effect of local rounding error propagation on the maximal attainable accuracy of the pipelined Conjugate Gradient method

Pipelined Krylov subspace methods typically offer improved strong scaling on par- allel HPC hardware compared to standard Krylov subspace methods for large and sparse linear systems. In pipelined methods the traditional synchronization bottleneck is mitigated by overlap- ping time-consuming global communications with useful computations. However, to achieve this communication-hiding strategy, pipelined methods introduce additional recurrence relations for a number of auxiliary variables that are required to update the approximate solution. This paper aims at studying the influence of local rounding errors that are introduced by the additional recurrences in the pipelined Conjugate Gradient (CG) method. Specifically, we analyze the impact of local round-off effects on the attainable accuracy of the pipelined CG algorithm and compare it to the traditional CG method. Furthermore, we estimate the gap between the true residual and the recursively computed residual used in the algorithm. Based on this estimate we suggest an automated residual replacement strategy to reduce the loss of attainable accuracy on the final iterative solution. The resulting pipelined CG method with residual replacement improves the maximal attainable accuracy of pipelined CG while maintaining the efficient parallel performance of the pipelined method. This conclusion is substantiated by numerical results for a variety of benchmark problems.

More information on these results can be found in [7].

7.2.4. Sparse supernodal solver using block low-rank compression: Design, performance and analysis

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

These contributions have been published in International Journal of Computational Science and Engineering (JoCS) [9].

7.2.5. Supernodes ordering to enhance Block Low-Rank compression in a sparse direct solver

Solving sparse linear systems appears in many scientific applications, and sparse direct linear solvers are widely used for their robustness. Still, both time and memory complexities limit the use of direct methods to solve larger problems. In order to tackle this problem, low-rank compression techniques have been introduced in direct solvers to compress large dense blocks appearing in the symbolic factorization. In this paper, we consider the Block Low-Rank compression (BLR) format and adress the problem of clustering unknowns

that come from separators issued from the nested dissection process. We show that methods considering only intra-separators connectivity (i.e., k-way or recursive bisection) as well as methods managing only interaction between separators have some limitations. We propose a new strategy that considers interactions between a separator and its children to pre-select some interactions while reducing the number of off-diagonal blocks in the symbolic structure. We demonstrate how this new method enhances the BLR strategies in the sparse direct supernodal solver PaStiX.

These contributions have been submitted in SIAM Journal on Matrix Analysis and Applications (SIMAX) [22].

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

At the core of numerical simulations for scientific computing applications, one typically needs to solve an equation either in the form of a linear system (Ax = b) or an eigenvalue problem (Ax = x) to determine the course of the simulation. A major breakthrough in this solution step is exploiting the inherent low-rank structure in the problem; an idea stemming from the observation that particles in the same spatial locality exhibit similar interactions with others in a distant cluster/region. This property has been exploited in many contexts such as fast multipole methods (FMM) and hierarchical matrices (H-matrices) in applications ranging from n-body simulations to electromagnetics, which amount to numerically compressing the matrix in order to reduce computational and memory costs. Recent theory along this direction involves employing tensor decomposition to quantize the matrix in the form of a tensor (through logical restructuring/reshaping) and use tensor decomposition to approximate it with a controllable global error. Once the matrix and vectors are compressed this way, one can similarly use the compressed tensor to carry out matrix-vector operations with significantly better compression rate than the H-matrix approach.

Despite these major recent breakthroughs in the theory and application of tensor-based methods, addressing large-scale real-world problems with these methods requires immense computational power, which necessitates highly optimized parallel algorithms and implementations. To this end, we have initiated the development of a tensor-based linear system and eigenvalue solver library called Celeste++(C++ library for Efficient low-rank Linear and Eigenvalue Solvers using Tensor decomposition) providing a complete framework for expressing a problem in tensor form, then effectuating all matrix-vector operations under this compressed form with tremendous computational and memory efficiency. The fruits of our preliminary studies led two project submissions at the national scale (ANR JCJC and CNRS PEPS JCJC, currently under evaluation) and one Severo Ochoa Mobility Grant for a collaboration visit to Barcelona Supercomputing Center (BSC). We also supervised an internship on the application of tensor solvers in the context of electromagnetic applications with very promising results for future work.

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 by Maria Predari, 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. 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.

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). This year we have worked on three main areas. First, we extend our shared memory parallelization to distributed memory using the message exchange paradigm. This new version should allow us to process larger systems quickly. To target the eigenvalues relevant for chemists, i. eigenvalues with an intensity. 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 paper is being written on these last two points.

7.5.1.2. Dislocation

We have focused on the improvements of the parallel collision detection and the data structure in the OPTIDIS code [11].

- The new collision detection algorithm to reliably handle junction formation for Dislocation Dynamics using hybrid OpenMP + MPI parallelism has been developed. The enhanced precision and reliability of this new algorithm allows the use of larger time-steps for faster simulations. Hierarchical methods for collision detection, as well as hybrid parallelism are also used to improve performance;
- A new distributed data structure has been developed to enhance the reliability and modularity of OPTIDIS. The new data structure provides an interface to modify safely and reliably the distributed dislocation mesh in order to enforce data consistency across all computation nodes. This interface also improves code modularity allowing the study of data layout performance without modifying the algorithms.

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

7.5.2.1. A geometric view of biodiversity: scaling to metagenomics

We have designed a new efficient dimensionality reduction algorithm in order to investigate new ways of accurately characterizing the biodiversity, namely from a geometric point of view, scaling with large environmental sets produced by NGS ($\sim$105 sequences). The approach is based on Multidimensional Scaling (MDS) that allows for mapping items on a set of n points into a low dimensional euclidean space given the set of pairwise distances. We compute all pairwise distances between reads in a given sample, run MDS on the distance matrix, and analyze the projection on first axis, by visualization tools. We have circumvented the quadratic complexity of computing pairwise distances by implementing it on a hyperparallel computer (Turing, a Blue Gene Q), and the cubic complexity of the spectral decomposition by implementing a dense random projection based algorithm. We have applied this data analysis scheme on a set of 10^5 reads, which are amplicons of a diatom environmental sample from Lake Geneva. Analyzing the shape of the point cloud paves the way for a geometric analysis of biodiversity, and for accurately building OTUs (Operational Taxonomic Units), when the data set is too large for implementing unsupervised, hierarchical, high-dimensional clustering.

More information on these results can be found in [19].

7.5.2.2. High performance simulation for ITER tokamak

Concerning the GYSELA global non-linear electrostatic code, a critical problem is the design of a more efficient parallel gyro-average operator for the deployment of very large (future) GYSELA runs. The main unknown of the computation is a distribution function that represents either the density of the guiding centers, either the density of the particles in a tokamak. The switch between these two representations is done thanks to the gyro-average operator. In the previous version of GYSELA, the computation of this operator was achieved thanks to a Padé approximation. In order to improve the precision of the gyro-averaging, a new parallel version based on an Hermite interpolation has been done (in collaboration with the Inria TONUS project-team and IPP Garching). The integration of this new implementation of the gyro-average operator has been done in GYSELA and the parallel benchmarks have been successful. This work is carried on in the framework of the PhD of Nicolas Bouzat (funded by IPL C2S@Exa) co-advised with Michel Mehrenberger from TONUS project-team and in collaboration with Guillaume Latu from CEA-IRFM. The scientific objectives of this work are first to consolidate the parallel version of this gyro-average operator, in particular by designing a scalable MPI+OpenMP parallel version and by using a new communication scheme, and second to design new numerical methods for the gyro-average, source and collision operators to deal with new physics in GYSELA. The objective is to tackle kinetic electron configurations for more realistic complex large simulations. This has been done by using a new data distribution for a new irregular mesh in order to take into account the complex geometries of modern tokamak reactors. All these contributions have been validated on a new object-oriented proptotype of GYSELA which uses a task based programming model. The PhD thesis of Nicolas Bouzat has been defended on December 17, 2018.

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

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

Numerically scalable hybrid solvers based on a fully algebraic coarse space correction have been theoretically studied within the PhD thesis of Louis Poirel defended on November 28, 2018. Some of the proposed numerical schemes have been integrated in the MaPHyS parallel package. In particular, multiple parallel strategies have been designed and their parallel efficiencies were assessed in two large application codes. The first one is Alya developed at BSC, that is a high performance computational mechanics code to solve coupled multi-physics / multi-scale problems, which are mostly coming from engineering applications. This activity was carried out in the framework of the EoCoE project. Thye second large code is AVIP jointly developed by CERFACS and Laboratoire de Physique des Plasmas at E´cole Polytechnique for the calculation of plasma propulsion. For this latter code, part of the parallel experiments were conduced on a PRACE Tier-0 machine within a PRACE Project Access.

KERDATA Project-Team

5. New Results

5.1. Convergence of HPC and Big Data

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

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

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

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

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

5.1.2. Increasing small files access performance with dynamic metadata replication

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

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

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

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

5.1.3. Modeling elastic storage

Participants: Nathanaël Cheriere, Gabriel Antoniu.

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

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

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

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

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

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

5.2. Scalable stream storage

5.2.1. KerA ingestion and storage

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

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

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

5.2.2. Tailwind: fast and atomic RDMA-based replication

Participants: Yacine Taleb, Gabriel Antoniu.

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

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

5.3. Hybrid edge/cloud processing

5.3.1. Edge benchmarking

Participants: Pedro Silva, Alexandru Costan, Gabriel Antoniu.

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

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

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

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

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

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

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

5.3.3. *Integrating KerA and Flink*

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

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

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

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

5.4.1. *HDF-based storage*

Participants: Hadi Salimi, Gabriel Antoniu.

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

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

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

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

Participants: Hadi Salimi, Gabriel Antoniu.

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

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

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

POLARIS Project-Team

7. New Results

7.1. Design of Experiments

A large amount of resources is spent writing, porting, and optimizing scientific and industrial High Performance Computing applications, which 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 a large budget of costly experimental measurements to provide good results, while rarely providing exploitable knowledge after optimization. In [40], we present a user-transparent (white-box) autotuning technique based on Design of Experiments 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 $10\times$ less measurements than random sampling.

7.2. Experimenting with Fog Infrastructures

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. Fog computing provides computing, storage and communication resources at the edge of the network, near the physical world. This section describes two independent contributions on how to study and develop FOG infrastructures. These contributions take place in the context of the Inria/Orange Labs joint laboratory.

- Despite its several advantages, Fog computing raises new challenges which slow its adoption down. In particular, there are currently few practical solutions allowing to exploit such infrastructure and to evaluate potential strategies. In [42], we propose a prototype orchestration architecture building on both Grid5000 and Fit-IoT lab (SILECS). This experimental testbed allows to realistically and rigorously evaluate orchestration strategies. In [20], we propose FITOR, an orchestration system for IoT applications in the Fog environment, which extends the actor-model based Calvin framework to cope with Fog environments 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.
- End devices nearing the physical world can have interesting properties such as short delays, responsiveness, optimized communications and privacy. However, these end devices have low stability and are prone to failures. There is consequently a need for failure management protocols for IoT applications in the Fog. The design of such solutions is complex due to the specificities of the environment, i.e., (i) dynamic infrastructure where entities join and leave without synchronization, (ii) high heterogeneity in terms of functions, communication models, network, processing and storage capabilities, and, (iii) cyber-physical interactions which introduce non-deterministic and physical world's space and time dependent events. In [29], [37], we present a fault tolerance approach taking into account these three characteristics of the Fog-IoT environment. Fault tolerance is achieved by saving the state of the application in an uncoordinated way. When a failure is

detected, notifications are propagated to limit the impact of failures and dynamically reconfigure the application. Data stored during the state saving process are used for recovery, taking into account consistency with respect to the physical world. The approach was validated through practical experiments on a smart home platform.

7.3. HPC Application Analysis and Visualization

- Programming paradigms in High-Performance Computing have been shifting towards task-based models which are capable of adapting readily to heterogeneous and scalable supercomputers. The performance of task-based application heavily depends on the runtime scheduling heuristics and on its ability to exploit computing and communication resources. Unfortunately, the traditional performance analysis strategies are unfit to fully understand task-based runtime systems and applications: they expect a regular behavior with communication and computation phases, while task-based applications demonstrate no clear phases. Moreover, the finer granularity of task-based applications typically induces a stochastic behavior that leads to irregular structures that are difficult to analyze. Furthermore, the combination of application structure, scheduler, and hardware information is generally essential to understand performance issues. The papers [36], [6] presents a flexible framework that enables one to combine several sources of information and to create custom visualization panels allowing to understand and pinpoint performance problems incurred by bad scheduling decisions in task-based applications. Three case-studies using StarPU-MPI, a task-based multi-node runtime system, are detailed to show how our framework can be used to study the performance of the well-known Cholesky factorization. Performance improvements include a better task partitioning among the multi-(GPU,core) to get closer to theoretical lower bounds, improved MPI pipelining in multi-(node,core,GPU) to reduce the slow start, and changes in the runtime system to increase MPI bandwidth, with gains of up to 13% in the total makespan.
- In the context of multi-physics simulations on unstructured and heterogeneous meshes, generating well-balanced partitions is not trivial. The computing cost per mesh element in different phases of the simulation depends on various factors such as its type, its connectivity with neighboring elements or its layout in memory with respect to them, which determines the data locality. Moreover, if different types of discretization methods or computing devices are combined, the performance variability across the domain increases. Due to all these factors, evaluate a representative computing cost per mesh element, to generate well-balanced partitions, is a difficult task. Nonetheless, load balancing is a critical aspect of the efficient use of extreme scale systems since idle-times can represent a huge waste of resources, particularly when a single process delays the overall simulation. In this context, we present in [16] some improvements carried out on an in-house geometric mesh partitioner based on the Hilbert Space-Filling Curve. We have previously tested its effectiveness by partitioning meshes with up to 30 million elements in a few tenths of milliseconds using up to 4096 CPU cores, and we have leveraged its performance to develop an autotuning approach to adjust the load balancing according to runtime measurements. In this paper, we address the problem of having different load distributions in different phases of the simulation, particularly in the matrix assembly and in the solution of the linear system. We consider a multi-partition approach to ensure a proper load balance in all the phases. The initial results presented show the potential of this strategy.

7.4. Energy Optimization and Smart Grids Simulation

Large-scale decentralized photovoltaic (PV) generators are currently being installed in many low-voltage distribution networks. Without grid reinforcements or production curtailment, they might create current and/or voltage issues. In [13], [45], we consider the use the advanced metering infrastructure (AMI) as the basis for PV generation control. We show that the advanced metering infrastructure may be used to infer some knowledge about the underlying network, and we show how this knowledge can be used by a simple feed-forward controller to curtail the solar production efficiently.

We developed a environment for co-simulating electrical networks, telecommunication networks and online learning algorithms [3]. One the outputs of this work was to allow us to perform realistic numerical simulations of active distribution networks. We used this simulator to compare our proposed controller with two other controller structures: open-loop, and feedback P (U) and Q(U). We demonstrate that our feed-forward controller –that requires no prior knowledge of the underlying electrical network– brings significant performance improvements as it can effectively suppress over-voltage and over-current while requiring low energy curtailment. This method can be implemented at low cost and require no specific information about the network on which it is deployed.

Finally, we study demand-Response (DR) programs, whereby users of an electricity network are encouraged by economic incentives to rearrange their consumption in order to reduce production costs. Such mechanisms are envisioned to be a key feature of the smart grid paradigm. Several recent works proposed DR mechanisms and used analytical models to derive optimal incentives. Most of these works, however, rely on a macroscopic description of the population that does not model individual choices of users. In in[4], we conduct a detailed analysis of those models and we argue that the macroscopic descriptions hide important assumptions that can jeopardize the mechanisms' implementation (such as the ability to make personalized offers and to perfectly estimate the demand that is moved from a timeslot to another). Then, we start from a microscopic description that explicitly models each user's decision. We introduce four DR mechanisms with various assumptions on the provider's capabilities. Contrarily to previous studies, we find that the optimization problems that result from our mechanisms are complex and can be solved numerically only through a heuristic. We present numerical simulations that compare the different mechanisms and their sensitivity to forecast errors. At a high level, our results show that the performance of DR mechanisms under reasonable assumptions on the provider's capabilities are significantly lower than those suggested by previous studies, but that the gap reduces when the population's flexibility increases.

7.5. Simulation of HPC Applications

Beside continuous development and contribution to the SimGrid project, the two following contributions have been published this year. Both build on the SMPI interface which allows to efficiently predict the performance of MPI applications.

- Finite-difference methods are commonplace in High Performance Computing applications. Despite their apparent regularity, they often exhibit load imbalance that damages their efficiency. In [9], we characterize the spatial and temporal load imbalance of Ondes3D, a typical finite-differences application dedicated to earthquake modeling. Our analysis reveals imbalance originating from the structure of the input data, and from low-level CPU optimizations. Ondes3D was successfully ported to AMPI/CHARM++ using over-decomposition and MPI process migration techniques to dynamically rebalance the load. However, this approach requires careful selection of the over-decomposition level, the load balancing algorithm, and its activation frequency. These choices are usually tied to application structure and platform characteristics. We have thus proposed a workflow that leverages the capabilities of SimGrid to conduct such study at low experimental cost. We rely on a combination of emulation, simulation, and application modeling that requires minimal code modification and manages to capture both spatial and temporal load imbalance to faithfully predict the performance of dynamic load balancing. We evaluate the quality of our simulation by comparing simulation results with the outcome of real executions and demonstrate how this approach can be used to quickly find the optimal load balancing configuration for a given application/hardware configuration.
- It is typical in High Performance Computing (HPC) courses to give students access to HPC platforms so that they can benefit from hands-on learning opportunities. Using such platforms, however, comes with logistical and pedagogical challenges. For instance, a logistical challenge is that access to representative platforms must be granted to students, which can be difficult for some institutions or course modalities; and a pedagogical challenge is that hands-on learning opportunities are constrained by the configurations of these platforms. A way to address these challenges is to

instead simulate program executions on arbitrary HPC platform configurations. In [19] we focus on simulation in the specific context of distributed-memory computing and MPI programming education. While using simulation in this context has been explored in previous works, our approach offers two crucial advantages. First, students write standard MPI programs and can both debug and analyze the performance of their programs in simulation mode. Second, large-scale executions can be simulated in short amounts of time on a single standard laptop computer. This is possible thanks to SMPI, an MPI simulator provided as part of SimGrid. After detailing the challenges involved when using HPC platforms for HPC education and providing background information about SMPI, we present SMPI Courseware. SMPI Courseware is a set of in-simulation assignments that can be incorporated into HPC courses to provide students with hands-on experience for distributed-memory computing and MPI programming learning objectives. We describe some these assignments, highlighting how simulation with SMPI enhances the student learning experience.

7.6. Mean Field and Refined Mean Field Methods

Mean field approximation is a popular means to approximate stochastic models that can be represented as a system of N interacting objects. It is know to be exact as N goes to infinity. In a recent series of paper, [24], [25], [7], we establish theoretical results and numerical methods that allows us to define an approximation that is much more accurate than the classical mean field approximation. This new approximation, that we call the *refined mean field approximation*, is based on the computation of an expansion term of the order $1/N$. By considering a variety of applications, that include coupon collector, load balancing and bin packing problems, we illustrate that the proposed refined mean field approximation is significantly more accurate that the classic mean field approximation for small and moderate values of N: relative errors are often below 1% for systems with $N = 10$.

In [23], [8], we improve this result in two directions. First, we show how to obtain the same result for the transient regime. Second, we provide a further refinement by expanding the term in $1/N^2$ (both for transient and steady-state regime). Our derivations are inspired by moment-closure approximation, a popular technique in theoretical biochemistry. We provide a number of examples that show: (1) that this new approximation is usable in practice for systems with up to a few tens of dimensions, and (2) that it accurately captures the transient and steady state behavior of such systems.

7.7. Optimization of Networks and Communication

This section describes two independent contributions on the analysis and optimization of networks and communication.

- Telecommunication networks are converging to a massively distributed cloud infrastructure interconnected with software defined networks. In the envisioned architecture, services will be deployed flexibly and quickly as network slices. Our paper [26] addresses a major bottleneck in this context, namely the challenge of computing the best resource provisioning for network slices in a robust and efficient manner. With tractability in mind, we propose a novel optimization framework which allows fine-grained resource allocation for slices both in terms of network bandwidth and cloud processing. The slices can be further provisioned and auto-scaled optimally based on a large class of utility functions in real-time. Furthermore, by tuning a slice-specific parameter, system designers can trade off traffic-fairness with computing-fairness to provide a mixed fairness strategy. We also propose an iterative algorithm based on the alternating direction method of multipliers (ADMM) that provably converges to the optimal resource allocation and we demonstrate the method's fast convergence in a wide range of quasi-stationary and dynamic settings.
- Distributed power control schemes in wireless networks have been well-examined, but standard methods rarely consider the effect of potentially random delays, which occur in almost every real-world network. We present in paper [33] Robust Feedback Averaging, a novel power control algorithm that is capable of operating in delay-ridden and noisy environments. We prove optimal convergence of this algorithm in the presence of random, time-varying delays, and present numerical

simulations that indicate that Robust Feedback Averaging outperforms the ubiquitous Foschini-Miljanic algorithm in several regimes.

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

Bringing transparency to algorithmic decision making systems and guaranteeing that the system satisfies properties of fairness and privacy is crucial in today's world. To start tackling this broad challenge, we focused on the case of online advertising and we had the following contributions.

- *Transparency properties for social media advertising and audit of Facebook's explanations.* In [15], we took a first step towards exploring the transparency mechanisms provided by social media sites, focusing on the two processes for which Facebook provides transparency mechanisms: the process of how Facebook infers data about users, and the process of how advertisers use this data to target users. We call explanations about those two processes *data explanations* and *ad explanations*, respectively.

  We identify a number of *properties* that are key for different types of explanations aimed at bringing transparency to social media advertising. We then evaluate empirically how well Facebook's explanations satisfy these properties and discuss the implications of our findings in view of the possible purposes of explanations. In particular, for *ad explanations*, we define five key properties: *personalization*, *completeness*, *correctness* (and the companion property of *misleadingness*), *consistency*, and *determinism*, and we show that Facebook's ad explanations are often *incomplete* and sometimes *misleading*. In particular, we observe that Facebook reveals only the most prevalent attribute used by the advertisers, which may allow malicious advertisers to easily obfuscate ad explanations from ad campaigns that are discriminatory or that target privacy-sensitive attributes. For *data explanations*, we define four key properties of the explanations: *specificity*, *snapshot completeness*, *temporal completeness*, and *correctness*; and we show that Facebook's explanations are *incomplete* and often *vague*; hence potentially limiting user control.

  Overall, our study provides a first step towards better understanding and improving transparency in social media advertising. During this work, we developed the tool AdAnalyst (https://adanalyst.mpi-sws.org/), which was instrumental for the study but also provides a transparency tool on its own for the large public, and is anticipated to be the basis of a number of further research studies in transparency.
- *Potential for discrimination in social media advertising.* Recently, online targeted advertising platforms like Facebook have been criticized for allowing advertisers to discriminate against users belonging to sensitive groups, i.e., to exclude users belonging to a certain race or gender from receiving their ads. Such criticisms have led, for instance, Facebook to disallow the use of attributes such as ethnic affinity from being used by advertisers when targeting ads related to housing or employment or financial services. In our paper [30], we systematically investigate the different targeting methods offered by Facebook (traditional attribute- or interest-based targeting, custom audience and lookalike audience) for their ability to enable discriminatory advertising and showed that a malicious advertiser can create highly discriminatory ads without using sensitive attributes (hence banning those features is inefficient to solve the problem). We argue that discrimination measures should be based on the targeted population and not on the attributes used for targeting and propose a discrimination metric in this direction.
- *Identification and resolution of privacy leakages in the Facebook's advertising platform.* In paper [31] we discovered that the information provided to advertisers through the custom audience feature (where an advertisers can upload PIIs (Personally Identifiable Information) of their customers and Facebook matches those with their users) was very severely leaking personal information. Specifically, it was making it possible for a malicious advertiser knowing the email address of a user to discover its phone number. Perhaps even worse, it was allowing a malicious advertiser to de-anonymize visitors of a website he controls. We discovered that the problem was due to the way Facebook computes estimates of the number of users matching a list of PIIs and proposed a solution based on not de-duplicating records with different PIIs belonging to the same users; and

we proved the robustness of our solution theoretically. Our work led to Facebook implementing a solution inspired by the one we proposed.

7.9. Optimization Methods

This section describes four independent contributions on optimization.

- In view of solving convex optimization problems with noisy gradient input, we analyze in the paper [11] the asymptotic behavior of gradient-like flows under stochastic disturbances. Specifically, we focus on the widely studied class of mirror descent schemes for convex programs with compact feasible regions, and we examine the dynamics' convergence and concentration properties in the presence of noise. In the vanishing noise limit, we show that the dynamics converge to the solution set of the underlying problem (a.s.). Otherwise, when the noise is persistent, we show that the dynamics are concentrated around interior solutions in the long run, and they converge to boundary solutions that are sufficiently "sharp". Finally, we show that a suitably rectified variant of the method converges irrespective of the magnitude of the noise (or the structure of the underlying convex program), and we derive an explicit estimate for its rate of convergence.
- We examine in paper [12] a class of stochastic mirror descent dynamics in the context of monotone variational inequalities (including Nash equilibrium and saddle-point problems). The dynamics under study are formulated as a stochastic differential equation driven by a (single-valued) monotone operator and perturbed by a Brownian motion. The system's controllable parameters are two variable weight sequences that respectively pre- and post-multiply the driver of the process. By carefully tuning these parameters, we obtain global convergence in the ergodic sense, and we estimate the average rate of convergence of the process. We also establish a large deviations principle showing that individual trajectories exhibit exponential concentration around this average.
- We develop in [17] 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.
- One of the most widely used training methods for large-scale machine learning problems is distributed asynchronous stochastic gradient descent (DASGD). However, a key issue in its implementation is that of delays: when a "worker" node asynchronously contributes a gradient update to the "master", the global model parameter may have changed, rendering this information stale. In massively parallel computing grids, these delays can quickly add up if a node is saturated, so the convergence of DASGD is uncertain under these conditions. Nevertheless, by using a judiciously chosen quasilinear step-size sequence, we show in [35] that it is possible to amortize these delays and achieve global convergence with probability 1, even under polynomially growing delays, reaffirming in this way the successful application of DASGD to large-scale optimization problems.

7.10. Multi-agent Learning and Distributed Best Response

This section describes several independent contributions on multi-agent learning.

- In [5], [22], [21], we study how fast can simple algorithms compute Nash equilibria. We study the case of random potential games for which we have designed and analyzed distributed algorithms to compute a Nash equilibrium. Our algorithms are based on best-response dynamics, with suitable revision sequences (orders of play). We compute the average complexity over all potential games

of best response dynamics under a random i.i.d. revision sequence, since it can be implemented in a distributed way using Poisson clocks. We obtain a distributed algorithm whose execution time is within a constant factor of the optimal centralized one. We also showed how to take advantage of the structure of the interactions between players in a network game: non- interacting players can play simultaneously. This improves best response algorithm, both in the centralized and in the distributed case.

- In [10], we study a class of evolutionary game dynamics defined by balancing a gain determined by the game's payoffs against a cost of motion that captures the difficulty with which the population moves between states. Costs of motion are represented by a Riemannian metric, i.e., a state-dependent inner product on the set of population states. The replicator dynamics and the (Euclidean) projection dynamics are the archetypal examples of the class we study. Like these representative dynamics, all Riemannian game dynamics satisfy certain basic desiderata, including positive correlation and global convergence in potential games. Moreover, when the underlying Riemannian metric satisfies a Hessian integrability condition, the resulting dynamics preserve many further properties of the replicator and projection dynamics. We examine the close connections between Hessian game dynamics and reinforcement learning in normal form games, extending and elucidating a well-known link between the replicator dynamics and exponential reinforcement learning.
- The paper [18] examines the long-run behavior of learning with bandit feedback in non-cooperative concave games. The bandit framework accounts for extremely low-information environments where the agents may not even know they are playing a game; as such, the agents' most sensible choice in this setting would be to employ a no-regret learning algorithm. In general, this does not mean that the players' behavior stabilizes in the long run: no-regret learning may lead to cycles, even with perfect gradient information. However, if a standard monotonicity condition is satisfied, our analysis shows that no-regret learning based on mirror descent with bandit feedback converges to Nash equilibrium with probability 1. We also derive an upper bound for the convergence rate of the process that nearly matches the best attainable rate for single-agent bandit stochastic optimization.
- In [34], we consider a game-theoretical multi-agent learning problem where the feedback information can be lost during the learning process and rewards are given by a broad class of games known as variationally stable games. We propose a simple variant of the classical online gradient descent algorithm, called reweighted online gradient descent (ROGD) and show that in variationally stable games, if each agent adopts ROGD, then almost sure convergence to the set of Nash equilibria is guaranteed, even when the feedback loss is asynchronous and arbitrarily corrrelated among agents. We then extend the framework to deal with unknown feedback loss probabilities by using an estimator (constructed from past data) in its replacement. Finally, we further extend the framework to accomodate both asynchronous loss and stochastic rewards and establish that multi-agent ROGD learning still converges to the set of Nash equilibria in such settings. Together, these results contribute to the broad lanscape of multi-agent online learning by significantly relaxing the feedback information that is required to achieve desirable outcomes.
- Regularized learning is a fundamental technique in online optimization, machine learning and many other fields of computer science. A natural question that arises in these settings is how regularized learning algorithms behave when faced against each other. In the paper [27], we study a natural formulation of this problem by coupling regularized learning dynamics in zero-sum games. We show that the system's behavior is Poincaré recurrent, implying that almost every trajectory revisits any (arbitrarily small) neighborhood of its starting point infinitely often. This cycling behavior is robust to the agents' choice of regularization mechanism (each agent could be using a different regularizer), to positive-affine transformations of the agents' utilities, and it also persists in the case of networked competition, i.e., for zero-sum polymatrix games.

7.11. Blotto games

The Colonel Blotto game is a famous game commonly used to model resource allocation problems in many domains ranging from security to advertising. Two players distribute a fixed budget of resources on multiple

battlefields to maximize the aggregate value of battlefields they win, each battlefield being won by the player who allocates more resources to it. The continuous version of the game –where players can choose any fractional allocation– has been extensively studied, albeit only with partial results to date. Recently, the discrete version –where allocations can only be integers– started to gain traction and algorithms were proposed to compute the equilibrium in polynomial time; but these remain computationally impractical for large (or even moderate) numbers of battlefields. In [32], [46], we propose an algorithm to compute very efficiently an approximate equilibrium for the discrete Colonel Blotto game with many battlefields. We provide a theoretical bound on the approximation error as a function of the game's parameters, in particular number of battlefields and resource budgets. We also propose an efficient dynamic programming algorithm to compute the best-response to any strategy that allows computing for each game instance the actual value of the error. We perform numerical experiments that show that the proposed strategy provides a fast and good approximation to the equilibrium even for moderate numbers of battlefields.

ROMA Project-Team

7. New Results

7.1. Birkhoff–von Neumann decomposition

The well-known Birkhoff-von Neumann (BvN) decomposition expresses a doubly stochastic matrix as a convex combination of a number of permutation matrices. For a given doubly stochastic matrix, there are many BvN decompositions, and finding the one with the minimum number of permutation matrices is NP-hard. There are heuristics to obtain BvN decompositions for a given doubly stochastic matrix. A family of heuristics are based on the original proof of Birkhoff and proceed step by step by subtracting a scalar multiple of a permutation matrix at each step from the current matrix, starting from the given matrix. At every step, the subtracted matrix contains nonzeros at the positions of some nonzero entries of the current matrix and annihilates at least one entry, while keeping the current matrix nonnegative. Our first result, which supports a claim of Brualdi [68], shows that this family of heuristics can miss optimal decompositions. We also investigate the performance of two heuristics from this family theoretically. The findings are published in a journal [10].

7.2. Parallel sparse matrix-vector multiply

There are three common parallel sparse matrix-vector multiply algorithms: 1D 3 row-parallel, 1D column-parallel and 2D row-column-parallel. The 1D parallel algorithms offer the advantage of having only one communication phase. On the other hand, the 2D parallel algorithm is more scalable but it suffers from two communication phases. In this work, we introduce a novel concept of heterogeneous messages where a heterogeneous message may contain both input-vector entries and partially computed output-vector entries. This concept not only leads to a decreased number of messages, but also enables fusing the input-and output-communication phases into a single phase. These findings are exploited to propose a 1.5D parallel sparse matrix-vector multiply algorithm which is called local row-column-parallel. This proposed algorithm requires a constrained fine-grain partitioning in which each fine-grain task is assigned to the processor that contains either its input-vector entry, or its output-vector entry, or both. We propose two methods to carry out the constrained fine-grain partitioning. We conduct our experiments on a large set of test matrices to evaluate the partitioning qualities and partitioning times of these proposed 1.5D methods. The findings are published in a journal [14].

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

We consider a variant of the well-known, NP-complete problem of minimum cut linear arrangement for directed acyclic graphs. In this variant, we are given a directed acyclic graph and we are asked to find a topological ordering such that the maximum number of cut edges at any point in this ordering is minimum. In our variant, the vertices and edges have weights, and the aim is to minimize the maximum weight of cut edges in addition to the weight of the last vertex before the cut. There is a known, polynomial time algorithm [78] for the cases where the input graph is a rooted tree. We focus on the instances where the input graph is a directed series-parallel graph, and propose a polynomial time algorithm, thus expanding the class of graphs for which a polynomial time algorithm is known. Directed acyclic graphs are used to model scientific applications where the vertices correspond to the tasks of a given application and the edges represent the dependencies between the tasks. In such models, the problem we address reads as minimizing the peak memory requirement in an execution of the application. Our work, combined with Liu's work on rooted trees addresses this practical problem in two important classes of applications. The findings are published in a journal [15].

7.4. Parallel Candecomp/Parafac decomposition of sparse tensors using dimension trees

Tensor factorization has been increasingly used to address various problems in many fields such as signal processing, data compression, computer vision, and computational data analysis. CANDECOMP/PARAFAC (CP) decomposition of sparse tensors has successfully been applied to many well-known problems in web search, graph analytics, recommender systems, health care data analytics, and many other domains. In these applications, computing the CP decomposition of sparse tensors efficiently is essential in order to be able to process and analyze data of massive scale. For this purpose, we investigate an efficient computation and parallelization of the CP decomposition for sparse tensors. We provide a novel computational scheme for reducing the cost of a core operation in computing the CP decomposition with the traditional alternating least squares (CP-ALS) based algorithm. We then effectively parallelize this computational scheme in the context of CP-ALS in shared and distributed memory environments, and propose data and task distribution models for better scalability. We implement parallel CP-ALS algorithms and compare our implementations with an efficient tensor factorization library, using tensors formed from real-world and synthetic datasets. With our algorithmic contributions and implementations, we report up to 3.95x, 3.47x, and 3.9x speedups in sequential, shared memory parallel, and distributed memory parallel executions over the state of the art, and up to 1466x overall speedup over the sequential execution using 4096 cores on an IBM BlueGene/Q supercomputer. The findings are published in a journal [13].

7.5. Approximation algorithms for maximum matchings in undirected graphs

We propose heuristics for approximating the maximum cardinality matching on undirected graphs. Our heuristics are based on the theoretical body of a certain type of random graphs, and are made practical for real-life ones. The idea is based on judiciously selecting a subgraph of a given graph and obtaining a maximum cardinality matching on this subgraph. We show that the heuristics have an approximation guarantee of around $0.866 - \log(n)/n$ for a graph with n vertices. Experiments for verifying the theoretical results in practice are provided. The findings are published in a conference proceedings [25].

7.6. SINA: A Scalable iterative network aligner

Given two graphs, network alignment asks for a potentially partial mapping between the vertices of the two graphs. This arises in many applications where data from different sources need to be integrated. Recent graph aligners use the global structure of input graphs and additional information given for the edges and vertices. We present SINA, an efficient, shared memory parallel implementation of such an aligner. Our experimental evaluations on a 32-core shared memory machine showed that SINA scales well for aligning large real-world graphs: SINA can achieve up to $28.5\times$ speedup, and can reduce the total execution time of a graph alignment problem with 2M vertices and 100M edges from 4.5 hours to under 10 minutes. To the best of our knowledge, SINA is the first parallel aligner that uses global structure and vertex and edge attributes to handle large graphs. The findings are published in a conference proceedings [34].

7.7. Acyclic partitioning of large 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 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 inter-part 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 multi-way 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 improvements, on average, around 59% compared to the current state of the art. The findings are published in a research report [50].

7.8. 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 by several researchers. In this work, we first devise heuristics for this problem by generalizing the existing cheap graph matching heuristics. Then, we 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. The findings are published in a research report [46].

7.9. Scaling matrices and counting the perfect matchings in graphs

We investigate efficient randomized methods for approximating the number of perfect matchings in bipartite graphs and general graphs. Our approach is based on assigning probabilities to edges. The findings are published in a research report [47].

7.10. 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. 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 better performance in cases with many communications. The findings are published in a research report [53].

7.11. Data-Locality Aware Dynamic Schedulers for Independent Tasks with Replicated Inputs

In this work we concentrate on a crucial parameter for efficiency in Big Data and HPC applications: data locality. We focus on the scheduling of a set of independant tasks, each depending on an input file. We assume that each of these input files has been replicated several times and placed in local storage of different nodes of a cluster, similarly of what we can find on HDFS system for example. We consider two optimization problems, related to the two natural metrics: makespan optimization (under the constraint that only local tasks are allowed) and communication optimization (under the constraint of never letting a processor idle in order to optimize makespan). For both problems we investigate the performance of dynamic schedulers, in particular the basic greedy algorithm we can for example find in the default MapReduce scheduler. First we theoretically study its performance, with probabilistic models, and provide a lower bound for communication metric and asymptotic behaviour for both metrics. Second we propose simulations based on traces from a Hadoop cluster to compare the different dynamic schedulers and assess the expected behaviour obtained with the theoretical study.

These findings have been presented at the CEBDA workshop [19].

7.12. Parallel scheduling of DAGs under memory constraints.

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

This work has been presented at the IPDPS 2018 conference [31] and an extended version has been submitted to the Elsevier JPDC journal [52].

7.13. Online Scheduling of Task Graphs on Hybrid 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 [61], 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 $\Theta(\sqrt{m/k})$. 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.

This work has been presented at the EuroPar 2018 conference [24].

7.14. Memory-aware tree partitioning on homogeneous platforms

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

This work has been presented as a short paper in the PDP 2018 conference [27].

7.15. Reliability-aware energy optimization for throughput-constrained applications on MPSoC.

Multi-Processor System-on-Chip (MPSoC) has emerged as a promising platform to meet the increasing performance demand of embedded applications. However, due to limited energy budget, it is hard to guarantee that applications on MPSoC can be accomplished on time with a required throughput. The situation becomes even worse for applications with high reliability requirements, since extra energy will be inevitably consumed by task re-executions or duplicated tasks. Based on Dynamic Voltage and Frequency Scaling (DVFS) and task duplication techniques, this paper presents a novel energy-efficient scheduling model, which aims at minimizing the overall energy consumption of MPSoC applications under both throughput and reliability constraints. The problem is shown to be NP-complete, and several polynomial-time heuristics are proposed to tackle this problem. Comprehensive simulations on both synthetic and real application graphs show that our proposed heuristics can meet all the given constraints, while reducing the energy consumption.

This findings have been presented at the ICPADS 2018 conference [26].

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

Scientific workloads are often described by Directed Acyclic task Graphs. Indeed, DAGs represent both a theoretical model and the structure employed by dynamic runtime schedulers to handle HPC applications. A natural problem is then to compute a makespan-minimizing schedule of a given graph. In this paper, we are motivated by task graphs arising from multifrontal factorizations of sparse matrices and therefore work under the following practical model. Tasks are malleable (i.e., a single task can be allotted a time-varying number of processors) and their speedup behaves perfectly up to a first threshold, then speedup increases linearly, but not perfectly, up to a second threshold where the speedup levels off and remains constant.

After proving the NP-hardness of minimizing the makespan of DAGs under this model, we study several heuristics. We propose model-optimized variants for PROPSCHEDULING, widely used in linear algebra application scheduling, and FLOWFLEX. GREEDYFILLING is proposed, a novel heuristic designed for our speedup model, and we demonstrate that PROPSCHEDULING and GREEDYFILLING are 2-approximation algorithms. In the evaluation, employing synthetic data sets and task graphs arising from multifrontal factorization, the proposed optimized variants and GREEDYFILLING significantly outperform the traditional algorithms, whereby GREEDYFILLING demonstrates a particular strength for balanced graphs.

These findings have been published in the IEEE TPDS journal [16].

7.17. Performance and scalability of the block low-rank multifrontal factorization on multicore architectures

Matrices coming from elliptic Partial Differential Equations have been shown to have a low-rank property which can be efficiently exploited in multifrontal solvers to provide a substantial reduction of their complexity. Among the possible low-rank formats, the Block Low-Rank format (BLR) is reasonably easy to use in a general purpose multifrontal solver and its potential compared to standard (full-rank) solvers has been demonstrated. Recently, new variants have been introduced and it was proved that they can further reduce the complexity but their performance remained to be analyzed. We develop a multithreaded BLR factorization, and analyze its efficiency and scalability in shared-memory multicore environments. We identify the challenges posed by the use of BLR approximations in multifrontal solvers and put forward several algorithmic variants of the BLR factorization that overcome these challenges by improving its efficiency and scalability. We illustrate the performance analysis of the BLR multifrontal factorization with numerical experiments on a large set of problems coming from a variety of real-life applications.

This work has been accepted for publication in the ACM Transactions on Mathematical Software [5].

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

The cost of the solution phase in sparse direct methods is sometimes critical. It can be larger than that of the factorization in applications where systems of linear equations with thousands of right-hand sides (RHS) must be solved. In this work, we focus on the case of multiple sparse RHS with different nonzero structures in each column. In this setting, vertical sparsity reduces the number of operations by avoiding computations on rows that are entirely zero, and horizontal sparsity goes further by performing each elementary solve operation only on a subset of the RHS columns. To maximize the exploitation of horizontal sparsity, we propose a new algorithm to build a permutation of the RHS columns. We then propose an original approach to split the RHS columns into a minimal number of blocks, while reducing the number of operations down to a given threshold. Both algorithms are motivated by geometric intuitions and designed using an algebraic approach, so that they can be applied to general systems. We demonstrate the effectiveness of our algorithms on systems coming from real applications and compare them to other standard approaches. We also give some perspectives and possible applications.

This work has been accepted for publication in the SIAM Journal on Scientific Computing [6].

7.19. Efficient use of sparsity by direct solvers applied to 3D controlled-source EM problems

Controlled-source electromagnetic (CSEM) surveying becomes a widespread method for oil and gas exploration, which requires fast and efficient software for inverting large-scale EM datasets. In this context, one often needs to solve sparse systems of linear equations with a *large* number of *sparse* right-hand sides, each corresponding to a given transmitter position. Sparse direct solvers are very attractive for these problems, especially when combined with low-rank approximations which significantly reduce the complexity and the cost of the factorization. In the case of thousands of right-hand sides, the time spent in the sparse triangular solve tends to dominate the total simulation time and here we propose several approaches to reduce it. A significant reduction is demonstrated for marine CSEM application by utilizing the sparsity of the right-hand sides (RHS) and of the solutions that results from the geometry of the problem. Large gains are achieved by restricting computations at the forward substitution stage to exploit the fact that the RHS matrix might have empty rows (*vertical sparsity*) and/or empty blocks of columns within a non-empty row (*horizontal sparsity*). We also adapt the parallel algorithms that were designed for the factorization to solve-oriented algorithms and describe performance optimizations particularly relevant for the very large numbers of right-hand sides of the CSEM application. We show that both the operation count and the elapsed time for the solution phase can be significantly reduced. The total time of CSEM simulation can be divided by approximately a factor of 3 on all the matrices from our set (from 3 to 30 million unknowns, and from 4 to 12 thousands RHSs).

These findings are described in a technical report [37] and will be submitted for publication.

7.20. A Generic Approach to Scheduling and Checkpointing Workflows

We dealt with scheduling and checkpointing strategies to execute scientific workflows on failure-prone large-scale platforms. To the best of our knowledge, this work was 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 revisited classical mapping heuristics such as HEFT and MINMIN and complement them with several checkpointing strategies. The objective was 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.

This findings have been presented at the ICPP 2018 conference [28].

7.21. Scheduling independent stochastic tasks under deadline and budget constraints

We studied 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.

This findings have been presented at the SBAC-PAD 2018 conference [23].

STORM Project-Team

7. New Results

7.1. InKS Programming Model

Existing programming models tend to tightly interleave algorithm and optimization in HPC simulation codes. This requires scientists to become experts in both the simulated domain and the optimization process and makes the code difficult to maintain and port to new architectures. The InKS programming model, developed within the context of the PhD. Thesis of Ksander Ejjaaouani [9], decouples these two concerns with distinct languages for each. The simulation algorithm is expressed in the InKS pia language with no concern for machine-specific optimizations. Optimizations are expressed using both a family of dedicated optimizations DSLs (InKS O) and plain C++. InKS O relies on the InKS pia source to assist developers with common optimizations while C++ is used for less common ones. Our evaluation demonstrates the soundness of the approach by using it on synthetic benchmarks and the Vlasov-Poisson equation. It shows that InKS offers separation of concerns at no performance cost.

7.2. Porting Chameleon on top of OpenMP

Chameleon is a dense linear algebra software relying on sequential task-based algorithms where sub-tasks of the overall algorithms are submitted to a Runtime system. Algorithms were implemented on top of several task-based runtime systems: QUARK, PaRSEC, and StarPU (for which there is also an optional heterogeneous implementation). In the context of PRACE-5IP, we introduced OpenMP as an alternative backend for these linear algebra kernels.

7.3. StarPU in Julia

Julia is a modern language for parallelism and simulation that aims to ease the effort for developping high performance codes. In this context, we have started to develop a StarPU binding inside Julia. It is now possible to launch StarPU kernels inside Julia, either given as libraries, or described in Julia directly. Julia has the advantage to simplify significantly the syntax required to express the task and data management in StarPU (defining a new scope for StarPU, using automatic desallocation of buffers, ...).

Besides, using the introspection properties of Julia, the kernels written in Julia are automatically translated in both C codes and CUDA codes. Some preliminary experimental results show encouraging speedups on some limited codes. This is a work in progress, developped with A.Juven and M.Keryell.

7.4. Simulation and Validation of Error Correction Code Algorithms

The AFF3CT Error Correction Code (ECC) development and experimentation toolchain reached a major milestone with the release of the 2.x branch. It incorporates a hefty set of new modules and capabilities:

- New code families: Reed-Solomon, Turbo Product Code (TCP);
- New decoders: Maximum Likelihood (ML), Chase, LDPC Approximate Min-Star (AMS), LDPC Vertical Layered, LDPC Peeling, LDPC Bit Flipping;
- New channels: Optical, Binary Erasure Channel (BEC), Binary Symmetric Channel (BSC);
- New modem: On-Off Keying (OOK).

This new branch comes with extensive documentation of all available parameters at any point in the chain (https://aff3ct.readthedocs.io). In the process of the new release, the source code has also be reorganized in a rational and compartimentalized way, in terms of modules, tasks and sockets. This refactoring streamlines the use AFF3CT as a library, building on the concept of tasks, with well defined input and output sockets.

7.5. Speeding-Up Error Correction Code Processing using a Portable SIMD Wrapper

Error correction code (ECC) processing has so far been performed on dedicated hardware for previous generations of mobile communication standards, to meet latency and bandwidth constraints. As the 5G mobile standard, and its associated channel coding algorithms, are now being specified, modern CPUs are progressing to the point where software channel decoders can viably be contemplated. A key aspect in reaching this transition point is to get the most of CPUs SIMD units on the decoding algorithms being pondered for 5G mobile standards. The nature and diversity of such algorithms requires highly versatile programming tools. We proposed the virtues and versatility of our MIPP SIMD wrapper in implementing a high performance portfolio of key ECC decoding algorithms on AFF3CT [8].

7.6. Runtime System Interoperability with StarPU

Parallel HPC applications increasingly build on multiple parallel libraries, which results in interferences if the parallel entities in the application and in the libraries it uses access computing resources in an uncoordinated manner. A set of resource management APIs has therefore been designed within the context of H2020 project INTERTWinE (see http://www.intertwine-project.eu/developer-hub/resource-manager), and implemented in the StarPU task-based runtime system developed by Team STORM, as well as in the OmpSs/Nanos 6 task-based runtime system developed at the Barcelona Supercomputing Center (BSC). It enables StarPU and OmpSs to interoperate within an application, along multiple scenarios such as *nested interoperability*, with a host runtime system executing parallel tasks over a guest runtime system, or *concurrent interoperability*, with several runtime systems dynamically sharing computing resources over the application lifespan.

7.7. Hierarchicals Tasks

The programming model of StarPU, namely the sequential task flow model, was successfuly used in several applicative areas and was able to achieve high performance. However, the submission process needed either to be completely static, in the sense that the whole task graph is submitted at once, or to be stopped from time to time in order to control the execution. To overcome these limitations, we have introduced a new paradigm which we call *hierarchical tasks* where the so-called *control tasks* allow to submit at runtime a task subgraph. By allowing the submission of some parts of the task graph to be delayed until the execution of the corresponding control task, this feature allows to timely and dynamically choose the right version of the computation task subgraph (e.g. OpenMP, StarPU, cuda, or sequential, etc. implementations).

The graph of control tasks provide a high-level description of the computations which allow to use and design sophisticated scheduling algorithms. Furthermore, the cost of managing the control graph being much smaller than the one of the computation task graph, relying on the hierarchical tasks scheme enhances the scalability of the runtime system and allows to parallelize the submission process. Finally, this mechanism represents an elegant way of tackling the granularity issues which represent a key problem for achieving high performance in a heterogeneous context. The specificity of our implementation is to nicely combine hierarchical tasks with data partitioning, without needless synchronisation points.

7.8. Load Balancing Management in a Distributed Task-Based Programming Model

Distributed task-based programming models such as StarPU optimize the execution of applications based on an initial distribution of data. The resulting computational load on each node may however evolve over the course of the application, to the point where this initial distribution of data leads becomes suboptimal. It becomes necessary to correct the distribution the distribution of data to rebalance the load among nodes. Tools such as Zoltan or ParMetis do exist to perform this rebalancing job. However, they cannot be employed without breaking the application execution flow, and force synchronizing steps in fundamentally asynchronous task parallelism paradigms. Within the context of the internship of Loïc Jouans, we proposed a mechanism to enable the detection of load imbalance as well as the application of corrective measures to rebalance it while preserving the execution asynchrony.

7.9. Task-based Execution Visualization

One of the purpose of task-based programming is to let asynchronous execution achieve extreme pipelining of operations. This however make it a real challenge to determine why an execution performs poorly, since the execution trace shows the mixture of unrelated tasks. With the the University of Grenoble, we have designed a visualization framework which allows to easily visualize different metrics of the execution trace and apply different techniques to reveal the execution behavior. This allowed to determine and fix some erratic behaviors for instance in the StarPU runtime system and OpenMPI communication library [13], [4]

7.10. Interprocedural Collectives Verification

The advent to exascale requires more scalable and efficient techniques to help developers to locate, analyze and correct errors in parallel applications. PARallel COntrol flow Anomaly CHecker (PARCOACH) is a framework that detects the origin of collective errors in applications using MPI and/or OpenMP. In MPI, such errors include collective operations mismatches. In OpenMP, a collective error can be a barrier not called by all tasks in a team. We have developed an extension of PARCOACH which improves its collective errors detection [11]. The new analysis is more precise and accurate than the previous one on different benchmarks and real applications.

7.11. Profile-Guided Scope-Based Data Allocation Method

The complexity of High Performance Computing nodes memory system increases in order to challenge application growing memory usage and increasing gap between computation and memory access speeds. As these technologies are just being introduced in HPC supercomputers no one knows if it is better to manage them with hardware or software solutions. Thus both are being studied in parallel. For both solutions, the problem consists in choosing which data to store on which memory at any time.

In this context, we propose a linear formulation of the data allocation problem. Moreover, we propose a new profile- guided scope-based approach which reduces the data allocation problem complexity, thus enhancing the precision of state of the art analyzes. Finally we have implemented our method in a framework made of GCC plugins, dynamic libraries and python scripts, allowing to test the method on several benchmarks. We have evaluated our method on an INTEL Knight's Landing processor. To this aim we have run LULESH, HydroMM, two hydrodynamic codes, and MiniFE, a finite element mini application. We have compared our framework performance over these codes to several straight- forward solutions: MCDRAM as a cache, in hybrid mode, in flat mode using numactl command and existing AutoHBW dynamic library [7]

7.12. Lightweight Containerization of Computing Resources

SwLoc is a library for flexible and generic partitioning of computing resources (CPU, accelerators). It allows applications to create contexts (i.e. resource partitions) and run parallel codes inside such lightweight containers. Many libraries developed using OpenMP, Pthreads or Intel TBB can ben executed concurrently with little or no modification. SwLoc also features dynamic context resizing capabilities that enables parallel applications to perform resource negotiation.

7.13. Adaptive Partitioning for Iterated Sequences of Irregular OpenCL Kernels

OpenCL defines a common parallel programming language for CPU and GPU devices, although writing tasks adapted to the devices, managing communication and load-balancing issues are left to the programmer. We propose [10] a static/dynamic approach for the execution of an iterated sequence of data-dependent kernels on a multi-device heterogeneous architecture. The method allows to automatically distribute irregular kernels onto multiple devices and tackles, without training, both load balancing and data transfers issues coming from hardware heterogeneity, load imbalance within the application itself and load variations between repeated executions of the sequence. Our evaluation on some benchmarks and a complex N-body application, SOTL, simulating the electromagnetic Coulomb force applied on particles, show the interest of our approach.

7.14. A compiler front-end for OpenMP's variants

OpenMP 5.0 introduced the concept of *variant*: a directive which can be used to indicate that a function is a variant of another existing *base function*, in a specific context (eg: `foo_gpu_nvidia` could be declared as a variant of `foo`, but only when executing on specific NVidia hardware).

In the context of PRACE-5IP, we want to leverage this construct to be able to take advantage of the StarPU heterogeneous scheduler through the interoperability layer between OpenMP and StarPU.

We started this work by implementing the necessary changes in the Clang front-end to support OpenMP's *variant*.

7.15. Combining Task-based Parallelism and Adaptive Mesh Refinement Techniques in Molecular Dynamics Simulations

Modern parallel architectures require applications to generate massive parallelism so as to feed their large number of cores and their wide vector units. We have revisited the extensively studied classical Molecular Dynamics N-body problem in the light of these hardware constraints. We have introduced Adaptive Mesh Refinement techniques to store particles in memory, and to optimize the force computation loop using multi-threading and vectorization-friendly data structures [14]. Our design is guided by the need for load balancing and adaptivity raised by highly dynamic particle sets, as typically observed in simulations of strong shocks resulting in material micro-jetting. We have analyzed performance results on several simulation scenarios, over 512 nodes equipped by Intel Xeon Phi Knights Landing (KNL) processors. Performance obtained with our OpenMP implementation outperforms state-of-the-art implementations (LAMMPS) on both steady and micro-jetting particles simulations. In the latter case, our implementation is 1.38 times faster on KNL.

These results were obtained in the context of joint work between Inria and CEA/DAM.

TADAAM Project-Team

7. New Results

7.1. Checkpointing Strategies for Adjoint Computation on Hierarchical Platforms

The Adjoint Computation problem can be split in two phases: the forward phase where functions are successively evaluated on a particular input, and a backward phase computing the gradient descent. In the backward phase, the outputs of the forward phase are used* for the corresponding backward computation. On very large problems, all the forward outputs can not be kept in the memory at the same time, and one has to decide which output should be checkpointed and which output should be recomputed later on. The goal is to minimize the number of recomputation when reversing an Adjoint Computation Graph.

Griewank and Walther proved that, for a given number of available checkpoints with negligible writing and reading costs, the schedule that minimizes the amount of recomputation uses a binomial checkpointing strategy. We have designed an optimal algorithm to tackle the more general problem where we don't have only one level of memory with negligible access cost, but a hierarchical storage architecture. Each level of memory has its own size, writing and reading cost. The problem becomes more complex, since, not only we have to decide if an output should be checkpointed, be we have to decide in which level of the memory it should be kept. A trade-off must be found between the cost of memory accesses and that of recomputations.

We have designed an exact algorithm providing the optimal checkpointing strategy for a given Adjoint Computation Graph size and a description of the Hierarchical Platform; as well as heuristics. These algorithms can be found in the Software DISK-REVOLVE and a paper describing them is under writing process.

7.2. Modeling Non-Uniform Memory Access 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 [33] (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 [5] (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.

This work has been achieved in collaboration with the authors of the CARM from Universidade de Lisboa.

7.3. Cross Platform Classification for Detecting Locality Sensitivity and Selecting Data and Threads Placement Strategy

Individual nodes composing High Performance Computing (HPC) systems embed complex multicore and manycore processors. At this scale, compute tasks and data placement can double or halve execution times with respectively trivial are wise placements. While state of the art placement solutions can offer good performance improvements, they failed to set up as standards in supercomputers software stack. Current solutions are rather directed toward data or thread driven static policies. Among existing or promising future placement solutions a deep evaluation of applications response to these had yet to be done in order to wisely choose the best one.

With a set of 37 HPC representative applications, three different HPC processors, and 51 state of the art characterization metrics we built thousands models to evaluate applications response to data and threads placement policies. Thanks to a thorough methodology, our models were able to predict applications sensitivity to locality and their preferred placement policy both on new platforms and new applications. In the first case we were able to achieve more than 75% accuracy while preferred policy predictions approach optimal speedups in the second case.

This work was conducted using the PlaFRIM experimental testbed, in collaboration with Thomas Ropars from Laboratoire d'Informatique de Grenoble.

Several leads can be taken toward an extension of this work. For instance, predictions can be improved with benchmark directed learning. Models interpretation can also be furthered studied to refine the design of application characterization metrics.

7.4. 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. [13]

7.5. Memory Footprint of Locality Information on Many-Core Platforms

Exploiting the power of HPC platforms requires knowledge of their increasingly complex hardware topologies. Multiple components of the software stack, for instance MPI implementations or OpenMP runtimes, now perform their own topology discovery to find out the available cores and memory, and to better place tasks based on their affinities.

We studied the impact of this topology discovery in terms of memory footprint. Storing locality information wastes an amount of physical memory that is becoming an issue on many-core platforms on the road to exascale.

We demonstrated that this information may be factorized between processes by using a shared-memory region. Our analysis of the physical and virtual memories in supercomputing architectures showed that this shared region can be mapped at the same virtual address in all processes, hence dramatically simplifying the software implementation. [19]

Our implementation in HWLOC and Open MPI showed a memory footprint that does not increase with the number of MPI ranks per node anymore. Moreover the job launch time decreased by more than a factor of 2 on an Intel Knights Landing Xeon Phi and on a 96-core NUMA platform.

7.6. New abstraction to manage hardware topologies in MPI applications

Since the end of year 2016, we have been working on new abstractions and mechanisms that can allow the programmer to take advantage of the underlying hardware topology in their parallel applications developed in MPI. For instance, taking into account the intricate network/memory hierarchy can lead to substantial improvements in communication performance and reduce altogether the overall execution time of the application. However, it is important to find the relevant level of abstraction, as too much details are not usable practically because the programmer is not a a hardware specialist most of the time. Also, MPI being hardware-agnostic,

it is important to find means to use the hardware specifics without being tied to a particular architecture or hardware design.

With these goals in mind, we proposed the HSPLIT (see Section 6.1) libray that implements a solution based on a well-known MPI concept, the *communicators* (that can be seen as groups of communicating processes) [7], [19]. With HSPLIT, each level in the hardware hierarchy is accessible through a dedicated communicator. In this way, the programmer can leverage the underlying hierarchy in their application quite simply . The current implementation of HSPLIT is based on both HWLOC and NETLOC.

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

Also, this work has lead to the joint developement of the Hippo software with the CERFACS. Thanks to this piece of software, hybrid OpenMP/MPI applications can leverage the underlying physical hierarchy in order to better place MPI processes and OpenMP threads. This is particularly useful in a context where the application is composed of several kernels that use their own placement and mapping policy for processes and threads to achieve the best performance. Thanks to HSPLIT and HWLOC, CERFACS is now able to write codes in a more portable fashion without to solely rely on interactions of OpenMP and MPI runtimes for mapping and binding of processes/threads management.

7.7. Scheduling Parallel Tasks under Multiple Resources: List Scheduling vs. Pack Scheduling

Scheduling in High-Performance Computing (HPC) has been traditionally centered around computing resources (e.g., processors/cores). The ever-growing amount of data produced by modern scientific applications start to drive novel architectures and new computing frameworks to support more efficient data processing, transfer and storage for future HPC systems. This trend towards data-driven computing demands the scheduling solutions to also consider other resources (e.g., I/O, memory, cache) that can be shared amongst competing applications. In this paper, we study scheduling HPC applications while exploring the availability of multiple resources that could impact their performance. The goal is to minimize the overall execution time, or makespan, for a set of moldable tasks under multi-resource constraints. Two scheduling paradigms, namely, list scheduling and pack scheduling, are compared through both theoretical analyses and experimental evaluation. Theoretically, we prove, for several algorithms falling in the two scheduling paradigms, tight approximation ratios that increase linearly with the number of resource types. As the complexity of the direct solutions grows exponentially with the number of resource types, we also design a strategy to indirectly solve the problem via a transformation to a single-resource problem, which can significantly reduce the algorithms' running times without compromising their approximation ratios. Experiments conducted on Intel Knights Landing with two types of resources (processor cores and high-bandwidth memory) and simulations designed on more resource types confirm the benefit of the transformation strategy and show that pack-based scheduling, despite having a slightly worse theoretical bound, offers a practically promising and easy-to-implement solution, especially when managing a large number of resources. [20]

7.8. Sizing Burst-Buffers efficiently

Burst-Buffers are high throughput, small size intermediate storage systems typically based on SSDs or NVRAM that are designed to be used as a potential buffer between the computing nodes of a supercomputer and its main storage system consisting of hard drives. Their purpose is to absorb the bursts of I/O that many HPC applications experience (for example for saving checkpoints or data from intermediate results). In this paper, we propose a probabilistic model for evaluating the performance of Burst-Buffers. From a model of application and a data management strategy, we build a Markov-chain-based model of the system, that allows us to quickly answer issues about dimensioning of the system: for a given set of applications, and for a given Burst-Buffers size and bandwidth, how often does the buffer overflow? We also provide extensive simulation results to validate our modeling approach. [12], [25]

7.9. Scheduling for Neurosciences

In this project in collaboration with the Vanderbilt University, we are interested in scheduling stochastic jobs (originating from Neuroscience applications) 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 depends on both the request duration and the actual execution time of the job. A reservation strategy is 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 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. [14], [27]

7.10. Process Affinity, Metrics and Impact on Performance

Process placement, also called topology mapping, is a well-known strategy to improve parallel program execution by reducing the communication cost between processes. It requires two inputs: the topology of the target machine and a measure of the affinity between processes. In the literature, the dominant affinity measure is the communication matrix that describes the amount of communication between processes. The goal of this work is to study the accuracy of the communication matrix as a measure of affinity. We have done an extensive set of tests with two fat-tree machines and a 3d-torus machine to evaluate several hypotheses that are often made in the literature and to discuss their validity. First, we have checked the correlation between algorithmic metrics and the performance of the application. Then, we have checked whether a good generic process placement algorithm never degrades performance. And finally, we have seen whether the structure of the communication matrix can be used to predict gain [16].

7.11. Scheduling bi-colored-chains

In high performance computing, platform are shared by concurrent applications, each able to work with immense amount of data. As the file system is shared, we need to tackle congestion problems. One way to avoid increased I/O duration is to schedule the tasks with regards to their requests.

We proposed a theoretical model, bi-colored chains, that models applications with two alternating phases on distinct resources. After showing that minimizing the makespan with this model is a NP-complete problem in most cases. We studied particular cases, especially periodic applications and periodic schedule and provided approximation algorithms. This model will be developed in a PhD that started this fall, and enrich with practical data from simulations.

An extended intership report is available here: [31].

7.12. Experimenting task-based runtimes on a legacy Computational Fluid Dynamics code with unstructured meshes

Advances in high performance computing hardware systems lead to higher levels of parallelism and optimizations in scientific applications and more specifically in computational fluid dynamics codes. To reduce the level of complexity that such architectures bring while attaining an acceptable amount of the parallelism offered by modern clusters, the task-based approach has gained a lot of popularity recently as it is expected to deliver portability and performance with a relatively simple programming model. In this work, we have reviewed and presented the process of adapting part of Code Saturne, a legacy code at EDF R&D into a task-based form using the PARSEC (Parallel Runtime Scheduling and Execution Control) framework. We have first shown show the adaptation of our prime algorithm to a simpler form to remove part of the complexity of our code and then present its task-based implementation. We then have compared performance of various forms of our code and discuss the perks of task-based runtimes in terms of scalability, ease of incremental deployment in a legacy CFD code, and maintainability [8].

7.13. Progress threads placement for overlapping MPI non-blocking collectives using simultaneous multi-threading

Non-blocking collectives have been proposed so as to allow communications to be overlapped with computation in order to amortize the cost of MPI collective operations. To obtain a good overlap ratio, communications and computation have to run in parallel. To achieve this, different hardware and software techniques exists. Using dedicated cores for progress threads is one of them. However, some CPUs provide Simultaneous Multi-Threading, which is the ability for a core to have multiple hardware threads running simultaneously, sharing the same arithmetic units. We propose [18], [3] to use SMT to run progress threads to avoid dedicated cores allocation. We have run benchmarks on Haswell processors, using its Hyper-Threading capability, and get good results for both performance and overlap for inter-node communications. However, we have shown that Simultaneous Multi-Threading for intra-communications leads to bad performances due to contention on cache.

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 [28], [17], [21], [3] 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. Multi-criteria graph partitioning

The inclusion of multi-constraint graph partitioning algorithms in SCOTCH resulted in the obtainment of balanced multi-constraint partitions for a simulation software used in an industrial context [15]. This protype version is being transferred into the trunk of the SCOTCH package. Much of this year's software development has been devoted to the refactoring of the multi-threading management of the sequential version of the SCOTCH library.

DIVERSE Project-Team

6. New Results

6.1. Results on Variability modeling and management

6.1.1. Variability and testing.

Many approaches for testing configurable software systems start from the same assumption: it is impossible to test all configurations. This motivated the definition of variability-aware abstractions and sampling techniques to cope with large configuration spaces. Yet, there is no theoretical barrier that prevents the exhaustive testing of all configurations by simply enumerating them, if the effort required to do so remains acceptable. Not only this: we believe there is lots to be learned by systematically and exhaustively testing a configurable system. We report on the first ever endeavor to test all possible configurations of an industry-strength, open source configurable software system, JHipster, a popular code generator for web applications. We built a testing scaffold for the 26,000+ configurations of JHipster using a cluster of 80 machines during 4 nights for a total of 4,376 hours (182 days) CPU time. We find that 35.70% configurations fail and we identify the feature interactions that cause the errors. We show that sampling testing strategies (like dissimilarity and 2-wise) (1) are more effective to find faults than the 12 default configurations used in the JHipster continuous integration; (2) can be too costly and exceed the available testing budget. We cross this quantitative analysis with the qualitative assessment of JHipster's lead developers. Publication at Empirical Software Engineering: [25] See also, in the rest of the report, the work on *Multimorphic Testing* that actually relies on variability techniques.

6.1.2. Variability and teaching.

Software Product Line (SPL) engineering has emerged to provide the means to efficiently model, produce, and maintain multiple similar software variants, exploiting their common properties, and managing their variabilities (differences). With over two decades of existence, the community of SPL researchers and practitioners is thriving as can be attested by the extensive research output and the numerous successful industrial projects. Education has a key role to support the next generation of practitioners to build highly complex, variability-intensive systems. Yet, it is unclear how the concepts of variability and SPLs are taught, what are the possible missing gaps and difficulties faced, what are the benefits, or what is the material available. Also, it remains unclear whether scholars teach what is actually needed by industry. We report on three initiatives we have conducted with scholars, educators, industry practitioners, and students to further understand the connection between SPLs and education, i.e., an online survey on teaching SPLs we performed with 35 scholars, another survey on learning SPLs we conducted with 25 students, as well as two workshops held at the International Software Product Line Conference in 2014 and 2015 with both researchers and industry practitioners participating. We build upon the two surveys and the workshops to derive recommendations for educators to continue improving the state of practice of teaching SPLs, aimed at both individual educators as well as the wider community. Finally, we are developing and maintaining a repository for teaching SPLs and variability. Publication at SPLC (journal first) [29], workshop SPLTea'18 http://spltea.irisa.fr/ and repository: https://teaching.variability.io

6.1.3. Variability and machine learning

We propose the use of a machine learning approach to infer variability constraints from an oracle that is able to assess whether a given configuration is correct. We propose an automated procedure to generate configurations, classify them according to the oracle, and synthesize cross-tree constraints. Specifically, based on an oracle (e.g. a runtime test) that tells us whether a given configuration meets the requirements (e.g. speed or memory footprint), we leverage machine learning to retrofit the acquired knowledge into a variability model of the system that can be used to automatically specialize the configurable system. We validate our approach on a set of well-known configurable software systems (Apache server, x264, etc.) Our results show

that, for many different kinds of objectives and performance qualities, the approach has interesting accuracy, precision and recall after a learning stage based on a relative small number of random samples. Publications: Temple et al. *Towards Adversarial Configurations for Software Product Lines*https://arxiv.org/abs/1805.12021, VaryLaTeX [30] a variability and learning-based tool to generate relevant paper variants written in latex.

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

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

6.2. Results on Software Language Engineering

6.2.1. Omniscient Debugging for Executable DSLs

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

6.2.2. Trace Comprehension Operators for Executable DSLs

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

6.2.3. Model Transformation Reuse across Metamodels

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

6.2.4. Modular Language Composition for the Masses

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

6.2.5. Shape-Diverse DSLs

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

6.2.6. Fostering metamodels and grammars

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

6.2.7. Automatic Production of End User Documentation for DSLs

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

6.3. Results on Heterogeneous and dynamic software architectures

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

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

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

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

6.3.2. A Temporal Model for Interactive Diagnosis of Adaptive Systems

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

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

on a cloud elasticity manager case study that such temporal action-aware context leads to improved reasoners while still be highly scalable.

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

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

6.4. Results on Diverse Implementations for Resilience

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

6.4.1. Privacy and Security

6.4.1.1. FP-STALKER: Tracking Browser Fingerprint Evolutions

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

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

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

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

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

6.4.2. Software Testing

6.4.2.1. A Comprehensive Study of Pseudo-tested Methods

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

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

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

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

6.4.2.3. Multimorphic Testing

The functional correctness of a software application is, of course, a prime concern, but other issues such as its execution time, precision , or energy consumption might also be important in some contexts. Systematically testing these quantitative properties is still extremely difficult, in particular, because there exists no method to

tell the developer whether such a test set is "good enough" or even whether a test set is better than another one. This work [41] proposes a new method, called Multimorphic testing, to assess the relative effectiveness of a test suite for revealing performance variations of a software system. By analogy with mutation testing, our core idea is to vary software parameters, and to check whether it makes any difference on the outcome of the tests: i.e. are some tests able to " kill " bad morphs (configurations)? Our method can be used to evaluate the quality of a test suite with respect to a quantitative property of interest, such as execution time or computation accuracy.

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

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

EASE Team

6. New Results

6.1. Smart City and ITS

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

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

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

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

6.2. Opportunistic and local communication/information sharing

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

Smart spaces (Smart-city, home, building, etc.) are complex environments made up of resources (cars, smartphones, electronic equipment, applications, servers, flows, etc.) that cooperate to provide a wide range of services to a wide range of users. They are by nature extremely fluctuating, heterogeneous, and unpredictable. In addition, applications are often mobile and have to migrate or are offered by mobile platforms such as smartphones or vehicles. To be relevant, applications must be able to adapt to users by understanding their environment and anticipating its evolutions. Communication between devices and information sharing is a key to achieve this goal. In recent years, many products have been developed based on the cloud. This raises privacy and network access issues.We believe that communication and information sharing should be able to take place directly when possible, more efficient, confidential, or when the network is not available. To achieve this, applications must be provided with technologies that enable the opportunistic and rapid exchange of information based on simple and widespread technologies.

Applications such as pervasive games (for ex. Pokemon Go), on the go data sharing, collaborative mobile app are often good candidates for opportunistic or dynamic interaction models. But they are not well supported by existing communication stacks, especially in context involving multiple technologies. Technological heterogeneity is not hidden, and high level properties associated with the interactions, such as proximity/range, or mobility-related parameters (speed, discovery latency) have to be addressed in an ad hoc manner.

We think that a good way to solve these issues is to offer an abstract interaction model that could be mapped over the common proximity communication technologies, in a similar way as MOM (Message Oriented Middleware) such as MQTT abstract communications in many IoT and pervasive computing scenarios. However, they typically requires IP level communication, which far beyond the capabilities of ultra low energy proximity communication such as RFID and BLE. Moreover, they often rely on a coordinator node that is not adapted in highly dynamic context involving ephemeral communications and mobile nodes.

To ease communication, we developed an opportunistic communication system that does not need any connection between participants, nor any preexisting infrastructure (e.g. WiFi network). The only condition for participants to exchange information is that they are close enough to each other. The communication protocol has been implemented over Bluetooth Low Energy advertisement packets. This protocol has been ported to ESP8266, ESP32 and Android platform. To ease information sharing, we started the implementation of an associative memory mechanism over BLE, as it is a common ground that can be shared with passive or semi passive communications (RFID, NFC). Such mechanism, although relatively low level, is still a very useful building block for opportunistic applications: it enables opportunistic data storage/sharing and signaling/synchronization (in space in particular). This approach is fully in line with more general trend developed in the team to build smart systems leveraging local resources and data oriented mediation. The communication protocol has been extended to allow REST-like operations. The implementation in C of the protocol and a storage base was done in such a way as to take little memory and run on small chips (ESP8266, ESP32). The storage base can be accessed either opportunistically using the BLE protocol or via a COAP protocol for longer or bigger exchanges.

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

This should lead to publishing on both infrastructure and application level aspects of the approach.

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

Participants: Alexandre Rio, Yoann Maurel [contact].

This work began in 2017 and is carried out as part of a broader collaboration between EASE, the Diverse Team and OKWind, a company specialized in the production of renewable sources of energy. OKWind proposes to deploy self-production units directly where the consumption . It has developed expertise in vertical-axis wind turbines,photovoltaic trackers, heat pump and energy storage devices. An interesting aspect of renewable energies is that they can be produced locally, close to the consumers, thus considerably reducing infrastructures and distribution costs.The autonomy of sites with micro-generation capabilities is then greatly increased by self-consumption of locally produced energy.

Designing solutions in favor of self-consumption for small industries or city districts is challenging. It consists in designing an energy production system made of solar panels, wind turbines, batteries that fit the annual weather prediction and the industrial or human activity. This raises several issues. How to precisely assess the consumption and production of energy on a given site with changing conditions? How to adequately size energy sources and energy storage (wind turbine, solar panel and batteries)? What methods to use to optimize consumption and, whenever possible, act on installations and activities to reduce energy costs?

We aim to design an integrated tool-suite to assist the engineers in dimensioning an Energy Management System (EMS) for an isolated site to reduce the construction of new network infrastructure and reduce its dependence on the grid. We advocate that the MDE is a very good candidate to integrate the various technological and business knowledge on the renewable energy production and consumption forecasting techniques, the planning of processes, energy costs, grid, and batteries. The development of a DSL to describe the relationships between activities, their planning, and the production and environmental factors would make possible to simulate a given site at a given location, to make assumptions on sizing, and would be a basis to forecast energy consumption so as to provide recommendations for the organization of activities. Using a DSL and components that clearly separate the different concerns would avoid code redundancies and would facilitate the work of domain experts.

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

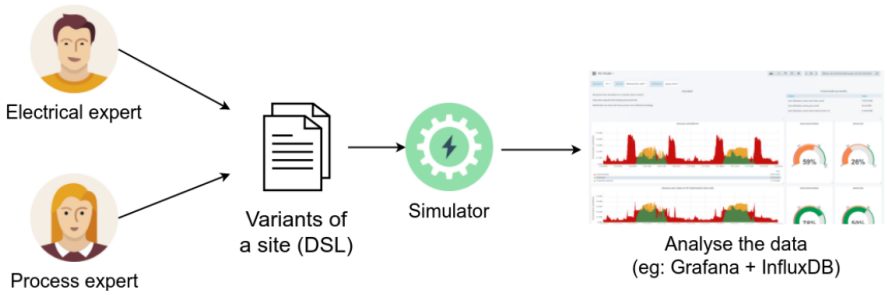

Figure 3. Experts express their concerns using the same DSL and can simulate various scenarios

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

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

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

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

6.4. Location assessment from local observations

Participants: Yoann Maurel, Paul Couderc [contact].

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

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

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

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

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

6.5. Introducing Data Quality to the Internet of Things

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

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

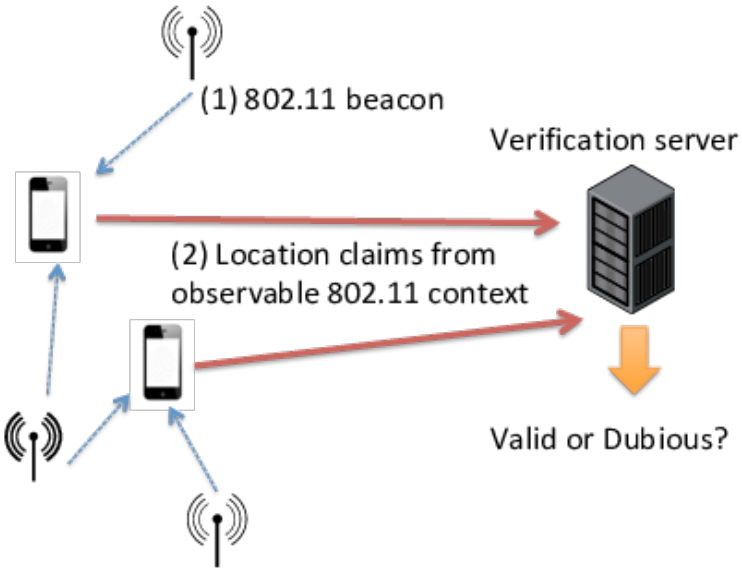

Figure 4. Location assessment from local observations: architecture

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

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

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

6.6. Risk Monitoring and Intrusion Detection

Participant: Jean-Marie Bonnin [contact].

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

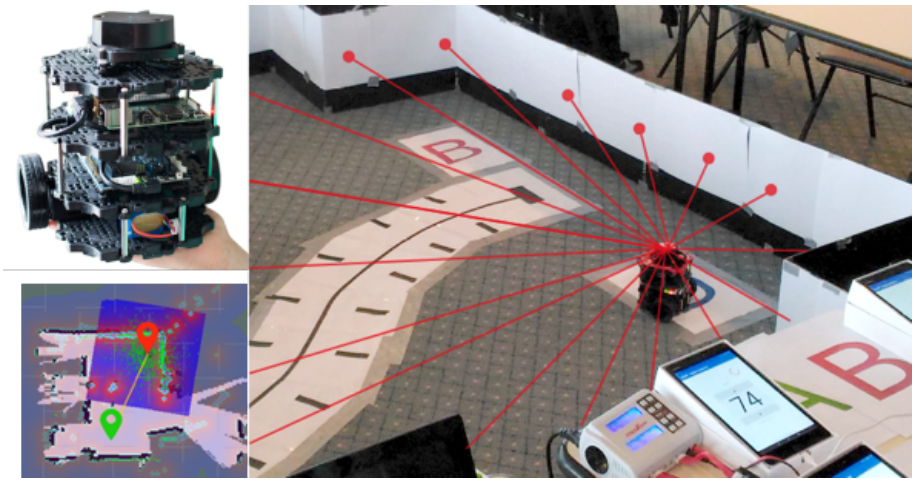

Figure 5. Demonstration at adhoc now 2018

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

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

6.7. Secure design of WoT services for Smart Cities

Participant: Jean-Marie Bonnin [contact].

The richness and the versatility of WebRTC, a new peer-to-peer, real-time and browser- based communication technology, allowed the imagination of new and innovative services. We analyzed the capabilities required to allow a participant in a WebRTC session to access the smart things belonging to his own environment as well as those of any other participant in the same session. The access to such environment (a Smart Space (SS)) can be either passive, for example by monitoring the contextual information provided by the sensors, or active by

requesting the execution of commands by the actuators, or a mixture of both. This approach deserves attention because it allows to solve in an original way various issues such as allowing experts to remotely exercise and provide their expertises. From a technical point of view the issue is not trivial because it requires a smooth and mastered articulation between two different technologies: WebRTC and the Internet of Things (IoT) / Web of Things (WoT) [6].

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

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

FOCUS Project-Team

6. New Results

6.1. Service-Oriented Computing

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

6.1.1. Orchestrations and choreographies

The practice of programming distributed systems is extremely error-prone, due to the complexity in correctly implementing separate components that, put together, enact an agreed protocol. Usage of contracts and session types in orchestration languages guarantees correct communication [22]. Asynchronous subtyping for binary session types has been thought to be decidable for 8 years, before we proved in previous work it to be undecidable. We have now highlighted some practically-relevant fragments of session types where asynchronous subtyping is indeed decidable [14].

We also studied practical aspects of choreographies, that is multiparty contracts. On the one hand we extended the classical proof of correctness of the projection of a choreography on one participant, which is the cornerstone of the theory of choreographies. Indeed, the classical proof considers projection to a model based on channels in the CCS style, while practice mainly relies on correlation sets. We bridged this gap by giving a correctness proof towards a real-world execution model based on correlation sets [31].

We then studied how to apply choreographies for cross-organizational system integration. More precisely, we proposed a software development process to build integrations composed by distributed, independent connectors whose global behaviour is correct by construction [30]. Choreographies and choreography projection are at the heart of the proposed development process.

6.1.2. Microservices

We continued the study of microservice-oriented computing started in past years, in particular by using our microservice-oriented language Jolie. We focused, in particular, on the use of Jolie in an Internet of Things (IoT) setting [28]. Technically, a key feature of Jolie is that it supports in a uniform way multiple service-oriented communication protocols such as HTTP and SOAP. We extended Jolie in order to support, uniformly as well, also lightweight protocols such as MQTT and CoAP, which are largely used in IoT. These are very different from service-oriented protocols which are point-to-point and based on TCP since MQTT is publish-subscribe while CoAP is based on UDP.

6.2. Models for Reliability

Participant: Ivan Lanese.

6.2.1. Reversibility

We have continued the study of reversibility started in the past years, concentrating on contracts and debugging, and applying the results related to debugging to the Erlang programming language. Concerning contracts, in [12] we further studied the retractable contracts that we defined in previous work. The main novelty consists in showing how retractable contracts can be obtained by taking standard contracts, making them reversible using the general approach presented by Phillips and Ulidowski [44], and then applying suitable control policies.

Concerning debugging, we highlighted in [18] the general approach that can be used to build a causal-consistent reversible debugger for a given language. We then instantiated this general approach to a relevant subset of the language Erlang, first defining uncontrolled and controlled reversible semantics for it [16], and then building an actual causal-consistent reversible debugger called CauDEr [32].

6.3. Probabilistic Systems and Resource Control

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

6.3.1. Probabilistic Rewriting and Computation

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 sized-type system. In 2018, Focus has produced another work along these lines, based on intersection types [23]. In the usual, pure, lambda-calculus, various notions of terminating terms can be characterised by way of intersection types, in such a way that the class of terminating terms *coincides* with the one of typable terms. The presence of probabilistic choices together with the aforementioned recursion theoretical limitations prevents the same scenario to happen in probabilistic lambda-calculi, i.e., lambda-calculi endowed with some form of probabilistic choice. Nevertheless, Breuvart and Dal Lago proved that capturing the probability of termination in an approximate way by means of intersection types is indeed possible [23].

In 2018, we have also been active in laying out a novel foundation for *probabilistic abstract reduction systems* (*probabilistic ARSs*). ARSs constitute a general framework to study fundamental properties of computations, such as termination or confluence. These properties are intricately related to the well-definedness of functions, and consequently, play key roles in the formal study of programming languages. Specifically, in collaboration with Yamada, Avanzini and Dal Lago [20] introduced a new notion of probabilistic computation by means of a reduction relation over multidistributions. This relation enables the seamless combination of non-deterministic and probabilistic choice, thereby, considerably simplifying earlier notions of reduction semantics by means of schedulers and Markov chains. On top of this, a partially flawed characterisation of positive almost sure termination by means of Lyapunov ranking functions, initially due to Bournez and Garnier, could be clarified. Moreover, the *interpretation method*, which is maybe the most fundamental technique to investigate termination and runtime complexity of term rewrite systems, could be lifted to probabilistic systems.

Finally, we have been able to propose a novel and natural way of giving the reduction semantics for Markovian process algebras [13], a model of concurrent interaction which is particularly appropriate to the performance analysis of concurrent systems.

6.3.2. Complexity Analysis of Functional Programs

A research topic which lies at the core of Focus since its inception is the complexity analysis of functional programs, through tools like implicit complexity and linear logic. During 2018, we have published an extended version of a paper in which we proved that the most general form of ramified recursion, a key tool in term rewriting, remains sound for polynomial time computation, although requiring some nontrivial machinery based on sharing and memoisation [11]. We have also started to investigate along a new and promising research direction concerning the efficient implementation of functional programming languages through randomised strategies. We have shown that even the simplest strategy is nontrivial for the pure, untyped lambda-calculus [25], being in certain cases more efficient than both innermost and outermost strategies.

6.3.3. Reasoning About Effectful and Concurrent Programs

Pure functional programs are relatively easy to reason about due to referential transparency, a property inherent to functional programs that renders their semantics close to the one of mathematical expressions.

Quite recently, functional programming has found its way into main stream programming languages. Thus functional programming is *combined* with various forms of computational effects, such as exceptions, state, or even nondeterministic choice. Since a couple of years, we are interested in studying the impact of effects on the metatheory of functional programming languages, with an eye to coinductive methodologies akin to those employed in concurrency, coinduction *in primis*. In 2018, Francesco Gavazzo has extended some of our previous work about a generic approach to behavioural equivalences for higher-order effectual languages to *metrics*, themselves a much more natural way to compare programs in many cases (e.g., in the presence of probabilistic choice). His contribution has been published in the top conference in logic in computer science in 2018 [29].

6.4. Verification Techniques

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

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

6.4.1. Deadlock detection

We have continued the work on deadlock detection of previous years, on languages of concurrent objects. Thus in [33] we have applied and refined previous techniques so to handle multi-threaded programs with reentrant locks. For this we have defined a simple calculus featuring recursion, threads and synchronizations that guarantee exclusive access to objects. We detect deadlocks by associating an abstract model to programs and we define an algorithm for verifying that a problematic object dependency (e.g. a circularity) between threads will not be manifested.

In [15] we give two different notions of deadlock for systems based on active objects and futures. One is based on blocked objects and conforms with the classical definition of deadlock. The other one is an extended notion of deadlock based on blocked processes which is more general than the classical one. We introduce a technique to prove deadlock freedom in which an abstract version of the program is translated into Petri nets. Extended deadlocks, and then also classical deadlock, can be detected via checking reachability of a certain forms of marking.

6.4.2. Proof techniques based on unique solutions

We study bisimilarity, a behavioural equivalence whose success is much due to the associated bisimulation proof method. In particular, we discuss different proof methods, based on unique solution of equations or of special forms of inequations called contractions, and inspired by Milner's theorem on unique solution of equations. The techniques are at least as powerful as the bisimulation proof method and its up-to context enhancements. The techniques can be transferred onto other behavioural equivalences, possibly contextual and non-coinductive. This enables a coinductive reasoning style on such equivalences. An overview paper on these techniques is [19].

The paper [36] discusses a rather comprehensive formalisation of the core of the theory of CCS in the HOL theorem prover (HOL4), with a focus towards the theory of unique solutions of contractions. (The formalisation consists of about 20,000 lines of proof scripts in Standard ML.) Some refinements of the theory itself are obtained. In particular we remove the constraints on summation, which must be weakly-guarded, by moving to rooted contraction, that is, the coarsest precongruence contained in the contraction preorder.

In [26] we apply the above techniques to study Milner's encoding of the call-by-value λ-calculus into the π-calculus. We show that, by tuning the encoding to two subcalculi of the π-calculus (Internal π and Asynchronous Local π), the equivalence on λ-terms induced by the encoding coincides with Lassen's eager normal-form bisimilarity, extended to handle η-equality. As behavioural equivalence in the π-calculus we consider contextual equivalence and barbed congruence. We also extend the results to preorders.

On a different, but related, strand of work [17], we study the tree structures that result when writing call-by-name functions as processes, and give general conditions under which this representation produces Lévy-Longo Trees and Böhm Trees, the best known tree structures on the lambda-calculus.

6.5. Computer Science Education

Participants: Michael Lodi, Simone Martini.

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

6.5.1. Computational thinking and constructionism

In the last ten years, the expression “computational thinking” has been used to talk about the introduction of CS in K-12 education. The expression was originally used in the 1980s by Seymour Papert, a pioneer in Math education using programming (he is the principal inventor of the LOGO programming language). We analysed [37] the original context in which the expression originated: the constructionist learning theory, that promotes an active way of learning by constructing meaningful computational artifacts. Papert aimed to teach Math and Physics, but we think CS too is a breeding ground for applying constructionist practices like creative learning, iterative and incremental development, learning by doing, learning by trial and error, project-based learning [35].

6.5.2. CS in the school curriculum

As there is no established practice in teaching CS, academics should facilitate the introduction of CS principles in the school curriculum, to avoid misconceptions and to focus mainly on scientific principles, rather than on technical aspects. Within a CINI (Italian National Interuniversity Consortium for Informatics) group, we designed a proposal [27] for CS teaching in Italian K-10 schools, that focuses on CS principles, and gives space to the use of digital technologies only as tools for self-expression through computation. When introducing a new discipline, often misconceptions arise. In a large sample of primary teachers, we investigate [38], [24] the ideas about the “buzzword” *coding*, that is more and more used to talk about CS at school. Only 60% of teachers correctly linked “coding” to “programming” (some of them implicitly), and many misconceptions (e.g. “coding is only for children”, or “coding is the transversal use of computational thinking at school”, “programming is only for professionals”) were found. After defining a curriculum, one should also provide some materials to concretely teach the discipline and ensure learning objectives will be achieved. We presented [21] the structure of a nationwide initiative by the Italian Ministry of Education: Problem Solving Olympics (OPS). Preliminary analysis of students’ results in the last five editions suggests the competition fosters learning of computational thinking knowledge and skills.

6.5.3. Growth mindset and teacher training

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. Teachers’ mindset directly affects students’ one. By contrast, applying constructionists approaches seems to foster a growth mindset. In facts, we found a statistically significative, albeit little, increase of pre-service primary teacher’s growth mindset after a “creative computing and computational thinking” course [34].

6.6. 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 this area a *portfolio solver* combines a variety of different constraint solvers for solving a given problem. This fairly recent approach enables to significantly boost the performance of single solvers, especially when multicore architectures are exploited. In [10] we give a brief overview of the portfolio solver sunny-cp, and we discuss its performance in the MiniZinc Challenge —the annual international competition for CP solvers —where it won two gold medals in 2015 and 2016.

INDES Project-Team

5. New Results

5.1. Information Flow Security

We have pursued our study on information flow security policies and enforcements. We have followed two main axes.

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

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

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

This study is described in [16].

A Better Facet of Dynamic Information Flow Control

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

In this year, we have published a paper [15], where we have extended Multiple Facets in three directions. First, we propose a new version of MF for arbitrary lattices, called Generalised Multiple Facets, or GMF. GMF strictly generalizes MF, which was originally proposed for a specific lattice of principals. Second, we propose a new optimization on top of GMF that further reduces the number of executions. Third, we strengthen the security guarantees provided by Multiple Facets by proposing a termination sensitive version that eliminates covert channels due to termination.

5.2. JavaScript Implementation

We have pursued the development of Hop.js and our study on efficient JavaScript implementation. We have followed three main axes.

Implementing Hop.js

Hop.js supports full ECMAScript 5 but it still lack many of the new features ECMAScript 2016 has introduced and now that are now well established in ECMAScript 2017. During the year, we have implemented many of these features (iterators, destructuring assignments, modules, etc.). Few constructs remain missing and hopefully will be added to the system by the end of the year (map, set, and proxies). Completing full ECMAScript 2017 is important as we now see more and more packages using these new features being made available and we consider that maintaining the ability to use all these publicly available resources is a prerequisite to a wide Hop.js adoption. We also consider that this is an important asset for Hop.js users, in particular, for the Denimbo company, an Inria startup using Hop.js extensively.

Ahead-of-time JavaScript compilation

Hop.js differs from most JavaScript implementations by many aspects because contrary to all fast and popular JavaScript engines that use just-in-time compilation, Hop.js relies on static compilation, *a.k.a.,* ahead-of-time (AOT) compilation. It is an alternative approach that can combine good speed and lightweight memory footprint, and that can accommodate read-only memory constraints that are imposed by some devices and some operating systems. Unfortunately the highly dynamic nature of JavaScript makes it hard to compile statically and all existing AOT compilers have either gave up on good performance or full language support.

Indeed, JavaScript is hard to compile, much harder than languages like C, Java, and even harder than Scheme and ML two other close functional languages. This is because a JavaScript source code accepts many more possible interpretations than other languages do. It forces JavaScript compilers to adopt a defensive position by generating target codes that can cope with all the possible, even unlikely, interpretations because general compilers can assume very little about JavaScript programs. The situation is worsened further by the *raise as little errors as possible* principle that drives the design of the language. JavaScript functions are not required to be called with the declared number of arguments, fetching an unbound property is permitted, assigning undeclared variables is possible, etc.

All these difficulties are considered serious enough to prevent classic static compilers to deliver efficient code for a language as dynamic and as flexible as JavaScript. We do not share this point of view. We think that by carefully combining classical analyses, by developing new ones when needed, and by crafting a compiler where the results of the high-level analyses are propagated up to the code generation, it is possible for AOT compilation to be in the same range of performance as fast JIT compilers. This is what we attempt to demonstrate with this study. Of course, our ambition is not to produce a compiler strictly as fast as the fastest industrial JavaScript implementations. This would require much more engineering strength than what we can afford. Instead, we only aim at showing that static compilation can have performances reasonably close to those of fastest JavaScript implementations. *Reasonably close* is of course a subjective notion, that everyone is free to set for himself. For us, it means a compiler showing half the performances of the fastest implementations.

The version of the Hop.js AOT compiler we have developed during the year contains new typing analyses and heuristics that compensate for the lack of information JavaScript source codes contain. A first analysis, named *occurrence typing*, that elaborates on top of older techniques developed for the compilation of the Scheme programming language, extracts as much as possible syntactic information directly out of the source code. This analysis alone would give only rough approximations of the types used by the program but its main purpose is to feed the compiler with sufficient information so that it can deploys more efficient supplemental analyses. Probably the most original one is the analysis that we have named *hint typing* or *which typing* that consists in assigning types to variables and to function arguments according to the efficiency of the generated code. In other words, the *which typing* assign types for which the compiler will be able to deliver its best code instead of assigning types that might denote all the possible values variables and arguments may have during all possible executions. We have shown that these *whiched* types correspond very frequently to the implicit *intentional* types programmers had in mind when they wrote their programs. These analyses and the optimizations they enable are implemented in Hop.js version 3.2.0 available on the Inria pages and from Github. They are described in [17] paper.

Property caches: Property caches are a well-known technique invented over 30 years ago to improve dynamic object accesses. They have been adapted to JavaScript, which they have greatly contributed to accelerate.

However, this technique is applicable only when some constraints are satisfied by the objects, the properties, and the property access sites. We have started a study to try to improve it on two common usage patterns: *prototype accesses* and *megamorphic accesses*. We have built a prototypical implementation in Hop.js that has let us measure the impact of the technique we propose. We have observed that they effectively complement traditional caches and that they reduce cache misses and consequently accelerate execution. Moreover, they do not cause a slowdown in the handling of the other usage patterns. We are now at completing this study by polishing the implementation and by publishing a paper exposing and evaluating the new techniques.

5.3. Web Reactive Programming

During the year, we have continued our effort in designing and implementing the HipHop.js programming language, we have applied it to interactive music composition, and we have studied security of reactive systems.

HipHop.js

Web applications react to many sort of events. Let them be GUI events, multimedia events, or network events on client code or IO and system events on the server, they are all triggered asynchronously. JavaScript, the hegemonic programming language of the Web, handles them using low level constructs based on *listeners*, a synonym for *callback*. To improve on the so-called *callback hell*, the recent versions of the language have proposed new constructs that raise the programming abstraction level (promises and `async/await`). They enable a programming style, closer to traditional sequential programming, which helps developing and maintaining applications. However, the improvements they propose rely exclusively on syntactic extensions. They do not change the programming model. For that reason, complex orchestration problems that imply all sorts of synchronization, preemption, and parallelism remain as complex to program as before. We think that orchestration should be reconsidered more globally and from the ground. The solution we propose consists in embedding a DSL specialized on orchestration inside the traditional Web development environment, in our case, Hop.js, the Web programming language that the team develop.

The orchestration DSL we propose is called HipHop.js. It is a reactive synchronous language. More precisely, it is an adaptation of the Esterel programming language to the Web. The motivations for choosing Esterel are diverse. First, and most important, Esterel is powerful enough to handle all the orchestration patterns we are considering. Second, the team, via its partnership with Colège de France, has high expertise in the design and development of Esterel-like languages, which constitutes a highly valuable asset for our development.

Esterel is powerful enough to handle all the orchestration patterns we are considering but Esterel has been designed and developed in a context baring no resemblance with the Web. Esterel was considering static execution models while the Web assumes permanent evolutions and modifications of the running programs. Esterel was considering sequential imperative languages for its embedding, while the Web is considering dynamic functional languages (i.e., JavaScript). Esterel was assuming static execution contexts where *a-priori* validity proof were enforced before hand while the Web assumes highly dynamic runtime executions so that only dynamic verifications are doable. For all these reasons, adapting Esterel and transforming it to form HipHop.js has needed a deep revamping and a deep paradigm shift.

During the year of 2018, we have finalized and completed the design of the language that is now almost stabilized. It follows previous version developed in C. Vidal's PhD studies [13], [18]. The version 0.3.x has been made available at the URL http://hop-dev.inria.fr. It has been used to implement our first orchestration demanding applications, in particular, an interactive music composition application. Our next steps will consist in completing the design and implementation of the language and a minimal development environment without which only experts can use the system. We of course also need to publicize the system and describe its design and internal in various academic publications.

Interactive music composition: the Skini platform

In the sixties, the philosopher Umberto Eco, and musicians such as K. Stockausen, K. Pendereski, L. Berio questioned about the relationship between composers, musicians, and the way we perceive music. Eco used the wording "Open Work", and showed that, the vision of the world evolved from a static world to a more

blurred perception. According to this new perception and in a shift comparable to the evolution of physics from Copernic to Einstein, some contemporary artists tried to express this complexity through works where the performer and the audience have a concrete impact on the work. Since the sixties, the development of audience participation for collaborative music production has become a more and more active field. Thanks to the large device market and web based technology development such as web audio API, "Open Work" got a broader meaning with systems allowing individual interaction. Nevertheless it is still difficult to find systems proposing frameworks dedicated to music composition of interactive performances with a clear composition scheme and ease of use. This is our motivation for developing a framework, called Skini, designed for composing, simulating, and executing interactive performances. Skini is based on elementary music patterns, automatic control of the patterns activation made possible thanks to Hop and Hiphop. Skini was first used for a concert that took place at the very end of 2017 in the contemporary Musical Festival of Nice (MANCA) followed in 2018 by performances during the "Portes ouvertes" of Inria, the "Fête de la Science" and the Synchron conference.

In 2018: The Skini's user interface has been revamped. We have tried several interfaces for the pattern activation and focused on a simple one in order to make the interface more intuitive and fluid. We have added an important feature called the "distributed sequencer", which allows the audience not only to activate patterns but also to create them. We have added a new level of interaction, the scrutator, which allows global actions by the audience on the orchestration. The complete system is now synchronized with an external Midi clock. We have developed a first version of stand alone Midi control of the pattern. The system has followed the evolution of the Hiphop syntax and now implements the last version for the control of orchestration. We improved the synchronisation system and the processes for implementing the orchestration.

5.4. Session Types

Session types describe communication protocols between two or more participants by specifying the sequence of exchanged messages, together with 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 now been embedded into a range of functional, concurrent, and object-oriented programming languages.

We have pursued our work on session types along three main directions.

Multiparty Reactive Sessions

Ensuring that communication-centric systems interact according to an intended protocol is an important but difficult 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 reactive systems. *Session-based concurrency* is the model of concurrent computation induced by session types, a rich typing discipline designed to specify the structure of interactions.

In this work, we propose a process calculus for multiparty sessions enriched with features from SRP. In this calculus, protocol participants may broadcast messages, suspend themselves while waiting for a message, and also react to events.

Our main contribution is a session type system for this calculus, which enforces session correctness in terms of communication safety and protocol fidelity, and also ensures a time-related property, which we call input timeliness, which entails livelock-freedom. Our type system departs significantly from existing ones, specifically as it captures the notion of "logical instant" typical of SRP. This work is currently under submission.

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 has the ability to roll back to a past state. Allowing computations to reverse is a means to improve system flexibility and reliability. In the setting of concurrent process calculi, reversible computations have been first studied for CCS, then for the π-calculus, and only recently for session calculi.

Following up on our previous work on concurrent reversible sessions [29], 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 the choice. This operator is inspired by the notion of *connecting action* recently introduced by Hu and Yoshida to describe protocols with optional participants. We argue that this choice operator allows for a natural description of typical communication protocols. Our calculus also 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. We present a session type system for the calculus and show that it enforces the expected properties of session fidelity, forward progress and backward progress. This work has been accepted for journal publication.

Multiparty sessions with Internal Delegation

We have investigated a new form of delegation for multiparty session calculi. Usually, delegation 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 sessions. As a consequence, properties such as deadlock-freedom or lock-freedom are difficult to ensure in the presence of delegation. Here 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 (internal delegation) and may be captured by its global type. To increase flexibility in the use of delegation, our calculus uses connecting communications, which allow optional participants in the branches of choices. By these means, we are 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 is under submission.

5.5. Measurement and Detection of Web Tracking

Detecting Web Trackers via Analyzing 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 [23] and EasyPrivacy [24] (EL&EP) are the most popular publicly maintained blacklist of know advertising and tracking domains, used by the popular browser extensions AdBlock Plus [20] and uBlockOrigin [28]. Disconnect [22] is another very popular list for detecting domains known for tracking, used in Disconnect browser extension [21] and in integrated tracking protection of Firefox browser [25]. 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.

In this work, to detect trackers, we propose a new technique based on the analysis of invisible pixels [0]. These images are routinely used by trackers in order 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 HTML tag into a webpage. Since invisible pixels do not provide any useful functionality, we consider them *perfect suspects for tracking*.

[0] By "invisible pixels" we mean 1x1 pixel images or images without content.

By using an Inria cluster and setting up a distributed crawler, we have collected a dataset of invisible pixels from 829,349 webpages. By analyzing this dataset, we observed that invisible pixels are widely used: more than 83% of pages incorporate at least one invisible pixel.

Overall, we made the following key contributions:

- We define a new classification of Web tracking behaviors based on the analysis of invisible pixels. By analyzing behavior associated to the delivery of invisible pixels, we propose a new fine-grained classification of tracking behaviors, that consists of 8 categories of tracking. To our knowledge, *we are the first to analyse tracking behavior based on invisible pixels that are present on 83% of the webpages*.
- We apply our classification to a full dataset and uncover new collaborations between third-party domains. We detect new relationships between third-party domains beyond basic cookie syncing detected in the past. In particular, we discovered that *first to third party cookie syncing* is the most prevalent tracking behavior performed by 50,812 distinct domains. Finally, we find that 76.23% of requests responsible for tracking originate from loading other resources than invisible images. To our knowledge, *we are the first to discover a highly prevalent first to third party syncing behavior detected on 51.54% of all crawled domains*.
- We show that the consumer protection lists cannot be considered as ground truth to identify trackers. We find out that the browser extensions based on EasyList and EasyPrivacy (EL&EP) and Disconnect each miss 22% of tracking requests we detect. Moreover, if we combine all the lists, 238,439 requests originated from 7,773 domains are unknown to these lists and hence still track users on 5,098 webpages even if tracking protection is installed. We also detect instances of cookie syncing in domains unknown to these lists and therefore likely unrelated to advertising. To our knowledge, *we are the first to detect that EL&EP and also Disconnect lists used in majority of Web Tracking detection literature are actually missing tracking requests to 7,773 distinct domains.*

This working paper [19] is currently under submission at an international conference.

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.

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.

Measuring Uniqueness of Browser Extensions and Web Logins

Web browser is the tool people use to navigate through the Web, and privacy research community has studied various forms of *browser fingerprinting*. Researchers have shown that a user's browser has a number of inherent "physical" characteristics that can be used to uniquely identify her browser and hence to track it across the Web. Fingerprinting of users' devices is similar to physical biometric traits of people, where only physical characteristics are studied.

Similar to previous demonstrations of user uniqueness based on their behavior, *behavioral characteristics*, such as browser settings and the way people use their browsers can also help to uniquely identify Web users. For example, a user installs web browser extensions she prefers, such as AdBlock, LastPass, or Ghostery to enrich her Web experience. Also, while browsing the Web, she logs in her preferred social networks, such as Gmail, Facebook or LinkedIn. In this work, we study *users' uniqueness* based on their behavior and preferences on the Web: we analyze how unique are Web users based on their *browser extensions and logins*.

In this work, we performed the first large-scale study of user uniqueness based on browser extensions and Web logins, collected from more than 16,000 users who visited our website https://extensions.inrialpes.fr/. Our experimental website identifies installed Google Chrome extensions via Web Accessible Resources. and detects websites where the user is logged in by methods that rely on URL redirection and CSP violation reports. Our website is able to detect the presence of 13K Chrome extensions (the number of detected extensions varied monthly between $12,164$ and $13,931$), covering approximately 28% of all free Chrome extensions [0] . We also detect whether the user is connected to one or more of 60 different websites. Our main contributions are:

- A large scale study on *how unique users are based on their browser extensions and website logins*. We discovered that 54.86% of users that have installed at least one detectable extension are unique; 19.53% of users are unique among those who have logged into one or more detectable websites; and 89.23% are unique among users with at least one extension and one login. Moreover, we discover that 22.98% of users could be uniquely identified by web logins, even if they disable JavaScript.
- We study the privacy dilemma on Adblock and privacy extensions, that is, *how well these extensions protect their users against trackers and how they also contribute to uniqueness*. We evaluate the statement "the more privacy extensions you install, the more unique you are" by analyzing how users' uniqueness increases with the number of privacy extensions she installs; and by evaluating the tradeoff between the privacy gain of the blocking extensions such as Ghostery [26] and Privacy Badger [27].

We furthermore show that browser extensions and web logins can be exploited to fingerprint and track users by only checking a limited number of extensions and web logins. We have applied an advanced fingerprinting algorithm [30] that carefully selects a limited number of extensions and logins. For example, we show that 54.86% of users are unique based on all 16,743 detectable extensions. However, by testing 485 carefully chosen extensions we can identify more than 53.96% of users. Besides, detecting 485 extensions takes only 625ms.

Finally, we give suggestions to the end users as well as website owners and browser vendors on how to protect the users from the fingerprinting based on extensions and logins.

This paper has been published at at WPES international workshop affiliated with ACM CCS 2018 [14].

[0]The list of detected extensions and websites are available on our website: https://extensions.inrialpes.fr/faq.php

PHOENIX-POST Team

6. New Results

6.1. Towards context-aware assistive applications for aging in place via real-life-proof activity detection

Assisted living applications can support aging in place efficiently when their context-awareness is based on a real-life-proof approach to activity detection. Recently, Caroux et al. proposed a new approach to monitoring activities dedicated to older adults, named "activity verification". This approach uses a knowledge-driven framework that draws from the literature on older adults. The purpose of the present study is to address the limitations of this approach by scaling it up and by demonstrating that it is applicable to context-aware assistive applications for aging in place. First, an experimental study was conducted in which this approach was used to monitor a large range of daily activities, for a long period (8 weeks of experimentation) and involving several participants (7 participants). Second, this approach was used to validate two examples of context-aware assisted living applications, via simulation, based on real-life sensor log data. Results showed that the applicability of the "activity verification" approach scales up to a large range of daily activities by extending this approach (with accuracy values ranging between 0.82 and 1.00 depending on the activity of interest). Its inter-participant and intra-participant consistencies were demonstrated. Its limitations were addressed and the applicability to context-aware assistive applications for aging in place running on a dedicated platform was demonstrated.

6.2. Are visual cues helpful for virtual spatial navigation and spatial memory in patients with mild cognitive impairment or Alzheimer's disease?

Objective: To evaluate whether visual cues are helpful for virtual spatial navigation and memory in Alzheimer's disease (AD) and patients with mild cognitive impairment (MCI). Method: 20 patients with AD, 18 patients with MCI and 20 age-matched healthy controls (HC) were included. Participants had to actively reproduce a path that included 5 intersections with one landmark at each intersection that they had seen previously during a learning phase. Three cueing conditions for navigation were offered: salient landmarks, directional arrows and a map. A path without additional visual stimuli served as control condition. Navigation time and number of trajectory mistakes were recorded. Results: With the presence of directional arrows, no significant difference was found between groups concerning the number of trajectory mistakes and navigation time. The number of trajectory mistakes did not differ significantly between patients with AD and patients with MCI on the path with arrows, the path with salient landmarks and the path with a map. There were significant correlations between the number of trajectory mistakes under the arrow condition and executive tests, and between the number of trajectory mistakes under the salient landmark condition and memory tests. Conclusion: Visual cueing such as directional arrows and salient landmarks appears helpful for spatial navigation and memory tasks in patients with AD and patients with MCI. This study opens new research avenues for neuro-rehabilitation, such as the use of augmented reality in real-life settings to support the navigational capabilities of patients with MCI and patients with AD.

6.3. Early detection of mild cognitive impairment with in-home monitoring technologies using functional measures: A systematic review

Introduction: The aging of the world population is accompanied by a substantial increase in neurodegenerative disorders such as dementia. Early detection of dementia, i.e. at the mild cognitive impairment (MCI) stage, could be an essential condition for slowing down the loss of autonomy and quality of life caused by the disease, as it would provide a critical window for the implementation of early pharmacological and non-pharmacological interventions. However, the current assessments for MCI have several limitations. In this

context, approaches involving smart home technologies offer many attractive advantages, including the continuous measurement of functional abilities in ecological environments. Objective: This systematic review aims to investigate the current state of knowledge on the effectiveness of smart home technologies for the early detection of MCI through the monitoring of everyday life activities. Methods: A systematic search of publications in Medline, EMBASE, CINAHL was conducted. Results: Sixteen studies were included in this review. Twelve studies were based on real-life monitoring, with several sensors installed in participants' actual homes, and four studies included scenario-based evaluations in which the participants had to complete various tasks in a research lab apartment. In real-life monitoring, the most used indicators of MCI were walking speed and activity/motion in the house. In scenario-based evaluation, time of completion, quality of activity completion, number of errors, amount of assistance needed, and task-irrelevant behaviors during the performance of everyday activities predicted MCI in participants. Discussion: Despite technological limitations and the novelty of the field, smart home technologies represent a promising potential for the early screening of MCI and could support clinicians in geriatric care.

6.4. A Language for Online State Processing of Binary Sensors, Applied to Ambient Assisted Living

There is a large variety of binary sensors in use today, and useful context-aware services can be defined using such binary sensors. However, the currently available approaches for programming context-aware services do not conveniently support binary sensors. Indeed, no existing approach simultaneously supports a notion of state, central to binary sensors, offers a complete set of operators to compose states, allows to define reusable abstractions by means of such compositions, and implements efficient online processing of these operators. This paper proposes a new language for event processing specifically targeted to binary sensors. The central contributions of this language are a native notion of state and semi-causal operators for temporal state composition including: Allen's interval relations generalized for handling multiple intervals, and temporal filters for handling delays. Compared to other approaches such as CEP (complex event processing), our language provides less discontinued information, allows less restricted compositions, and supports reusable abstractions. We implemented an interpreter for our language and applied it to successfully rewrite a full set of real Ambient Assisted Living services. The performance of our prototype interpreter is shown to compete well with a commercial CEP engine when expressing the same services.

6.5. Implementing a semi-causal domain-specific language for context detection over binary sensors

In spite of the fact that many sensors in use today are binary (i.e. produce only values of 0 and 1), and that useful context-aware applications are built exclusively on top of them, there is currently no development approach specifically targeted to binary sensors. Dealing with notions of state and state combinators, central to binary sensors, is tedious and error-prone in current approaches. For instance, developing such applications in a general programming language requires writing code to process events, maintain state and perform state transitions on events, manage timers and/or event histories. In another paper, we introduced a domain specific language (DSL) called Allen, specifically targeted to binary sensors. Allen natively expresses states and state combinations, and detects contexts on line, on incoming streams of binary events. Expressing state combinations in Allen is natural and intuitive due to a key ingredient: semi-causal operators. That paper focused on the concept of the language and its main operators, but did not address its implementation challenges. Indeed, online evaluation of expressions containing semi-causal operators is difficult, because semi-causal sub-expressions may block waiting for future events, thus generating unknown values, besides 0 and 1. These unknown values may or may not propagate to the containing expressions, depending on the current value of the other arguments. This paper presents a compiler and runtime for the Allen language, and shows how they implement its state combining operators, based on reducing complex expressions to a core subset of operators, which are implemented natively. We define several assisted living applications both in Allen and in a general scripting language. We show that the former are much more concise in Allen, achieve more effective code reuse, and ease the checking of some domain properties.

6.6. Towards Truly Accessible MOOCs for Persons with Cognitive Disabilities: Design and Field Assessment

MOOCs are playing an increasingly important role in education systems. Unfortunately, MOOCs are not fully accessible. In this paper, we propose design principles to enhance the accessibility of MOOC players, especially for persons with cognitive disabilities. These principles result from a participatory design process gathering 7 persons with disabilities and 13 expert professionals. They are also inspired by various design approaches (Universal Design for Learning, Instructional Design, Environmental Support). We also detail the creation of a MOOC player offering a set of accessibility features that users can alter according to their needs and capabilities. We used it to teach a MOOC on digital accessibility. Finally, we conducted a field study to assess learning and usability outcomes for persons with cognitive and non-cognitive impairments. Results support the effectiveness of our player for increasing accessibility.

6.7. Assistive Computing: a Human-Centered Approach to Developing Computing Support for Cognition

The growing population of cognitively impaired individuals calls for the emergence of a research area dedicated to developing computing systems that address their needs. The nature of this research area requires to bridge the many disciplines needed to develop human-centered, assistive computing systems. Such bridging may seem unattainable considering the conceptual and practical gaps between the related disciplines and the challenges of propagating human-related concerns throughout the many stages of the development process of assistive technologies. As a consequence, existing assistive technologies lack a proper needs analysis; their development is often driven by technology concerns, resulting in ill-designed and stereotype-biased systems; and, most of them are not tested for their effectiveness in assisting users. In this paper, we propose a systematic exploration of this vast challenge. First, we decline Assistive Computing as a research area and propose key principles to drive its study. Then, we introduce a tool-based methodology dedicated to developing assistive computing support, integrating a range of disciplines from human-related sciences to computer science. This methodology is purposefully pragmatic in that it leverages, aggregates and revisits numerous research results, concretizing it with a range of examples. More generally, our goal is i) to provide a framework to conduct research in the area of Assistive Computing and ii) to identify the necessary bridges between disciplines to account for all the dimensions of such systems.

RMOD Project-Team

7. New Results

7.1. Dynamic Languages: Virtual Machines

Assessing primitives performance on multi-stage execution. Virtual machines, besides the interpreter and just-in-time compiler optimization facilities, also include a set of primitive operations that the client language can use. Some of these are essential and cannot be performed in any other way. Others are optional: they can be expressed in the client language but are often implemented in the virtual machine to improve performance when the just-in-time compiler is unable to do so (start-up performance, speculative optimizations not implemented or not mature enough, etc.). In a hybrid runtime, where code is executed by an interpreter and a just-in-time compiler, the implementor can choose to implement optional primitives in the client language , in the virtual machine implementation language (typically C or C++), or on top of the just-in-time compiler back-end. This raises the question of the maintenance and performance trade-offs of the different alternatives. As a case study, we implemented the String comparison optional primitive in each case. The paper describes the different implementations, discusses the maintenance cost of each of them and evaluates for different string sizes the execution time in Cog, a Smalltalk virtual machine. [18]

Fully Reflective Execution Environments: Virtual Machines for More Flexible Software. VMs are complex pieces of software that implement programming language semantics in an efficient, portable, and secure way. Unfortunately, mainstream VMs provide applications with few mechanisms to alter execution semantics or memory management at run time. We argue that this limits the evolvability and maintainability of running systems for both, the application domain, e.g., to support unforeseen requirements, and the VM domain, e.g., to modify the organization of objects in memory. This work explores the idea of incorporating reflective capabilities into the VM domain and analyzes its impact in the context of software adaptation tasks. We characterize the notion of a fully reflective VM, a kind of VM that provides means for its own observability and modifiability at run time. This enables programming languages to adapt the underlying VM to changing requirements. We propose a reference architecture for such VMs and present TruffleMATE as a prototype for this architecture. We evaluate the mechanisms TruffleMATE provides to deal with unanticipated dynamic adaptation scenarios for security, optimization, and profiling aspects. In contrast to existing alternatives, we observe that TruffleMATE is able to handle all scenarios, using less than 50 lines of code for each, and without interfering with the application's logic. [2]

7.2. Dynamic Languages: Language Constructs for Modular Design

Dynamic Software Update from Development to Production. Dynamic Software Update (DSU) solutions update applications while they are executing. These solutions are typically used in production to minimize application downtime, or in integrated development environments to provide live programming support. Each of these scenarios presents different challenges, forcing existing solutions to be designed with only one of these use cases in mind. For example, DSUs for live programming typically do not implement safe point detection or instance migration, while production DSUs require manual generation of patches and lack IDE integration. Also, these solutions have limited ability to update themselves or the language core libraries, and some of them present execution penalties outside the update window. We propose a DSU (gDSU) that works for both live programming and production environments. Our solution implements safe update point detection using call stack manipulation and a reusable instance migration mechanism to minimize manual intervention in patch generation. Moreover, it also offers updates of core language libraries and the update mechanism itself. This is achieved by the incremental copy of the modified objects and an atomic commit operation. We show that our solution does not affect the global performance of the application and it presents only a run-time penalty during the update window. Our solution is able to apply an update impacting 100,000 instances in 1 second. In this 1 second, only during 250 milliseconds the application is not responsive. The rest of the time the application runs normally while gDSU is looking for the safe update point. The update only requires to copy the elements that are modified. [6]

Implementing Modular Class-based Reuse Mechanisms on Top of a Single Inheritance VM. Code reuse is a good strategy to avoid code duplication and speed up software development. Existing object-oriented programming languages propose different ways of combining existing and new code such as e.g., single inheritance, multiple inheritance, Traits or Mixins. All these mechanisms present advantages and disadvantages and there are situations that require the use of one over the other. To avoid the complexity of implementing a virtual machine (VM), many of these mechanisms are often implemented on top of an existing high-performance VM, originally meant to run a single inheritance object-oriented language. These implementations require thus a mapping between the programming model they propose and the execution model provided by the VM. Moreover, reuse mechanisms are not usually composable, nor it is easy to implement new ones for a given language. We propose a modular meta-level runtime architecture to implement and combine different code reuse mechanisms. This architecture supports dynamic combination of several mechanisms without affecting runtime performance in a single inheritance object-oriented VM. It includes moreover a reflective Meta-Object Protocol to query and modify classes using the programming logical model instead of the underlying low-level runtime model. Thanks to this architecture, we implemented Stateful Traits, Mixins, CLOS multiple inheritance, CLOS Standard Method Combinations and Beta prefixing in a modular and composable way. [15]

7.3. Software Reengineering

A Reflexive and Automated Approach to Syntactic Pattern Matching in Code Transformations. Empowering software engineers often requires to let them write code transformations. However existing automated or tool-supported approaches force developers to have a detailed knowledge of the internal representation of the underlying tool. While this knowledge is time consuming to master, the syntax of the language, on the other hand, is already well known to developers and can serve as a strong foundation for pattern matching. Pattern languages with metavariables (that is variables holding abstract syntax subtrees once the pattern has been matched) have been used to help programmers define program transformations at the language syntax level. The question raised is then the engineering cost of metavariable support. Our contribution is to show that, with a GLR parser, such patterns with metavariables can be supported by using a form of runtime reflexivity on the parser internal structures. This approach allows one to directly implement such patterns on any parser generated by a parser generation framework, without asking the pattern writer to learn the AST structure and node types. As a use case for that approach we describe the implementation built on top of the SmaCC (Smalltalk Compiler Compiler) GLR parser generator framework. This approach has been used in production for source code transformations on a large scale. We will express perspectives to adapt this approach to other types of parsing technologies. [12]

Relational Database Schema Evolution: An Industrial Case Study. Modern relational database management systems provide advanced features allowing, for example, to include behavior directly inside the database (stored procedures). These features raise new difficulties when a database needs to evolve (e.g. adding a new table). To get a better understanding of these difficulties, we recorded and studied the actions of a database architect during a complex evolution of the database at the core of a software system. From our analysis, problems faced by the database architect are extracted, generalized and explored through the prism of software engineering. Six problems are identified: (1) difficulty in analyzing and visualizing dependencies between database's entities, (2) difficulty in evaluating the impact of a modification on the database, (3) replicating the evolution of the database schema on other instances of the database, (4) difficulty in testing database's functionalities, (5) lack of synchronization between the IDE's internal model of the database and the database actual state and (6) absence of an integrated tool enabling the architect to search for dependencies between entities, generate a patch or access up to date PostgreSQL documentation. We suggest that techniques developed by the software engineering community could be adapted to help in the development and evolution of relational databases. [10]

A Quality-oriented Approach to Recommend Move Method Refactorings. Refactoring is an important activity to improve software internal structure. Even though there are many refactoring approaches, very few consider their impact on the software quality. We propose a software refactoring approach based on quality

attributes. We rely on the measurements of the Quality Model for Object Oriented Design (QMOOD) to recommend Move Method refactorings that improve software quality. In a nutshell, given a software system S, our approach recommends a sequence of refactorings R1,R2,...,Rn that result in system versions S1, S2,..., Sn, where quality (Si+1) > quality (Si). We empirically calibrated our approach, using four systems, to find the best criteria to measure the quality improvement. We performed three types of evaluation to verify the usefulness of our implemented tool, named QMove. First, we applied our approach on 13 open-source systems achieving an average recall of 84.2%. Second, we compared QMove with two state-of-art refactoring tools (JMove and JDeodorant) on the 13 previously evaluated systems, and QMove showed better recall, precision, and f-score values than the others. Third, we evaluated QMove, JMove, and JDeodorant in a real scenario with two proprietary systems on the eyes of their software architects. As result, the experts positively evaluated a greater number of QMove recommendations. [14]

7.4. Dynamic Languages: Debugging

Collectors. Observing and modifying object-oriented programs often means interacting with objects. At runtime, it can be a complex task to identify those objects due to the live state of the program. Some objects may exist for only a very limited period of time, others can be hardly reachable because they are never stored in variables. To address this problem we present Collectors. They are dedicated objects which can collect objects of interest at runtime and present them to the developer. Collectors are non-intrusive, removable code instrumentations. They can be dynamically specified and injected at runtime. They expose an API to allow their specification and the access to the collected objects. We present an implementation of Collectors in Pharo, a Smalltalk dialect. We enrich the Pharo programming and debugging environment with tools that support the Collectors API. We illustrate the use of these API and tools through the collection and the logging of specific objects in a running IOT application. [9]

Rotten Green Tests: a First Analysis. Unit tests are a tenant of agile programming methodologies, and are widely used to improve code quality and prevent code regression. A passing (green) test is usually taken as a robust sign that the code under test is valid. However, we have noticed that some green tests contain assertions that are never executed; these tests pass not because they assert properties that are true, but because they assert nothing at all. We call such tests Rotten Green Tests. Rotten Green Tests represent a worst case: they report that the code under test is valid, but in fact do nothing to test that validity, beyond checking that the code does not crash. We describe an approach to identify rotten green tests by combining simple static and dynamic analyses. Our approach takes into account test helper methods, inherited helpers, and trait compositions, and has been implemented in a tool called DrTest. We have applied DrTest to several test suites in Pharo 7.0, and identified many rotten tests, including some that have been sleeping in Pharo for at least 5 years. [22]

Mining inline cache data to order inferred types in dynamic languages. The lack of static type information in dynamically-typed languages often poses obstacles for developers. Type inference algorithms can help, but inferring precise type information requires complex algorithms that are often slow. A simple approach that considers only the locally used interface of variables can identify potential classes for variables, but popular interfaces can generate a large number of false positives. We propose an approach called inline-cache type inference (ICTI) to augment the precision of fast and simple type inference algorithms. ICTI uses type information available in the inline caches during multiple software runs, to provide a ranked list of possible classes that most likely represent a variable's type. We evaluate ICTI through a proof-of-concept that we implement in Pharo Smalltalk. The analysis of the top-n+2 inferred types (where n is the number of recorded run-time types for a variable) for 5486 variables from four different software systems shows that ICTI produces promising results for about 75% of the variables. For more than 90% of variables, the correct run-time type is present among first six inferred types. Our ordering shows a twofold improvement when compared with the unordered basic approach, i.e., for a significant number of variables for which the basic approach offered ambiguous results, ICTI was able to promote the correct type to the top of the list. [22]

7.5. Blockchain

Ethereum Query Language Blockchains store a massive amount of heterogeneous data which will only grow in time. When searching for data on the Ethereum platform, one is required to either access the records (blocks) directly by using a unique identifier, or sequentially search several records to find the desired information. Therefore, we propose the Ethereum Query Language (EQL), a query language that allows users to retrieve information from the blockchain by writing SQL-like queries. The queries provide a rich syntax to specify data elements to search information scattered through several records. We claim that EQL makes it easier to search, acquire, format, and present information from the blockchain. [7]

SmartInspect: solidity smart contract inspector. Solidity is a language used for smart contracts on the Ethereum blockchain. Smart contracts are embedded procedures stored with the data they act upon. Debugging smart contracts is a really difficult task since once deployed, the code cannot be reexecuted and inspecting a simple attribute is not easily possible because data is encoded. We address the lack of inspectability of a deployed contract by analyzing contract state using decompilation techniques driven by the contract structure definition. Our solution, SmartInspect, also uses a mirror-based architecture to represent locally object responsible for the interpretation of the contract state. SmartInspect allows contract developers to better visualize and understand the contract stored state without needing to redeploy, nor develop any ad-hoc code. [8]

Preliminary Steps Towards Modeling Blockchain Oriented Software Even though blockchain is mostly popular for its cryptocurrency, smart contracts have become a very prominent blockchain application. Smart contracts are like classes that can be called by client applications outside the blockchain. Therefore it is possible to develop blockchain-oriented software (BOS) that implements part of the business logic in the blockchain by using smart contracts. Currently, there is no design standard to model BOS. Since modeling is an important part of designing a software, developers may struggle to plan their BOS. We show three complementary modeling approaches based on well-known software engineering models and apply them to a BOS example. Our goal is to start the discussion on specialized blockchain modeling notations. [13]

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. In this chapter, 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. [21]

STACK Team

7. New Results

7.1. Resource Management

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

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

7.1.1. Geo-distributed Infrastructures

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

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

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

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

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

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

In [17], we discuss the limitations of meta-data management in Fog/Edge infrastructures. A few storage systems have been proposed to store data in those infrastructures. Most of them are relying on a Distributed Hash Table (DHT) to store the location of objects which is not efficient because the node storing the location of the data may be placed far away from the object replicas. In this paper, we propose to replace the DHT by a tree-based approach mapping the physical topology. Servers look for the location of an object by requesting

successively their ancestors in the tree. Location records are also relocated close to the object replicas not only to limit the network traffic when requesting an object, but also to avoid an overload of the root node. We also propose to modify the Dijkstra's algorithm to compute the tree used. Finally, we evaluate our approach using the object store InterPlanetary FileSystem (IPFS) on Grid'5000 using both a micro experiment with a simple network topology and a macro experiment using the topology of the French National Research and Education Network (RENATER). We show that the time to locate an object in our approach is less than 15 ms on average which is around 20% better than using a DHT.

7.1.2. Cloud and HPC convergence

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

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

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

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

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

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

7.2. Programming Support

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

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

7.2.1. Deployment and reconfiguration in the Cloud

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

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

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

7.2.2. Autonomic computing and self-management

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

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

7.3. Energy-aware computing

Participants: Jean-Marc Menaud, Shadi Ibrahim, Thomas Ledoux, Emile Cadorel, Yewan Wang, Jonathan Pastor.

Energy consumption is one of the major challenges of modern datacenters and supercomputers. Our works in Energy-aware computing can be categorized into two subdomains: Software level (SaaS, PaaS) and Infrastructure level (IaaS).

At Software level, we worked on the general Cloud applications architecture and HPC applications.

In particular, in his habilitation thesis [2], Thomas Ledoux shows that dynamic reconfiguration in Cloud computing can provide an answer to an important societal challenge, namely digital and energetic transitions. Unlike current work providing solutions in the lower layers of the Cloud to improve the energy efficiency of data centers, Thomas Ledoux advocates a software eco-elasticity approach on the high layers of the Cloud. Inspired by both the concept of frugal innovation (Jugaad) and the mechanism of energy brownout, he proposes a number of original artifacts – such as Cloud SLA, eco-elasticity in the SaaS layers, virtualization of energy or green energy-aware SaaS applications, etc. – to reduce the carbon footprint of Cloud architectures.

However, by applying Green Programming techniques, developers have to iteratively implement and test new versions of their software, thus evaluating the impact of each code version on their energy, power and performance objectives. This approach is manual and can be long, challenging and complicated, especially for High Performance Computing applications. In [21], we formally introduces the definition of the Code Version Variability (CVV) leverage and present a first approach to automate Green Programming (i.e., CVV usage) by studying the specific use-case of an HPC stencil-based numerical code, used in production. This approach is based on the automatic generation of code versions thanks to a Domain Specific Language (DSL), and on the automatic choice of code version through a set of actors. Moreover, a real case study is introduced and evaluated though a set of benchmarks to show that several trade-offs are introduced by CVV. Finally, different kinds of production scenarios are evaluated through simulation to illustrate possible benefits of applying various actors on top of the CVV automation. While this work takes HPC applications as a use-case the presented automated green programming technique could be applied to any kind of production application onto any kind of infrastructures.

In general, many Big Data processing applications nowadays run on large-scale multi-tenant clusters. Due to hardware heterogeneity and resource contentions, straggler problem has become the norm rather than the exception in such clusters. To handle the straggler problem, speculative execution has emerged as one of the most widely used straggler mitigation techniques. Although a number of speculative execution mechanisms have been proposed, as we have observed from real-world traces, the questions of “when” and “where” to launch speculative copies have not been fully discussed and hence cause inefficiencies on the performance and energy of Big Data applications. In [29], we propose a performance model and an energy consumption model to reveal the performance and energy variations with different speculative execution solutions. We further propose a window-based dynamic resource reservation and a heterogeneity-aware copy allocation technique to answer the “when” and “where” questions for speculative executions. Evaluations using real-world traces show that our proposed technique can improve the performance of Big Data applications by up to 30% and reduce the overall energy consumption by up to 34%.

At infrastructure level, we worked on power and thermal management from server to datacenter. In fact, with the advent of Cloud Computing, the size of datacenters is ever increasing and the management of servers and their power consumption and heat production have become challenges. The management of the heat produced by servers has been experimentally less explored than the management of their power consumption. It can be partly explained by the lack of a public testbed that provides reliable access to both thermal and power metrics of server rooms. In [34], [20], [19] we had describe SeDuCe, a testbed that targets research on power and thermal management of servers, by providing public access to precise data about the power consumption and the thermal dissipation of 48 servers integrated in Grid’5000 as the new ecotype cluster. We presented the chosen software and hardware architecture for the SeDuCe testbed. Future work will focus on two areas: adding renewable energy capabilities to the SeDuCe testbed, and improving the precision of temperature sensors.

If SeDuCe testbed is focused on the management of the power consumption and heat produced by servers at room level, we realized in [30], [24], [23] , studies on power consumption (and heat impact) of physical servers. First, we characterized some potential factors on the power variation of the servers, such as: original fabrication, position in the rack, voltage variation and temperature of components on motherboard. The results show that certain factors, such as original fabrication, ambient temperature and CPU temperature, have noticeable effects on the power consumption of servers. The experimental results emphasize the importance

of adding these external factors into the metric, so as to build an energy predictive model adaptable in real situations.

7.4. Security and Privacy

Participants: Mario Südholt, Mohammad Mahdi Bazm, Fatima-Zahra Boujdad, Jean-Marc Menaud.

This year the team has provided two major contributions on security and privacy challenges in distributed systems. First, we have developed our models and techniques for the detection and mitigation of side-channel attacks. Second, we have provided a first model and implementation techniques for secure and privacy-preserving distributed biomedical analyses, notably genomic ones.

7.4.1. Side-channel attacks and trusted Fog/Edge infrastructures

In [4], we investigate Cloud computing infrastructures, which are based on the sharing of hardware resources among different clients. The infrastructures leverage virtualization to share physical resources among several self-contained execution environments like virtual machines and Linux containers. Isolation is a core security challenge for such a paradigm. It may be threatened through side-channels, created due to the sharing of physical resources like caches of the processor or by mechanisms implemented in the virtualization layer. Side-channel attacks (SCAs) exploit and use such leaky channels to obtain sensitive data like kernel information. We clarify the nature of this threat for cloud infrastructures. Current SCAs are done locally and exploit isolation challenges of virtualized environments to retrieve sensitive information. We also introduce the concept of distributed side-channel attack (DSCA). We explore how such attacks can threaten isolation of any virtualized environments. Finally, we study a set of different applicable countermeasures for attack mitigation in cloud infrastructures.

In [9], we investigate Fog and Edge computing for the provision of large pools of resources at the edge of the network that may be used for distributed computing. Fog infrastructure heterogeneity also results in complex configuration of distributed applications on computing nodes. Linux containers are a mainstream technique allowing to run packaged applications and micro services. However, running applications on remote hosts owned by third parties is challenging because of untrusted operating systems and hardware maintained by third parties. To meet such challenges, we may leverage trusted execution mechanisms. In this work, we propose a model for distributed computing on Fog infrastructures using Linux containers secured by Intel's Software Guard Extensions (SGX) technology. We implement our model on a Docker and OpenSGX platform. The result is a secure and flexible approach for distributed computing on Fog infrastructures.

In [10], we contribute to the research on cache-based side-channel attacks and show the security impact of these attacks on cloud computing. The detection of cache-based side-channel attacks has received more attention in IaaS cloud infrastructures because of improvements in the attack techniques. However, detection of such attacks requires high resolution information, and it is also a challenging task because of the fine-granularity of the attacks. In this paper, we present an approach to detect cross-VM cache-based side-channel attacks through using hardware fine-grained information provided by Intel Cache Monitoring Technology (CMT) and Hardware Performance Counters (HPCs) following the Gaussian anomaly detection method. The approach shows a high detection rate with a 2% performance overhead on the computing platform.

7.4.2. Secure and privacy-aware biomedical analyses

In [11], we study the need for the sharing of genetic data, for instance, in genome-wide association studies, which is incessantly growing. In parallel, serious privacy concerns rise from a multi-party access to genetic information. Several techniques , such as encryption, have been proposed as solutions for the privacy-preserving sharing of genomes. However, existing programming means do not support guarantees for privacy properties and the performance optimization of genetic applications involving shared data. We propose two contributions in this context. First, we present new cloud-based architectures for cloud-based genetic applications that are motivated by the needs of geneticians. Second, we propose a model and implementation for the composition of watermarking with encryption, fragmentation, and client-side computations for the secure and privacy-preserving sharing of genetic data in the cloud.

7.5. Experiment-Driven Resarch

Participants: Adrien Lebre, Bastien Confais, Ronan-Alexandre Cherrueau, Matthieu Simonin, Thuy-Linh Nguyen.

Because STACK members have to perform a significant number of evaluations of complex software stack at large scale, the team contributes to the recent area of software-defined experiments and reproducible research.

In [36], [16] we propose a new approach to ensure reproducibility and repeatability of scientific experiments. Similar to the LAMP stack that considerably eased the web developers life, we advocate the need of an analogous software stack to help the experimenters making reproducible research. In 2018, we propose the EnosStack, an open source software stack especially designed for reproducible scientific experiments. EnosStack enables to easily describe experimental workflows meant to be re-used, while abstracting the underlying infrastructure running them. Being able to switch experiments from a local to a real testbed deployment greatly lower code development and validation time. In this paper, we describe the abstractions that have driven its design, before presenting a real experiment we deployed on Grid'5000 to illustrate its usefulness. We also provide all the experiment code, data and results to the community.

Similar to the previous work, we discuss in [37] a large experimental campaign that allows us to understand in details the boot duration of both virtualization techniques under various storage devices and resources contentions. While many studies have been focusing on reducing the time to manipulate Virtual Machine/Container images in order to optimize provisioning operations in a Cloud infrastructure, only a few studies have considered the time required to boot these systems. Some previous researches showed that the whole boot process can last from a few seconds to few minutes depending on co-located workloads and the number of concurrent deployed machines. The paper explains how we analyzed thoroughly the boot time of VMs, Dockers on top of bare-metal servers, and Dockers inside VMs. We discuss a methodology that enables us to perform fully-automatized and reproducible experimental campaigns on a scientific testbed. Thanks to this methodology, we conducted more than 14.400 experiments on Grid'5000 testbed for a bit more than 500 hours. The results we collected provide an important information related to the boot time behavior of these two virtualization technologies.

In [33], we presented the first experiment that has been done, as far as we know, on top of the Grid'5000 and FIT testbeds. More precisely, we discuss how we evaluated a new storage service for edge/IoT scenarios. Our proof-of-concept relies on the Interplanetary Object Store (IPFS), a Scale-Out NAS deployed on each site and a tree-based approach for the meta-data management [17]. This proposal enables (i) IoT devices to write locally on their closest site and (ii) to relocate automatically the objects on the sites they are requested, leading to low access times. The contribution of this work is a discussion of our attempt of using the two platforms simultaneously as well as the problems we encountered to interconnect them. Our ultimage goal is to give guidelines on how can researchers perform evaluations in a realistic environment : IoT devices comes from the FIT/IoT-lab and Fog nodes from Grid'5000.

AGORA Project-Team

7. New Results

7.1. Wireless network deployment

Participants : Walid Bechkit, Amjed Belkhiri, Jad Oueis, Hervé Rivano, Razvan Stanica, Fabrice Valois

7.1.1. UAVs positionning

Mobile base stations mounted on unmanned aerial vehicles (UAVs) provide viable wireless coverage solutions in challenging landscapes and conditions, where cellular/WiFi infrastructure is unavailable. Operating multiple such airborne base stations, to ensure reliable user connectivity, demands intelligent control of UAV movements, as poor signal strength and user outage can be catastrophic to mission critical scenarios. In [17], we propose a deep reinforcement learning based solution to tackle the challenges of base stations mobility control. We design an Asynchronous Advantage Actor-Critic (A3C) algorithm that employs a custom reward function, which incorporates SINR and outage events information, and seeks to provide mobile user coverage with the highest possible signal quality. Preliminary results reveal that our solution converges after 4 × 10 5 steps of training, after which it outperforms a benchmark gradient-based alternative, as we attain 5dB higher median SINR during an entire test mission of 10,000 steps.

7.1.2. Network functions placement

Emerging mobile network architectures (e.g., aerial networks, disaster relief networks) are disrupting the classical careful planning and deployment of mobile networks by requiring specic 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 [5], 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. We propose a novel centrality metric, the ow centrality, which measures a node capacity of receiving the total amount of ows in the network. We show that in order to maximize the amount of exchanged trac 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 ow centrality. We rst compare our proposed metric to other state of the art centralities. Then, we highlight the signicant trac loss occurring when the core network entity is not placed on the node with the maximum ow centrality, which could reach 55

7.1.3. Mobile edge computing orchestration

Orchestrating network and computing resources in Mobile Edge Computing (MEC) is an important item in the networking research agenda. In [12], we propose a novel algorithmic approach to solve the problem of dynamically assigning base stations to MEC facilities, while taking into consideration multiple time-periods, and computing load switching and access latency costs. In particular, leveraging on an existing state of the art on mobile data analytics, we propose a methodology to integrate arbitrary time-period aggregation methods into a network optimization framework. We notably apply simple consecutive time period aggregation and agglomerative hierarchical clustering. Even if the aggregation and optimization methods represent techniques which are different in nature, and whose aim is partially overlapping, we show that they can be integrated in an efficient way. By simulation on real mobile cellular datasets, we show that, thanks to the clustering, we can scale with the number of time-periods considered, that our approach largely outperforms the case without time-period aggregations in terms of MEC access latency, and at which extent the use of clustering and time aggregation affects computing time and solution quality.

7.1.4. On User Mobility in Dynamic Cloud Radio Access Networks

The development of virtualization techniques enables an architectural shift in mobile networks, where resource allocation, or even signal processing, become software functions hosted in a data center. The centralization of computing resources and the dynamic mapping between baseband processing units (BBUs) and remote antennas (RRHs) provide an increased flexibility to mobile operators, with important reductions of operational costs. Most research efforts on Cloud Radio Access Networks (CRAN) consider indeed an operator perspective and network-side performance indicators. The impact of such new paradigms on user experience has been instead overlooked. In [20], we shift the viewpoint, and show that the dynamic assignment of computing resources enabled by CRAN generates a new class of mobile terminal handover that can impair user quality of service. We then propose an algorithm that mitigates the problem, by optimizing the mapping between BBUs and RRHs on a time-varying graph representation of the system. Furthermore, we show that a practical online BBU-RRH mapping algorithm achieves results similar to an oracle-based scheme with perfect knowledge of future traffic demand. We test our algorithms with two large-scale real-world datasets, where the total number of handovers, compared with the current architectures, is reduced by more than 20%. Moreover, if a small tolerance to dropped calls is allowed, 30% less handovers can be obtained.

7.1.5. Wireless sensor network deployment for environmental monitoring

Air pollution has major negative effects on both human health and environment. Thus, air quality monitoring is a main issue in our days. In [9], we focus on the use of mobile WSN to generate high spatio-temporal resolution air quality maps. We address the sensors' online redeployment problem and we propose three redeployment models allowing to assess, with high precision, the air pollution concentrations. Unlike most of existing movement assisted deployment strategies based on network generic characteristics such as coverage and connectivity, our approaches take into account air pollution properties and dispersion models to offer an efficient air quality estimation. First, we introduce our proposition of an optimal integer linear program based on air pollution dispersion characteristics to minimize estimation errors. Then, we propose a local iterative integer linear programming model and a heuristic technique that offer a lower execution time with acceptable estimation quality. We evaluate our models in terms of execution time and estimation quality using a real data set of Lyon City in France. Finally, we compare our models' performances to existing generic redeployment strategies. Results show that our algorithms outperform the existing generic solutions while reducing the maximum estimation error up to 3 times.

7.2. Wireless data collection

Participants : Walid Bechkit, Ahmed Boubrima, Alexis Duque, Abdoul-Aziz Mbacke, Hervé Rivano, Razvan Stanica, Yosra Zguira

7.2.1. RFID paradigm

While RFID technology is gaining increased attention from industrial community deploying different RFID-based applications, it still suffers from reading collisions. As such, many proposals were made by the scientific community to try and alleviate that issue using different techniques either centralized or distributed, mono-channel or multi-channels, TDMA or CSMA. However, the wide range of solutions and their diversity make it hard to have a clear and fair overview of the different works. In [4], we propose a survey of the most relevant and recent known state-of-the- art anti-collision for RFID protocols. It provides a classification and performance evaluation taking into consideration different criteria as well as a guide to choose the best protocol for given applications depending on their constraints or requirements but also in regard to their deployment environments.

7.2.2. Anti-collision and routing protocol for RFID

In the midst of Internet of Things development, a first requirement was tracking and identification of those mentioned "things" which could be done thanks to Radio Frequency Identification. However, since then, the development of RFID allowed a new range of applications among which is remote sensing of environmental values. While RFID can be seen as a more efficient solution than traditional Wireless Sensor Networks,

two main issues remain: first reading collisions and second proficient data gathering solution. In [18], we examine the implementation of two applications: for industrial IoT and for smart cities, respectively. Both applications, in regards to their requirements and configuration, challenge the operation of a RFID sensing solution combined with a dynamic wireless data gathering over multi-hops. They require the use of both mobile and fixed readers to cover the extent of deployment area and a quick retrieval of tag information. We propose a distributed cross-layer solution for improving the efficiency of the RFID system in terms of collision and throughput but also its proficiency in terms of tag information routing towards one or multiple sinks. Simulation results show that we can achieve high level of throughput while maintaining a low level of collision and a fairness of reader medium access above 95% in situations where readers can be fix and mobile, while tag information is routed with a data rate of 97% at worst and reliable delays for considered applications.

7.2.3. Routing priority information in RFID

Long being used for identification purposes, a new set of applications is now available thanks to the development of RFID technology. One of which is remote sensing of environmental values using passive RFID tags. This leap forward allowed a more energy efficient and cheaper solution for applications like logistics or urban infrastructure monitoring. Nevertheless, serious issues raised with the use of RFID: (i) reading collisions and (ii) gathering of tag information. Indeed, tags information retrieved by readers have to be transmitted towards a base station through a multi-hop scheme which can interfere with neighboring readers activity. In [19], we propose cross-layer solutions meant for both scheduling of readers' activity to avoid collisions, and a multi-hop routing towards base stations, to gather read tag data. This routing is performed with a data priority aware mechanism allowing end-to-end delay reduction of urgent data packets delivery up to 13% faster compared to standard ones. Using fuzzy logic, we combine several observed metrics to reduce the load of forwarding nodes and improve latency as well as data rate. We validate our proposal running simulations on industrial and urban scenarios.

7.2.4. Data collection in DTN networks

Intelligent Transport Systems (ITS) are an essential part of the global world. They play a substantial role for facing many issues such as traffic jams, high accident rates, unhealthy lifestyles, air pollution, etc. Public bike sharing system is one part of ITS and can be used to collect data from mobiles devices. In this paper, we propose an efficient, *Internet of Bikes*, IoB-DTN routing protocol based on data aggregation which applies the Delay Tolerant Network (DTN) paradigm to Internet of Things (IoT) applications running data collection on urban bike sharing system based sensor network. In [6], we propose and evaluate three variants of IoB-DTN: IoB based on spatial aggregation (IoB-SA), IoB based on temporal aggregation (IoB-TA) and IoB based on spatiotemporal aggregation (IoB-STA). The simulation results show that the three variants offer the best performances regarding several metrics, comparing to IoB-DTN without aggregation and the low-power long-range technology, LoRa type. In an urban application, the choice of the type of which variant of IoB should be used depends on the sensed values.

7.2.5. Data sensing in Internet of Bikes

Following the trend of the Internet of Thing, public transport systems are seen as an efficient bearer of mobile devices to generate and collect data in urban environments. Bicycle sharing system is one part of the city's larger transport system. In [23], we study the *Internet of Bikes* IoB-DTN protocol which applies the Delay Tolerant Network (DTN) paradigm to the Internet of Things (IoT) applications running on urban bike sharing system based sensor network. We evaluate the performances of the protocol with respect to the transmission power. Performances are measured in terms of delivery rate, delivery delay, throughput and energy cost. We also compare the multi-hop IoB-DTN protocol to a low-power wide-area network (LPWAN) technology. LPWAN have been designed to provide cost-effective wide area connectivity for small throughput IoT applications: multiyear lifetime and multi-kilometer range for battery-operated mobile devices. This work aims at providing network designers and managers insights on the most relevant technology for their urban applications that could run on bike sharing systems. To the best of our knowledge, this work is the first to provide a detailed performance comparison between multi-hop and long range DTN-like protocol being applied to mobile network IoT devices running a data collection applications in an urban environment.

7.2.6. Reducing IoT traffic through data aggregation mechanisms

Intelligent Transport Systems (ITS) are an essential part of the global world. They play a substantial role for facing many issues such as traffic jams, high accident rates, unhealthy lifestyles, air pollution, etc. Public bike sharing system is one part of ITS and can be used to collect data from mobiles devices. In this paper, we propose an efficient, " Internet of Bikes " , IoB-DTN routing protocol based on data aggregation which applies the Delay Tolerant Network (DTN) paradigm to Internet of Things (IoT) applications running data collection on urban bike sharing system based sensor network. In [6], we propose and evaluate three variants of IoB-DTN: IoB based on spatial aggregation (IoB-SA), IoB based on temporal aggregation (IoB-TA) and IoB based on spatiotemporal aggregation (IoB-STA). The simulation results show that the three variants offer the best performances regarding several metrics, comparing to IoB-DTN without aggregation and the low-power long-range technology, LoRa type. In an urban application, the choice of the type of which variant of IoB should be used depends on the sensed values.

7.2.7. Environmental modeling

Wireless sensor networks (WSN) are widely used in environmental applications where the aim is to sense a physical parameter such as temperature, humidity, air pollution, etc. Most existing WSN-based environmental monitoring systems use data interpolation based on sensor measurements in order to construct the spatiotemporal field of physical parameters. However, these fields can be also approximated using physical models which simulate the dynamics of physical phenomena. In [11], we focus on the use of wireless sensor networks for the aim of correcting the physical model errors rather than interpolating sensor measurements. We tackle the activity scheduling problem and design an optimization model and a heuristic algorithm in order to select the sensor nodes that should be turned off to extend the lifetime of the network. Our approach is based on data assimilation which allows us to use both measurements and the physical model outputs in the estimation of the spatiotemporal field. We evaluate our approach in the context of air pollution monitoring while using a dataset from the Lyon city, France and considering the characteristics of a monitoring system developed in our lab. We analyze the impact of the nodes' characteristics on the network lifetime and derive guidelines on the optimal scheduling of air pollution sensors.

7.2.8. Multi-robot routing for evolving missions

In [22], we propose Dynamic Multi Robot-Routing (DMRR), as a continuous adaptation of the multi-robot target allocation process (MRTA) to new discovered targets. There are few works addressing dynamic target allocation. Existing methods are lacking the continuous integration of new targets, handling its progressive effects, but also lacking dynamic support (e.g. parallel allocations, participation of new robots). This work proposes a framework for dynamically adapting the existing robot missions to new discovered targets. Missions accumulate targets continuously, so the case of a saturation bound for the mission costs is also considered. Dynamic saturation-based auctioning (DSAT) is proposed for allocating targets, providing lower time complexities (due to parallelism in allocation). Comparison is made with algorithms ranging from greedy to auction-based methods with provable sub-optimality. The algorithms are tested on exhaustive sets of inputs, with random configurations of targets (for DMRR with and without a mission saturation bound). The results for DSAT show that it outperforms state-of-the-art methods, like standard sequential single-item auctioning (SSI) or SSI with regret clearing.

7.2.9. Measuring information using VLC

The use of visible light for bidirectional communication between regular smartphones and the small LEDs integrated in most consumer electronics nowadays raises new challenges. In [13], we enhance the state of the art with an efficient image processing algorithm to accurately detect the LEDs and decode their signal in real time. We propose an efficient decoding algorithm, which can detect the LED position, process and decode the signal on average in 18.4 ms, for each frame, on a Nexus 5 unrooted smartphone. Thus, this implementation is convenient for low latency indoor localization or real-time transmission with a moving receiver. Also, as the ROI detection is the most complex step of the algorithm, scenarios with several transmitters can be envisaged, enabling MIMO-like transmissions. We also present smart mechanisms and protocols to build a robust flash-to-LED communications channel using off- the-shelf smartphones and small LEDs. Our experimental evaluation

shows a throughput of 30 bit/s, which is suitable for feedback, wake-up or even some limited communication purposes. We believe that such bidirectional VLC communication system will be a great opportunity for smart and connected consumer electronic products, providing bidirectional smartphone- to-device communication at lower cost.

COATI Project-Team

7. New Results

7.1. Network Design and Management

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

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

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

7.1.1. Software Defined Networks (SDN)

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

7.1.1.1. Bringing Energy Aware Routing Closer to Reality With SDN Hybrid Networks

Energy-aware routing aims at reducing the energy consumption of Internet service provider (ISP) networks. The idea is to adapt routing to the traffic load to turn off some hardware. However, it implies to make dynamic changes to routing configurations which is almost impossible with legacy protocols. The software defined network (SDN) paradigm bears the promise of allowing a dynamic optimization with its centralized controller. In [34], we propose smooth energy aware routing (SENAtoR), an algorithm to enable energy-aware routing in a scenario of progressive migration from legacy to SDN hardware. Since in real life, turning off network devices is a delicate task as it can lead to packet losses, SENAtoR also provides several features to safely enable energy saving services: tunneling for fast rerouting, smooth node disabling, and detection of both traffic spikes and link failures. We validate our solution by extensive simulations and by experimentation. We show that SENAtoR can be progressively deployed in a network using the SDN paradigm. It allows us to reduce the energy consumption of ISP networks by 5%–35% depending on the penetration of SDN hardware while diminishing the packet loss rate compared to legacy protocols.

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

7.1.1.2. Energy-Aware Routing in Software-Defined Network using Compression

Over past few years, many applications have been built using SDN such as server load balancing, virtual-machine migration, traffic engineering and access control. In [31], we focus on using SDN for energy-aware routing (EAR). Since traffic load has a small influence on the power consumption of routers, EAR allows putting unused links into sleep mode to save energy. SDN can collect traffic matrix and then computes routing solutions satisfying QoS while being minimal in energy consumption. However, prior works on EAR have assumed that the SDN forwarding table switch can hold an infinite number of rules. In practice, this assumption does not hold since such flow tables are implemented in Ternary Content Addressable Memory (TCAM) which is expensive and power-hungry. We consider the use of wildcard rules to compress the forwarding tables. In [31], we propose optimization methods to minimize energy consumption for a backbone network while respecting capacity constraints on links and rule space constraints on routers. In details, we present two exact formulations using Integer Linear Program (ILP) and introduce efficient heuristic algorithms. Based on simulations on realistic network topologies, we show that using this smart rule space allocation, it is possible to save almost as much power consumption as the classical EAR approach.

7.1.1.3. Complexity of Compressing Two Dimensional Routing Tables with Order

Motivated by routing in telecommunication network using Software Defined Network (SDN) technologies, we consider the following problem of finding short routing lists using aggregation rules. We are given a set of communications $\mathcal{X}$, which are distinct pairs $(s,t) \subseteq S \times T$, (typically S is the set of sources and T the set of destinations), and a port function $\pi : \mathcal{X} \to P$ where P is the set of ports. A *routing list*$\mathcal{R}$ is an ordered list of triples which are of the form (s,t,p), $(*,t,p)$, $(s,*,p)$ or $(*,*,p)$ with $s \in S$, $t \in T$ and $p \in P$. It *routes* the communication (s,t) to the port $r(s,t) = p$ which appears on the first triple in the list $\mathcal{R}$ that is of the form (s,t,p), $(*,t,p)$, $(s,*,p)$ or $(*,*,p)$. If $r(s,t) = \pi(s,t)$, then we say that (s,t) is *properly routed* by $\mathcal{R}$ and if all communications of $\mathcal{X}$ are properly routed, we say that $\mathcal{R}$*emulates*$(\mathcal{X},\pi)$. The aim is to find a shortest routing list emulating $(\mathcal{X},\pi)$. In [30], we carry out a study of the complexity of the two dual decision problems associated to it. Given a set of communication $\mathcal{X}$, a port function π and an integer k, the first one called ROUTING LIST (resp. the second one, called LIST REDUCTION) consists in deciding whether there is a routing list emulating $(\mathcal{X},\pi)$ of size at most k (resp. $|\mathcal{X}| - k$). We prove that both problems are NP-complete. We then give a 3-approximation for LIST REDUCTION, which can be generalized to higher dimensions. We also give a 4-approximation for ROUTING LIST in the fundamental case when there are only two ports (i.e. $|P| = 2$), $\mathcal{X} = S \times T$ and $|S| = |T|$.

7.1.2. Provisioning Service Function Chains

7.1.2.1. Optimal Network Service Chain Provisioning

Service chains consist of a set of network services, such as firewalls or application delivery controllers, which are interconnected through a network to support various applications. While it is not a new concept, there has been an extremely important new trend with the rise of Software-Defined Network (SDN) and Network Function Virtualization (NFV). The combination of SDN and NFV can make the service chain and application provisioning process much shorter and simpler. In [33], [48], we study the provisioning of service chains jointly with the number/location of Virtual Network Functions (VNFs). While chains are often built to support multiple applications, the question arises as how to plan the provisioning of service chains in order to avoid data passing through unnecessary network devices or servers and consuming extra bandwidth and CPU cycles. It requires choosing carefully the number and the location of the VNFs. We propose an exact mathematical model using decomposition methods whose solution is scalable in order to conduct such an investigation. We conduct extensive numerical experiments, and show we can solve exactly the routing of service chain requests in a few minutes for networks with up to 50 nodes, and traffic requests between all pairs of nodes. Detailed analysis is then made on the best compromise between minimizing the bandwidth requirement and minimizing the number of VNFs and optimizing their locations using different data sets.

7.1.2.2. Energy-Efficient Service Function Chain Provisioning

Network Function Virtualization (NFV) is a promising network architecture concept to reduce operational costs. In legacy networks, network functions, such as firewall or TCP optimization, are performed by specific

hardware. In networks enabling NFV coupled with the Software Defined Network (SDN) paradigm, Virtual Network Functions (VNFs) can be implemented dynamically on generic hardware. This is of primary interest to implement energy efficient solutions, in order to adapt the resource usage dynamically to the demand. In [35], we study how to use NFV coupled with SDN to improve the energy efficiency of networks. We consider a setting in which a flow has to go through a Service Function Chain, that is several network functions in a specific order. We propose an ILP formulation, an ILP-based heuristic, as well as a decomposition model that relies on joint routing and placement configuration to solve the problem. We show that virtualization provides between 22% to 62% of energy savings for networks of different sizes.

7.1.2.3. Placement of Service Function Chains with Ordering Constraints

A Service Function Chain (SFC) is an ordered sequence of network functions, such as load balancing, content filtering, and firewall. With the Network Function Virtualization (NFV) paradigm, network functions can be deployed as pieces of software on generic hardware, leading to a flexibility of network service composition. Along with its benefits, NFV brings several challenges to network operators, such as the placement of virtual network functions. In [49], [50], [62], we study the problem of how to optimally place the network functions within the network in order to satisfy all the SFC requirements of the flows. Our optimization task is to minimize the total deployment cost. We show that the problem can be seen as an instance of the Set Cover Problem, even in the case of ordered sequences of network functions. It allows us to propose two logarithmic factor approximation algorithms which have the best possible asymptotic factor. Further, we devise an optimal algorithm for tree topologies. Finally, we evaluate the performances of our proposed algorithms through extensive simulations. We demonstrate that near-optimal solutions can be found with our approach.

7.1.2.4. Resource Requirements for Reliable Service Function Chaining

We study in [51], [49] the problem of deploying reliable Service Function Chains over a virtualized network function architecture. While there is a need for reliable service function chaining, there is a high cost to pay for it in terms of bandwidth and VNF processing requirements. We investigate two different protection mechanisms and discuss their resource requirements, as well as the latency of their paths. For each mechanism, we develop a scalable exact mathematical model using column generation.

7.1.2.5. Path protection in optical flexible networks with distance-adaptive modulation formats

Thanks to a flexible frequency grid, Elastic Optical Networks (EONs) will support a more efficient usage of the spectrum resources. On the other hand, this efficiency may lead to even more disruptive effects of a failure on the number of involved connections with respect to traditional networks. In [52], we study the problem of providing path protection to the lightpaths against a single fiber failure event in the optical layer. Our optimization task is to minimize the spectrum requirements for the protection in the network. We develop a scalable exact mathematical model using column generation for both shared and dedicated path protection schemes. The model takes into account practical constraints such as the modulation format, regenerators, and shared risk link groups. We demonstrate the effectiveness of our model through extensive simulation on two real-world topologies of different sizes. Finally, we compare the two protection schemes under different scenario assumptions, studying the impact of factors such as number of regenerators and demands on their performances.

7.1.2.6. Reconfiguring Service Functions chains with a make-before-break approach

The centralized routing model of SDN jointly with the possibility of instantiating VNFs on-demand open the way for a more efficient operation and management of networks. In [58], we consider the problem of reconfiguring network connections 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.

7.1.3. Capacity defragmentation

Optical multilayer optimization continuously reorganizes layer 0-1-2 network elements to handle both existing and dynamic traffic requirements in the most efficient manner. This delays the need to add new resources for new requests, saving CAPEX and leads to optical network defragmentation.

In [46], [47], we focus on Layer 2, i.e., on capacity defragmentation at the Optical Transport Network (OTN) layer when routes (e.g., LSPs in MPLS networks) are making unnecessarily long detours to evade congestion. Reconfiguration into optimized routes can be achieved by redefining the routes, one at a time, so that they use the vacant resources generated by the disappearance of services using part of a path that transits the congested section. For the Quality of Service, it is desirable to operate under Make-Before-Break (MBB), with the minimum number of rerouting. The challenge is to identify the rerouting order, one connection at a time, while minimizing the bandwidth requirement. We propose in [46], [47] an exact and scalable optimization model for computing a minimum bandwidth rerouting scheme subject to MBB in the OTN layer of an optical network. Numerical results show that we can successfully apply it on networks with up to 30 nodes, a very significant improvement with the state of the art. We also provide some defragmentation analysis in terms of the bandwidth requirement vs. the number of reroutings.

In [37], we focus on wavelength defragmentation in WDM networks. We propose a MBB wavelength defragmentation process which minimizes the bandwidth requirement of the resulting provisioning. Comparisons with minimum bandwidth provisioning that is not subject to MBB show that, on average, the best seamless lightpath rerouting is never more than 5% away (less than 1% on average) from an optimal lightpath provisioning.

7.1.4. 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 [16], 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.

7.1.5. Optimizing drone coverage

In the context of a collaboration with Tahiry Razafindralambo from the University of la Réunion we have studied several problems related to deployment of drones in order to collect data generated from sensors. Those problems may be seen as belonging to the category of "cover" problems and we have designed and proposed efficient formulations using linear programming models with columns generation.

Drones (Unmanned Aerial Vehicles, UAV) can be used to provide anytime and anywhere network access to targets located on the ground, using air-to-ground and air-to-air communications through directional antennas. In [43] we study how to deploy these drones to cover a set of fixed targets. It is a complex problem since each target should be covered, while minimizing (i) the deployment cost and (ii) the drones altitudes to ensure good communication quality. We also consider connectivity between the drone and a base station in order to collect and send information to the targets, which is not considered in many similar studies. We provide an efficient optimal program to solve the problem and show the trade-off analysis due to conflicting objectives. We propose a fair trade-off optimal solution and also evaluate the cost of adding connectivity to the drone deployment.

In [41], [42] we introduce a Linear Programming (LP) model for the problem of data gathering with mobile drones. The goal is to deploy a connected set of Unmanned Aerial Vehicles (UAVs) continuously monitoring mobile sensors and reporting information to a fixed base station for efficient data collection. We propose an effective optimization model reducing the number of variables of the problem and solved using column generation. Results show that our model is tractable for large topologies with several hundreds of possible 3D locations for the UAVs deployment and provides integer solutions with the generated columns very close to the optimum. Moreover, the deployment changes among time remains low in terms of number of UAVs and cost, to maintain connectivity and minimize the data collection delay to the base station.

We also have studied a problem arising when one will to recharge wireless sensor networks using drones and wireless power transfer; in [44] we consider the optimal energy replenishment problem (OERP). The goal is to operate a given number of flying drones in order to efficiently recharge wireless sensor nodes. We present a linear program that maximizes the amount of harvested energy to the sensors. We show that the model is solved to optimality in a few seconds for sensor networks with up to 50 nodes. The small number of available drones is shown to be optimally deployed at low altitude in order to efficiently recharge the batteries of at least half of the sensor nodes.

7.1.6. Other results in wireless networks

7.1.6.1. 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 [64], 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.

7.1.6.2. Gossiping with interference in radio chain networks

In [53], we study the problem of gossiping with interference constraint in radio chain networks. Gossiping (or total exchange information) is a protocol where each node in the network has a message and wants to distribute its own message to every other node in the network. The gossiping problem consists in finding the minimum running time (makespan) of a gossiping protocol and efficient algorithms that attain this makespan. The network is assumed to be synchronous, the time is slotted into steps, and each device is equipped with a half duplex interface; so, a node cannot both receive and transmit during a step. We use a binary asymmetric model of interference based on the distance in the communication digraph. We determine exactly the minimum number of rounds R needed to achieve a gossiping when transmission network is a dipath P_n on $n \geq 3$ nodes and the interference distance is $d_I = 1$.

7.2. Graph Algorithms

Participants: Julien Bensmail, Jean-Claude Bermond, Nathann Cohen, David Coudert, Frédéric Giroire, Frédéric Havet, Fionn Mc Inerney, Nicolas Nisse, Stéphane Pérennes.

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

Parameterized complexity theory has enabled a refined classification of the difficulty of NP-hard optimization problems on graphs with respect to key structural properties, and so to a better understanding of their true difficulties. More recently, hardness results for problems in P were established under reasonable complexity theoretic assumptions such as: Strong Exponential Time Hypothesis (SETH), 3SUM and All-Pairs Shortest-Paths (APSP). According to these assumptions, many graph theoretic problems do not admit truly subquadratic algorithms, nor even truly subcubic algorithms (Williams and Williams, FOCS 2010 [83] and Abboud *et al.* SODA 2015 [67]). A central technique used to tackle the difficulty of the above mentioned problems is fixed-parameter algorithms for polynomial-time problems with *polynomial dependency* in the fixed parameter (P-FPT). This technique was rigorously formalized by Giannopoulou et al. (IPEC 2015) [74], [75]. Following that, it was continued by Abboud *et al.* (SODA 2016) [68], by Husfeldt (IPEC 2016) [76] and Fomin *et al.* (SODA 2017) [73], using the treewidth as a parameter. Applying this technique to *clique-width*, another important graph parameter, remained to be done.

In [45] we study several graph theoretic problems for which hardness results exist such as *cycle problems* (triangle detection, triangle counting, girth), *distance problems* (diameter, eccentricities, Gromov hyperbolicity, betweenness centrality) and *maximum matching*. We provide hardness results and fully polynomial FPT algorithms, using clique-width and some of its upper-bounds as parameters (split-width, modular-width and P_4-sparseness). We believe that our most important result is an $\mathcal{O}(k^4 \cdot n + m)$-time algorithm for computing a maximum matching where k is either the modular-width or the P_4-sparseness. The latter generalizes many algorithms that have been introduced so far for specific subclasses such as cographs, P_4-lite graphs, P_4-extendible graphs and P_4-tidy graphs. Our algorithms are based on preprocessing methods using modular decomposition, split decomposition and primeval decomposition. Thus they can also be generalized to some graph classes with unbounded clique-width.

7.2.1.2. Revisiting Decomposition by Clique Separators

We study in [26] the complexity of decomposing a graph by means of clique separators. This common algorithmic tool, first introduced by Tarjan [79], allows to cut a graph into smaller pieces, and so, it can be applied to preprocess the graph in the computation of optimization problems. However, the best-known algorithms for computing a decomposition have respective $\mathcal{O}(nm)$-time and $\mathcal{O}(n^{(3+\alpha)/2}) = o(n^{2.69})$-time complexity, with $\alpha < 2.3729$ being the exponent for matrix multiplication. Such running times are prohibitive for large graphs. In [26], we prove that for every graph G, a decomposition can be computed in $\mathcal{O}(T(G) + \min\{n^\alpha, \omega^2 n\})$-time with $T(G)$ and ω being respectively the time needed to compute a minimal triangulation of G and the clique-number of G. In particular, it implies that every graph can be decomposed by clique separators in $\mathcal{O}(n^\alpha \log n)$-time. Based on prior work from Kratsch and Spinrad [77], we prove in addition that decomposing a graph by clique-separators is as least as hard as triangle detection. Therefore, the existence of any $o(n^\alpha)$-time algorithm for this problem would be a significant breakthrough in the field of algorithmic. Finally, our main result implies that planar graphs, bounded-treewidth graphs and bounded-degree graphs can be decomposed by clique separators in linear or quasi-linear time.

7.2.1.3. Distance-preserving elimination orderings in graphs

For every connected graph G, a subgraph H of G is *isometric* if the distance between any two vertices in H is the same in H as in G. A *distance-preserving elimination ordering* of G is a total ordering of its vertex-set $V(G)$, denoted $(v_1, v_2, ..., v_n)$, such that any subgraph $G_i = G \setminus (v_1, v_2, ..., v_i)$ with $1 \leq i < n$ is isometric. This kind of ordering has been introduced by Chepoi in his study on weakly modular graphs [71]. In [27], we prove that it is NP-complete to decide whether such ordering exists for a given graph | even if it has diameter at most 2. Then, we prove on the positive side that the problem of computing a distance-preserving ordering when there exists one is fixed-parameter-tractable in the treewidth. Lastly, we describe a heuristic in order to compute a distance-preserving ordering when there exists one that we compare to an exact exponential time algorithm and to an ILP formulation for the problem.

7.2.1.4. Complexity of computing strong pathbreadth

The strong pathbreadth of a given graph G is the minimum ρ such that G admits a Robertson and Seymour's path decomposition where every bag is the complete ρ-neighbourhood of some vertex in G. In [29] [0], we prove that deciding whether a given graph has strong pathbreadth at most one is NP-complete. This answers negatively to a conjecture of Leitert and Dragan [78].

7.2.1.5. Improving matchings in trees, via bounded-length augmentations

In [13] Due to a classical result of Berge, it is known that a matching of any graph can be turned into a maximum matching by repeatedly augmenting alternating paths whose ends are not covered. In a recent work, Nisse, Salch and Weber considered the influence, on this process, of augmenting paths with length at most k only. Given a graph G, an initial matching $M \subseteq E(G)$ and an odd integer k, the problem is to find a longest sequence of augmenting paths of length at most k that can be augmented sequentially from M. They proved that, when only paths of length at most $k = 3$ can be augmented, computing such a longest sequence can be done in polynomial time for any graph, while the same problem for any $k \geq 5$ is *NP*-hard. Although the latter result remains true for bipartite graphs, the status of the complexity of the same problem for trees is not known.

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

7.2.2. Dynamics of formation of communities in social networks

We consider in [40] 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 in this paper 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 $O(|V| \min \alpha(G^-), \sqrt{|V|})$ 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 [40], we disprove this. Specifically, we prove that for any $k \geq 4$, the maximum time of convergence is an $\Omega(|V|^{\Theta(\log |V|)})$.

7.2.3. Application to bioinformatics

For a (possibly infinite) fixed family of graphs $\mathcal{F}$, we say that a graph G *overlays* $\mathcal{F}$ on a hypergraph H if $V(H)$ is equal to $V(G)$ and the subgraph of G induced by every hyperedge of H contains some member of $\mathcal{F}$ as a spanning subgraph. While it is easy to see that the complete graph on $|V(H)|$ overlays $\mathcal{F}$ on a hypergraph H whenever the problem admits a solution, the MINIMUM $\mathcal{F}$-OVERLAY problem asks for such a graph with at most k edges, for some given $k \in \mathbb{N}$. This problem allows to generalize some natural problems which may arise in practice. For instance, if the family $\mathcal{F}$ contains all connected graphs, then MINIMUM $\mathcal{F}$-OVERLAY corresponds to the MINIMUM CONNECTIVITY INFERENCE problem (also known as SUBSET INTERCONNECTION DESIGN problem) introduced for the low-resolution reconstruction of macro-molecular assembly in structural biology, or for the design of networks.

[0] Work done while G. Ducoffe was a member of COATI and published this year.

In [23], we prove a strong dichotomy result regarding the polynomial vs. NP-complete status with respect to the considered family $\mathcal{F}$. Roughly speaking, we show that the easy cases one can think of (e.g. when edgeless graphs of the right sizes are in $\mathcal{F}$, or if $\mathcal{F}$ contains only cliques) are the only families giving rise to a polynomial problem: all others are $\mathcal{NP}$-complete. We then investigate the parameterized complexity of the problem and give similar sufficient conditions on $\mathcal{F}$ that give rise to W[1]-hard, W[2]-hard or FPT problems when the parameter is the size of the solution. This yields an FPT/W[1]-hard dichotomy for a relaxed problem, where every hyperedge of H must contain some member of $\mathcal{F}$ as a (non necessarily spanning) subgraph.

7.3. Games on Graphs

Participants: Julien Bensmail, Nicolas Nisse, Fionn Mc Inerney, Stéphane Pérennes.

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

7.3.1. Spy-game on graphs and eternal domination

In [24] we define and study the following two-player game on a graph G. Let $k \in \mathbf{N}^*$. A set of k guards is occupying some vertices of G while one spy is standing at some node. At each turn, first the spy may move along at most s edges, where $s \in \mathbf{N}^*$ is his speed. Then, each guard may move along one edge. The spy and the guards may occupy the same vertices. The spy has to escape the surveillance of the guards, i.e., must reach a vertex at distance more than $d \in \mathbf{N}$ (a predefined distance) from every guard. Can the spy win against k guards? Similarly, what is the minimum distance d such that k guards may ensure that at least one of them remains at distance at most d from the spy? This game generalizes two well-studied games: Cops and robber games (when $s = 1$) and Eternal Dominating Set (when s is unbounded).

In [24], we consider the computational complexity of the problem, showing that it is NP-hard (for every speed s and distance d) and that some variant of it is PSPACE-hard in DAGs. Then, we establish tight tradeoffs between the number of guards, the speed s of the spy and the required distance d when G is a path or a cycle.

In order to determine the smallest number of guards necessary for this task, we analyze in [25] the game through a Linear Programming formulation and the fractional strategies it yields for the guards. We then show the equivalence of fractional and integral strategies in trees. This allows us to design a polynomial-time algorithm for computing an optimal strategy in this class of graphs. Using duality in Linear Programming, we also provide non-trivial bounds on the fractional guard-number of grids and torus. We believe that the approach using fractional relaxation and Linear Programming is promising to obtain new results in the field of combinatorial games.

In [60] we pursue the study of the eternal domination game (which is equivalent to the spy game when s is unbounded and $d = 0$) on strong grids $P_n \square P_m$. 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. We prove that, for all $n, m \in \mathbb{N}^*$ such that $m \geq n$, $\lceil \frac{nm}{9} \rceil + \Omega(n + m) = \gamma^{\infty}_{all}(P_n \boxtimes P_m) = \lceil \frac{nm}{9} \rceil + O(m\sqrt{n})$ (note that $\lceil \frac{nm}{9} \rceil$ is the domination number of $P_n \boxtimes P_m$).

7.3.2. Metric dimension & localization

The questions that we study there are variant of the usual *Metric Dimension* problem in which one wishes to identify the vertices of a graph from the knowledge of the distances to a few points. This is motivated by localization problems, e.g., in cellular networks. few anchors.

In [19] we introduce a generalization of metric dimension based on a pursuit graph game that resembles the famous Cops and Robbers game. In this game, an invisible target is hidden at some vertex of a graph (at each turn, it may move to a neighbor). At every step, $k \geq 1$ vertices of G can be probed which results in the knowledge of the distances between each of these vertices and the secret location of the target. We provide upper bounds on the related graph invariant $\zeta(G)$, defined as the least number of probes per turn needed to localize the robber on a graph G, for several classes of graphs (trees, bipartite graphs, etc). Our main result is that, surprisingly, there exists planar graphs of treewidth 2 and unbounded $\zeta(G)$. On a positive side, we prove that $\zeta(G)$ is bounded by the pathwidth of G. We then show that the algorithmic problem of determining $\zeta(G)$ is NP-hard in graphs with diameter at most 2. Finally, we show that at most one cop can approximate (arbitrary close) the location of the robber in the Euclidean plane. We further study this problem in [18] where, in particular, we prove that $\zeta(G) \leq 3$ in outer-planar graphs.

In [39], [56], [38], we address the sequential metric dimension when the invisible target is immobile. The objective of the game is to minimize the number of steps needed to locate the target whatever be its location. Precisely, given a graph G and two integers $k, \ell \geq 1$, the LOCALIZATION problem asks whether there exists a strategy to locate a target hidden in G in at most ℓ steps and probing at most k vertices per step. We first show that, in general, this problem is *NP*-complete for every fixed $k \geq 1$ (resp., $\ell \geq 1$). We then focus on the class of trees. On the negative side, we prove that the LOCALIZATION PROBLEM is *NP*-complete in trees when k and ℓ are part of the input. On the positive side, we design a $(+1)$-approximation for the problem in n-node trees, *i.e.*, an algorithm that computes in time $O(n \log n)$ (independent of k) a strategy to locate the target in at most one more step than an optimal strategy. This algorithm can be used to solve the LOCALIZATION PROBLEM in trees in polynomial time if k is fixed.

In [57] we try to understand the phenomena when one choose an orientation of an (undirected) graphs. Namely, 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 $\frac{nm}{2}$.

7.3.3. Orienting edges to fight fire in graphs

In [12], we investigate a new oriented variant of the Firefighter Problem. In the traditional Firefighter Problem, a fire breaks out at a given vertex of a graph, and at each time interval spreads to neighbouring vertices that have not been protected, while a constant number of vertices are protected at each time interval. In our version of the problem, the firefighters are able to orient the edges of the graph before the fire breaks out, but the fire could start at any vertex. We consider this problem when played on a graph in one of several graph classes, and give upper and lower bounds on the number of vertices that can be saved. In particular, when one firefighter is available at each time interval, and the given graph is a complete graph, or a complete bipartite graph, we present firefighting strategies that are provably optimal. We also provide lower bounds on the number of vertices that can be saved as a function of the chromatic number, of the maximum degree, and of the treewidth of a graph. For a sub-cubic graph, we show that the firefighters can save all but two vertices, and this is best possible.

7.3.4. Network decontamination

The Network Decontamination problem consists in 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. In this Chapter, we focus on networks modeled by graphs. Many different

objectives have been studied in this context, the main one being the minimization of the number of mobile agents necessary to clean a contaminated network. Another important aspect is that this optimization problem 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). Our chapter [61] consists of an overview of algorithmic results on graph decontamination and graph searching.

7.3.5. Hyperopic Cops and Robbers

We introduce in [17] a new variant of the game of Cops and Robbers played on graphs, where the robber is invisible unless outside the neighbor set of a cop. The hyperopic cop number is the corresponding analogue of the cop number, and we investigate bounds and other properties of this parameter. We characterize the cop-win graphs for this variant, along with graphs with the largest possible hyperopic cop number. We analyze the cases of graphs with diameter 2 or at least 3, focusing on when the hyperopic cop number is at most one greater than the cop number. We show that for planar graphs, as with the usual cop number, the hyperopic cop number is at most 3. The hyperopic cop number is considered for countable graphs, and it is shown that for connected chains of graphs, the hyperopic cop density can be any real number in $[0, 1/2]$.

7.4. Graph theory

Participants: Julien Bensmail, Frédéric Havet, William Lochet, Nicolas Nisse, Fionn Mc Inerney, Stéphane Pérennes, Bruce Reed.

COATI studies theoretical problems in graph theory. If some of them are directly motivated by applications, others are more fundamental.

7.4.1. Interval number in cycle convexity

Recently, Araujo et al. [Manuscript in preparation, 2017] introduced the notion of Cycle Convexity of graphs. In their seminal work, they studied the graph convexity parameter called hull number for this new graph convexity they proposed, and they presented some of its applications in Knot theory. Roughly, the *tunnel number* of a knot embedded in a plane is upper bounded by the hull number of a corresponding planar 4-regular graph in cycle convexity. In [4], we go further in the study of this new graph convexity and we study the interval number of a graph in cycle convexity. This parameter is, alongside the hull number, one of the most studied parameters in the literature about graph convexities. Precisely, given a graph G, its *interval number* in cycle convexity, denoted by $CCIHN(G)$, is the minimum cardinality of a set $S \subseteq V(G)$ such that every vertex $w \in V(G) \setminus S$ has two distinct neighbors $u, v \in S$ such that u and v lie in same connected component of $G[S]$, i.e. the subgraph of G induced by the vertices in S.

In [4] we provide bounds on $CCIHN(G)$ and its relations to other graph convexity parameters, and explore its behaviour on grids. Then, we present some hardness results by showing that deciding whether $CCIHN(G) \leq k$ is NP-complete, even if G is a split graph or a bounded-degree planar graph, and that the problem is W[2]-hard in bipartite graphs when k is the parameter. As a consequence, we obtain that $CCIHN(G)$ cannot be approximated up to a constant factor in the classes of split graphs and bipartite graphs (unless $P = NP$).

On the positive side, we present polynomial-time algorithms to compute $CCIHN(G)$ for outerplanar graphs, cobipartite graphs and interval graphs. We also present fixed-parameter tractable (FPT) algorithms to compute it for $(q, q-4)$-graphs when q is the parameter and for general graphs G when parameterized either by the treewidth or the neighborhood diversity of G.

Some of our hardness results and positive results are not known to hold for related graph convexities and domination problems. We hope that the design of our new reductions and polynomial-time algorithms can be helpful in order to advance in the study of related graph problems.

7.4.2. Steinberg-like theorems for backbone colouring

A function $f : V(G) \to \{1, ..., k\}$ is a (proper) k-colouring of G if $|f(u) - f(v)| \geq 1$, for every edge $uv \in E(G)$. The *chromatic number*$\chi(G)$ is the smallest integer k for which there exists a proper k-colouring of G. Given a graph G and a subgraph H of G, a circular q-backbone k-colouring f of (G, H) is a k-colouring of G such that $q \leq |c(u) - c(v)| \leq k - q$, for each edge $uv \in E(H)$. The *circular q-backbone chromatic number* of a graph pair (G, H), denoted $\mathrm{CBC}_q(G, H)$, is the minimum k such that (G, H) admits a circular q-backbone k-colouring. Steinberg conjectured that if G is planar and G contains no cycles on 4 or 5 vertices, then $\chi(G) \leq 3$. If this conjecture is correct, then one could deduce that $\mathrm{CBC}_2(G, H) \leq 6$, for any $H \subseteq G$. In [5], we first show that if G is a planar graph containing no cycle on 4 or 5 vertices and $H \subseteq G$ is a forest, then $\mathrm{CBC}_2(G, H) \leq 7$. Then, we prove that if $H \subseteq G$ is a forest whose connected components are paths, then $\mathrm{CBC}_2(G, H) \leq 6$.

7.4.3. Homomorphisms of planar signed graphs and absolute cliques

Homomorphisms are an important topic in graph theory, as example the chromatic number of a graph G is the minimum k such that G maps onto the complete graph K_k. A signed graph (G, Σ) is a (simple) graph with sign function $\Sigma E(G) \to \{-1, 1\}$. A closed-walk is unbalanced if it has an odd number of negative edges, it is balanced otherwise. Homomorphisms of signed graphs are mapping that preserve adjacency and balance of cycles. Naserasr, Rollova and Sopena (Journal of Graph Theory 2015) posed the important question of finding out the minimum size k such that any planar signed graph (G, Σ) admits a homomorphism to a signed graph with k vertices. The question can be seen as the counterpart of the 4 color theorem which implies that any palnar graph maps onto K_4. It is known that if this minimum value is equal to 10, then every planar signed graph maps to a particular unique signed graph (P^+9, Γ^+) with 10 vertices. A graph G is an underlying absolute signed clique if there exists a signed graph (G, Σ) which does not admit any homomorphism to any signed graph (H, Π) with $|V(H)| < |V(G)|$. In [66] we characterize all underlying absolute signed planar cliques up to spanning subgraph inclusion. Furthermore, we show that every signed planar graph having underlying graphs obtained by (repeated, finite) k-clique sums ($k \leq 3$) of underlying absolute signed planar cliques admits a homomorphism to (P^+9, Γ^+). Based on this evidence, we conjecture that every planar signed graph admits a homomorphism to $(P + 9, \Gamma^+)$.

7.4.4. Edge-partitioning a graph into paths: 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. We recently proved this conjecture with Merker [69].

The path case of the Barát-Thomassen conjecture (i.e $\forall k, m = |E| \, mod \, k = 0$ there exists $f(k)$ such that if the connectivity of G is larger than $f(k)$ then G can be partitionned into P_k) has also been studied, notably by Thomassen [80], [81], [82], and had been solved by Botler, Mota, Oshiro and Wakabayashi [70]. In [15] we 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 $f(k)$ has an edge-partition into paths of length k. We also show that 24 can be dropped to 4 when the graph is Eulerian.

7.4.5. Some Aspects of Arbitrarily Partitionable Graphs

An n-graph G is arbitrarily partitionable (AP) if, for every partition of n as $n = n_1 + ... + n_p$, there is a partition $(V_1, ..., V_p)$ of $V(G)$ such that for $i = 1, ..., pG[V_i]$ is connected and $|V_i| = n_i$. The property of being AP is related to other well-known graph notions, such as perfect matchings and Hamiltonian cycles (obviously Hamiltonian graph is AP), with which it shares several properties. In [65] This work we studying two aspects of AP graphs.

On the one hand, we consider the algorithmic aspects. We first establish the *NP*-hardness of the problem of partitioning a graph into connected subgraphs following a given sequence, for various new graph classes of interest. We then prove that the problem of deciding whether a graph is AP is *NP*-hard for several classes of graphs, confirming a conjecture of Barth and Fournier.

On the other hand, we consider the weakening of APness to sufficient conditions for Hamiltonicity. While previous works have suggested that such conditions can sometimes indeed be weakened, we point out cases for which this is not true. This is done by considering conditions for Hamiltonicity involving squares of graphs, and claw- and net-free graphs.

7.4.6. Incident Sum problems and the 1-2-3 Conjecture

How can one distinguish the adjacent vertices of a graph through an edge-weighting? In the last decades, this question has been attracting increasing attention, which resulted in the active field of distinguishing labelings. One of its most popular problems is the one where neighbours must be distinguishable via their incident sums of weights. An edge-weighting verifying this is said to be *proper*. The popularity of this notion arises mainly due to the influence of the famous 1-2-3 Conjecture (posed by Karoński, Łuczak and Thomason), which claims that proper weightings with weights in $\{1, 2, 3\}$ exist for graphs with no isolated edge.

The questions that we study aim at solving or at progressing toward the solution of the 1-2-3 conjecture and similar problems.

In [8] we study locally irregular decompositions of sub-cubic graphs. A graph G is locally irregular if every two adjacent vertices of G have different degrees (this corresponds to a uniform weight). A locally irregular decomposition of G is a partition $E_1, \cdots, E_k$ of the edge set $E(G)$ such that each $G[E_i]$ is locally irregular. Not all graphs admit locally irregular decompositions, but for those who are decomposable, it was conjectured by Baudon, Bensmail, Przybyło and Woźniak that the decomposition uses at most 3 locally irregular graphs. Towards that conjecture, it was recently proved by Bensmail, Merker and Thomassen that every decomposable graph decomposes into at most 328 locally irregular graphs. Our work focuses on the case of sub-cubic graphs, which form an important family of graphs in this context, as all non-decomposable graphs are sub-cubic. As a main result, we prove that decomposable sub-cubic graphs decompose into at most 5 locally irregular graphs, and at most 4 when the maximum average degree is less than $\frac{12}{5}$. We then consider weaker decomposition, where subgraphs can also include regular connected components, and prove the relaxations of the conjecture above for sub-cubic graphs.

In [9] we pursue recent works generalizing "Neighbour Sum problems" (e.g. the well-known 1-2-3 Conjecture, or the notion of locally irregular decomposition) to digraphs. We introduce and study several variants of the 1-2 Conjecture for digraphs and for every such variant, we state conjectures concerning the number of weights necessary to obtain a desired total-weighting in any digraph. We verify some of these conjectures, while we obtain close results towards the solution of the ones that are still open.

In [10] we study a variant of the classical 1-2-3 Conjecture. This conjecture asks whether every graph but K_2 can be 3-edge-weighted so that every two adjacent vertices u and v can be distinguished via the sum of their incident weights, that is the incident sums of u and v differ by at least 1. In this work we investigate the consequences on the 1-2-3 Conjecture of requiring a stronger distinction condition, that is requiring the incident sums to differ by at least 2. Our conjecture is that every graph but K_2 admits a 5-edge-weighting permitting to distinguish the adjacent vertices in this stronger way. We prove this conjecture for several classes of graphs, including bipartite graphs and cubic graphs. We then consider algorithmic aspects, and show that it is *NP*-complete to determine the smallest k such that a given bipartite graph admits such a k-edge-weighting. In contrast, we show that the same problem can be solved in polynomial time when the graph is a tree.

In [11] we prove a 1-2-3-4 result for the 1-2-3 Conjecture in 5-regular graphs. Currently the best-known result toward te 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 (.i.e in our context total means that both the vertices and the edges are assigned weights). Our work, 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 adjacent vertices.

In [63] we investigate another aspect of edge weighting that allow to distinguish adjacent vertices (we shall call them *proper*). Namely we study the minimum number of distinct neighbourhood sums we can produce using such proper weightings. Clearly, this minimum number is bounded below by the chromatic number $\chi(G)$

of G. When using weights in Z, we show that we can always produce proper edge-weightings generating $\chi(G)$ distinct sums but in the peculiar case where G is a balanced bipartite graph, in which case exactly $\chi(G)+1$ distinct sums have to be generated. 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.

The 1-2-3 Conjecture has recently been investigated from a decompositional angle, via so-called locally irregular decompositions, which are edge-partitions into locally irregular subgraphs. Through several recent studies, it was shown that this concept is quite related to the 1-2-3 Conjecture. However, the full connection between all those concepts was not clear. In [55], we propose an approach that generalizes all concepts above, involving coloured weights and sums. As a consequence, we get another interpretation of several existing results related to the 1-2-3 Conjecture. We also propose new related conjectures, to which we give some support.

7.4.7. Identifying codes

For G a graph or a digraph, let $\mathrm{id}(G)$ be the minimum size of an identifying code of G if one exists, and $\mathrm{id}(G) = +\infty$ otherwise. For a graph G, let $\mathrm{idor}(G)$ be the minimum of $\mathrm{id}(D)$ overall orientations D of G. In [20], we give some lower and upper bounds on $\mathrm{idor}(G)$. In particular, we show that $\mathrm{idor}(G) \leq 3/2\mathrm{id}(G)$ for every graph G. We also show that computing $\mathrm{idor}(G)$ is NP-hard, while deciding whether $\mathrm{idor}(G) \leq |V(G)| - k$ is polynomial-time solvable for every fixed integer k.

7.5. Digraph theory

Participants: Julien Bensmail, Frédéric Havet, Nicolas Nisse, William Lochet.

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.

7.5.1. 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 [59], 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 is polynomial-time solvable for all fixed positive integer. On the other hand, deciding whether a graph admits an ear decomposition without ears of length in F is NP-complete for any finite set 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 is polynomial-time solvable for all positive integer ; deciding whether a digraph admits a handle decomposition without handles of length in F is NP-complete for any finite set F of positive integers (and minimizing the number of handles of length in F is not approximable up to $n(1-\varepsilon)$); 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 A of integers, deciding whether a digraph has a handle decomposition with all handles of length in A is NP-complete, unless there exists $h \in \mathbf{N}$ such that $A = \{1, \cdots, h\}$.

7.5.2. Substructures in digraphs

We study substructures in digraphs. We study all kind of substructures: subdigraphs (induced or not), subdivision, immersion, minors, etc. We are both interested in the algorithmic point of view, that is determining the complexity of finding a (fixed or given) substructure in a given graph, and the structural point of view, that is finding sufficient conditions to guarantee the existence of a substructure.

In [32], we study the algorithmic complexity of the problem of deciding if a digraph contains a subdivision of a fixed digraph F. Up to 5 exceptions, we completely classify for which 4-vertex digraphs F, the F-subdivision problem is polynomial-time solvable and for which it is NP-complete. While all NP-hardness proofs are made by reduction from some version of the 2-linkage problem in digraphs, some of the polynomial-time solvable cases involve relatively complicated algorithms.

In [25], [22] we study conditions under which a digraph contain a subdivision of an oriented cycle. An oriented cycle is an orientation of a undirected cycle. We first show that for any oriented cycle C, there are digraphs containing no subdivision of C (as a subdigraph) and arbitrarily large chromatic number. In contrast, we show that for any C a cycle with two blocks, every strongly connected digraph with sufficiently large chromatic number contains a subdivision of C. We prove a similar result for the antidirected cycle on four vertices (in which two vertices have out-degree 2 and two vertices have in-degree 2). We study the existence of more general structures than cycles. A $(k_1 + k_2)$-bispindle is the union of $k_1(x, y)$-dipaths and $k_2(y, x)$-dipaths, all these dipaths being pairwise internally disjoint. The above-mentioned results on cycle with two blocks [25] can be restated as follows: for every $(1, 1)$-bispindle B, there exists an integer k such that every strongly connected digraph with chromatic number greater than k contains a subdivision of B. In [21], we investigate generalizations of this result by first showing constructions of strongly connected digraphs with large chromatic number without any $(3, 0)$-bispindle or $(2, 2)$-bispindle. We then consider $(2, 1)$-bispindles. Let $B(k_1, k_2; k_3)$ denote the $(2, 1)$-bispindle formed by three internally disjoint dipaths between two vertices x, y, two (x, y)-dipaths, one of length k_1 and the other of length k_2, and one (y, x)-dipath of length k_3. We conjecture that for any positive integers k_1, k_2, k_3, there is an integer $g(k_1, k_2, k_3)$ such that every strongly connected digraph with chromatic number greater than $g(k_1, k_2, k_3)$ contains a subdivision of $B(k_1, k_2; k_3)$. As evidence, we prove this conjecture for $k_2 = 1$ (and k_1, k_3 arbitrary).

In [36], we prove the existence of a function $h(k)$ such that every simple digraph with minimum outdegree greater than $h(k)$ contains an immersion of the transitive tournament on k vertices. This solves a conjecture of Devos, McDonald, Mohar and Scheide [72].

In [3], we study χ-bounded families of oriented graphs. A famous conjecture of Gyárfás and Sumner states for any tree T and integer k, if the chromatic number of a graph is large enough, either the graph contains a clique of size k or it contains T as an induced subgraph. We present some results and open problems about extensions of this conjecture to oriented graphs. In particular, we conjecture that for every oriented star S and integer k, if the chromatic number of a digraph is large enough, either the digraph contains a clique of size k or it contains S as an induced subgraph. As an evidence, we prove that for any oriented star S, every oriented graph with sufficiently large chromatic number contains either a transitive tournament of order 3 or S as an induced subdigraph. We then study for which sets $\mathcal{P}$ of orientations of P_4 (the path on four vertices) similar statements hold. We establish some positive and negative results.

7.5.3. Partitions of digraphs

We also study partitions of digraphs. Again we are interested in the algorithmic point of view, that is determining the complexity of finding a partition satisfying some properties in a digraph, and the structural point of view, that is finding sufficient conditions to guarantee the existence of such a partition.

For a given 2-partition (V_1, V_2) of the vertices of a (di)graph G, we study in [7] properties of the spanning bipartite subdigraph $B_G(V_1, V_2)$ of G induced by those arcs/edges that have one end in each V_i, $i \in \{1, 2\}$. 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 $\mathcal{NP}$-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 $B_D(V_1, V_2)$ is strongly connected. This holds even if we require the input to be a highly connected Eulerian digraph.

The dichromatic number $\overrightarrow{\chi}(D)$ of a digraph D is the least number k such that the vertex set of D can be partitioned into k parts each of which induces an acyclic subdigraph. Introduced by Neumann-Lara in 1982, this digraph invariant shares many properties with the usual chromatic number of graphs and can be seen as the natural analog of the graph chromatic number. In [14], we study the list dichromatic number of digraphs, giving evidence that this notion generalizes the list chromatic number of graphs. We first prove that the list dichromatic number and the dichromatic number behave the same in many contexts, such as in small digraphs (by proving a directed version of Ohba's Conjecture), tournaments, and random digraphs. We then consider bipartite digraphs, and show that their list dichromatic number can be as large as $\Omega(\log_2 n)$. We finally give a Brooks-type upper bound on the list dichromatic number of digon-free digraphs.

DANTE Project-Team

7. New Results

7.1. Graph Signal Processing and Machine Learning

Participants: Paulo Gonçalves, Esteban Bautista Ruiz, Mikhail Tsitsvero, Sarah de Nigris.

7.1.1. Analytic signal in many dimensions

In a series of two articles [30] and [54] (in collaboration with P. Borgnat), we extended analytic signal to the multidimensional case. First we showed how to obtain separate phase-shifted components and how to combine them into instantaneous amplitude and phase. Secondly we defined the proper hypercomplex analytic signal as a holomorphic hypercomplex function on the boundary of polydisk in the hypercomplex space. Next it was shown that the correct phase-shifted components can be obtained by positive frequency restriction of the Scheffers-Fourier transform based on the commutative and associative algebra generated by the set of elliptic hypercomplex numbers. Moreover we demonstrated that for d > 2 there is no corresponding Clifford-Fourier transform that allows to recover phase-shifted components correctly. Finally the euclidean-domain construction of instantaneous amplitude was extended to manifold and manifold-like graphs and point clouds.

7.1.2. BGP Zombies: an Analysis of Beacons Stuck Routes

Joint work with Romain Fontugne (IIJ Research Lab, Japan) and Patrice Abry (CNRS, Physics Lab of ENS de Lyon) [25].

Network operators use the Border Gateway Protocol (BGP) to control the global visibility of their networks. When withdrawing an IP prefix from the Internet, an origin network sends BGP withdraw messages, which are expected to propagate to all BGP routers that hold an entry for that address space in their routing table. Yet network operators occasionally report issues where routers maintain routes to IP prefixes withdrawn by their origin network. We refer to this problem as BGP zombies and characterize their appearance using RIS BGP beacons, a set of prefixes withdrawn every four hours at predetermined times. Across the 27 monitored beacon prefixes, we observe usually more than one zombie outbreak per day. But their presence is highly volatile, on average a monitored peer misses 1.8% withdraws for an IPv4 beacon (2.7% for IPv6). We also discovered that BGP zombies can propagate to other ASes, for example, zombies in a transit network are inevitably affect- ing its customer networks. **We employ a graph-based semi-supervised machine learning technique to estimate the scope of zombies propagation**, and found that most of the observed zombie outbreaks are small (i.e. on average 10% of monitored ASes for IPv4 and 17% for IPv6). We also report some large zombie outbreaks with almost all monitored ASes affected.

7.1.3. Design of graph filters and filterbanks

Book chapter [43], co-authored with Nicolas Tremblay (CNRS, UGA Gipsa-Lab) and Pierre Borgnat (CNRS, Physics Lab, ENS de Lyon).

Basic operations in graph signal processing consist in processing signals indexed on graphs either by filtering them or by changing their domain of representation, in order to better extract or analyze the important information they contain. The aim of this chapter is to review general concepts underlying such filters and representations of graph signals. We first recall the different Graph Fourier Transforms that have been developed in the literature, and show how to introduce a notion of frequency analysis for graph signals by looking at their variations. Then, we move to the introduction of graph filters, that are defined like the classical equivalent for 1D signals or 2D images, as linear systems which operate on each frequency band of a signal. Some examples of filters and of their implementations are given. Finally, as alternate representations of graph signals, we focus on multiscale transforms that are defined from filters. Continuous multiscale transforms such as spectral wavelets on graphs are reviewed, as well as the versatile approaches of filterbanks on graphs. Several variants of graph filterbanks are discussed, for structured as well as arbitrary graphs, with a focus on the central point of the choice of the decimation or aggregation operators.

7.2. Optimization

Participant: Marion Foare.

7.2.1. A new proximal method for joint image restoration and edge detection with the Mumford-Shah model

Joint work with Nelly Pustelnik (CNRS, Physics Lab of ENS de Lyon) and Laurent Condat (CNRS, GIPSA Lab) [24].

In this paper, we propose an adaptation of the PAM algorithm to the minimization of a nonconvex functional designed for joint image denoising and contour detection. This new functional is based on the Ambrosio–Tortorelli approximation of the well-known Mumford–Shah functional. We motivate the proposed approximation, offering flexibility in the choice of the possibly non-smooth penalization, and we derive closed form expression for the proximal steps involved in the algorithm. We focus our attention on two types of penalization: 1-norm and a proposed quadratic-1 function. Numerical experiments show that the proposed method is able to detect sharp contours and to reconstruct piecewise smooth approximations with low computational cost and convergence guarantees. We also compare the results with state-of-the-art re-laxations of the Mumford–Shah functional and a recent discrete formulation of the Ambrosio–Tortorelli functional.

7.2.2. Semi-Linearized Proximal Alternating Minimization for a Discrete Mumford–Shah Model

Joint work with Nelly Pustelnik (CNRS, Physics Lab of ENS de Lyon) and Laurent Condat (CNRS, GIPSA Lab) [51].

The Mumford–Shah model is a standard model in image segmentation and many approximations have been proposed in order to approximate it. The major interest of this functional is to be able to perform jointly image restoration and contour detection. In this work, we propose a general formulation of the discrete counterpart of the Mumford–Shah functional, adapted to nonsmooth penalizations, fitting the assumptions required by the Proximal Alternating Linearized Minimization (PALM), with convergence guarantees. A second contribution aims to relax some assumptions on the involved functionals and derive a novel Semi-Linearized Proximal Alternated Minimization (SL-PAM) algorithm, with proved convergence. We compare the performances of the algorithm with several nonsmooth penalizations, for Gaussian and Poisson denoising, image restoration and RGB-color denoising. We compare the results with state-of-the-art convex relaxations of the Mumford–Shah functional, and a discrete version of the Ambrosio–Tortorelli functional. We show that the SL-PAM algorithm is faster than the original PALM algorithm, and leads to competitive denoising, restoration and segmentation results.

7.2.3. Discrete Mumford-Shah on graph for mixing matrix estimation

Joint work with Yacouba Kaloga (Physics Lab of ENS de Lyon), Nelly Pustelnik (CNRS, Physics Lab of ENS de Lyon) and Pablo Jensen (CNRS, Physics Lab of ENS de Lyon) [53].

The discrete Mumford-Shah formalism has been introduced for the image denoising problem, allowing to capture both smooth behavior inside an object and sharp transitions on the boundary. In the present work, we propose first to extend this formalism to graphs and to the problem of mixing matrix estimation. New algorithmic schemes with convergence guarantees relying on proximal alternating minimization strategies are derived and their efficiency (good estimation and robustness to initialization) are evaluated on simulated data, in the context of vote transfer matrix estimation.

7.3. Wireless & Wired Networks

Participants: Thomas Begin, Anthony Busson, Isabelle Guérin Lassous.

7.3.1. Conflict graph-based model for IEEE 802.11 networks: A Divide-and-Conquer approach

WLANs (Wireless Local Area Networks) based on the IEEE 802.11 standard have become ubiquitous in our daily lives. We typically augment the number of APs (Access Points) within a WLAN to extend its coverage and transmission capacity. This leads to network densification, which in turn demands some form of coordination between APs so as to avoid potential misconfigurations. In our article [20], we describe a performance modeling method that can provide guidance for configuring WLANs and be used as a decision-support tool by a network architect or as an algorithm embedded within a WLAN controller. The proposed approach estimates the attained throughput of each AP, as a function of the WLAN's conflict graph, the AP loads, the frame sizes, and the link transmission rates. Our modeling approach employs a Divide-and-Conquer strategy which breaks down the original problem into multiple sub-problems, whose solutions are then combined to provide the solution to the original problem. We conducted extensive simulation experiments using the ns-3 simulator that show the model's accuracy is generally good with relative errors typically less than 10%. We then explore two issues of WLAN configuration: choosing a channel allocation for the APs and enabling frame aggregation on APs.

7.3.2. Video on Demand in IEEE 802.11p-based Vehicular Networks: Analysis and Dimensioning

This is a joint work with A. Boukerche. In [31], 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. Videos are delivered through IEEE 802.11p Road Side Units deployed along the highway. In this paper, we propose a simple analytical and yet accurate solution to estimate (at the speed of a click) two key performance parameters for a VoD platform: (i) the total amount of data down-loaded by a vehicle over its journey and (ii) the total "interruption time" , which corresponds to the time a vehicle spends with the playback of its video interrupted because of an empty buffer. After validating its accuracy against a set of simulations run with ns-3, we show an example of application of our analytical solution for the sizing of an IEEE 802.11p-based VoD platform.

7.3.3. An accurate and efficient modeling framework for the performance evaluation of DPDK-based virtual switches

This is a joint work with B. Baynat, G. Artero Gallardo and V. Jardin [4]. Data plane development kit (DPDK) works as a specialized library that enables virtual switches to accelerate the processing of incoming packets by, among other things, balancing the incoming flow of packets over all the CPU cores and processing packets by batches to make a better use of the CPU cache. Although DPDK has become a de facto standard, the performance modeling of a DPDK-based vSwitch remains a challenging problem. In this paper, we present an analytical queueing model to evaluate the performance of a DPDK-based vSwitch. Such a virtual equipment is represented by a complex polling system in which packets are processed by batches, i.e., a given CPU core processes several packets of one of its attached input queues before switching to the next one. To reduce the complexity of the associated model, we develop a general framework that consists in decoupling the polling system into several queueing subsystems, each one corresponding to a given CPU core. We resort to servers with vacation to capture the interactions between subsystems. Our proposed solution is conceptually simple, easy to implement and computationally efficient. Tens of comparisons against a discrete-event simulator show that our models typically deliver accurate estimates of the performance parameters of interest (e.g., attained throughput, packet latency or loss rate). We illustrate how our models can help in determining an adequate setting of the vSwitch parameters using several real-life case studies.

7.3.4. Association optimization in Wi-Fi networks

Densification of Wi-Fi networks has led to the possibility for a wireless station to choose between several access points (APs), improving coverage, wireless link quality and mobility. But densification of APs may generate interference, contention and decrease the global throughput as these APs have to share a limited number of channels. The recent trend in which Wi-Fi networks are managed in a centralized way offers

the opportunity to alleviate this problem through a global optimization of the resource usage. In particular, optimizing the association step between APs and stations can increase the overall throughput and fairness between stations. In this work, we propose an original solution to this optimization problem. First, we propose a mathematical model to evaluate and forecast the throughput achieved for each station for a given association. The best association is then defined as the one that maximizes a logarithmic utility function using the stations' throughputs predicted by the model. The use of a logarithmic utility function allows to achieve a good trade-off between overall throughput and fairness. A heuristic based on a local search algorithm is used to propose approximate solutions to this optimization problem. This approach has the benefit to be tuned according to the CPU and time constraints of the WLAN controller. A comparison between different heuristic versions and the optimum solution shows that the proposed heuristic offers solutions very close to the optimum with a significant gain of time.

In the first place, we consider a saturated network. Even if such traffic conditions are rare, the optimization of the association step under this assumption has the benefit to fairly share the bandwidth between stations. Nevertheless, traffic demands may be very different from one station to another and it may be more useful to optimize associations according to the stations' demands. In a second step, we propose an optimization of the association step based on the stations' throughputs and the channel busy time fraction (BTF). The latter is defined as the proportion of time the channel is sensed busy by an AP. We propose an analytical model that predicts BTF for any configuration. Associations are optimized in order to minimize the greatest BTF in the network. This original approach allows the Wi-Fi manager to unload the most congested AP, increase the throughput for most of the stations, and offer more bandwidth to stations that need it. We present a local search technique that finds local optima to this optimization problem. This heuristic relies on an analytical model that predicts BTF for any configuration. The model is based on a Markov network and a Wi-Fi conflict graph. NS-3 simulations including a large set of scenarios highlight the benefits of our approach and its ability to improve the performance in congested and non-congested Wi-Fi networks.

Lastly, we consider the latest amendments of the IEEE 802.11 standard. The main challenges are to propose models that take into account recent enhancements such as spatial multiplexing (MIMO) at the physical layer and frame aggregation mechanism at the MAC layer. To assess these new features, we derive an association optimization approach based on a new metric, named Hypothetical Busy Time Fraction (H-BTF), that combines the classical Busy Time Fraction (BTF) and the frame aggregation mechanism [3].

7.3.5. Transient analysis of idle time in VANETs using Markov-reward models

The development of analytical models to analyze the behavior of vehicular ad hoc networks (VANETs) is a challenging aim. Adaptive methods are suitable for many algorithms (*e.g.* choice of forwarding paths, dynamic resource allocation, channel control congestion) and services (*e.g.* provision of multimedia services, message dissemination). These adaptive algorithms help the network to maintain a desired performance level. However, this is a difficult goal to achieve, especially in VANETs due to fast position changes of the VANET nodes. Adaptive decisions should be taken according to the current conditions of the VANET. Therefore, evaluation of transient measures is required for the characterization of VANETs. In the literature, different works address the characterization and measurement of the idle (or busy) time to be used in different proposals to attain a more efficient usage of wireless network. We focus on the idle time of the link between two VANET nodes. Specifically, we have developed an analytical model based on a straightforward Markov reward chain (MRC) to obtain transient measurements of this idle time. Numerical results from the analytical model fit well with simulation results [12].

7.4. Performance Evaluation of Communication Networks

Participants: Thomas Begin, Philippe Nain, Isabelle Guérin Lassous.

7.4.1. First-Come-First-Served Queues with Multiple Servers and Customer Classes

This a joint work with A. Brandwajn [5]. We present a simple approach to the solution of a multi-server FCFS queueing system with several classes of customers and phase-type service time distributions. The proposed

solution relies on solving a single two-class model in which we distinguish one of the classes and we aggregate the remaining customer classes. We use a reduced state approximation to solve this two-class model. We propose two types of aggregation: exact, in which we merge the phase-type service time distributions exactly, and approximate, in which we simplify the phase-type distribution for the aggregated class by matching only its first two moments. The proposed approach uses simple mathematics and is highly scalable in terms of the number of servers, the number of classes, as well as the number of phases per class. Our approach applies both to queues with finite and infinite buffer space.

7.4.2. A study of systems with multiple operating levels, probabilistic thresholds and hysteresis

This a joint work with A. Brandwajn [6]. Current architecture of many computer systems relies on dynamic allocation of a pool of resources according to workload conditions to meet specific performance objectives while minimizing cost (e.g., energy or billing). In such systems, different levels of operation may be defined, and switching between operating levels occurs at certain thresholds of system congestion. To avoid rapid oscillations between levels of service, "hysteresis" is introduced by using different thresholds for increasing and decreasing workload levels, respectively. We propose a model of such systems with general arrivals, arbitrary number of servers and operating levels where each higher operating level may correspond to an arbitrary number of additional servers and soft (i.e. non-deterministic) thresholds to account for "inertia" in switching between operating levels. In our model, request service times are assumed to be memoryless and server processing rates may be a function of the current operating level and of the number of requests (users) in the system. Additionally, we allow for delays in the activation of additional operating levels. We use simple mathematics to obtain a semi-numerical solution of our model. We illustrate the versatility of our model using several case study examples inspired by features of real systems. In particular, we explore optimal thresholds as a tradeoff between performance and energy consumption.

7.4.3. Covert cycle stealing in an M/G/1 queue

Consider an M/G/1 queue where arriving jobs are under control of a party (Willie). There exists a second party, Alice who may or may not want to introduce a sequence of jobs to be serviced. Her goal is to prevent Willie from being able to distinguish between these two cases. The question that we address is: can Alice introduce her stream of jobs covertly, i.e., prevent Willie from distinguishing between the two possibilities, either her introducing the stream or not, and if so, at what rate can she introduce her jobs? We present a square-root law on the amount of service Alice can receive covertly. The covertness criterion is that the probabilities of false alarm and missed detection is arbitrarily close to one. One result we have established is the following: consider exponential service times for Alice's jobs and Willies' jobs with rate μ_1 and μ_2, respectively. During n Willie's job busy periods, Alice can submit covertly $O(\sqrt{n})$ jobs if $\mu_1 < 2\mu_2$, $O(\sqrt{n/\log n})$ jobs if $\mu_1 = 2\mu_2$, and $O(n^{\mu_1/\mu_2})$ jobs if $\mu_1 > 2\mu_2$. This is the first time that such a phase transition has been observed in this context. This ongoing research, carried out by P. Nain in collaboration with D. Towsley (Univ. Massachusetts) and B. Jiang (Shanghai Jiao Tong Univ.), has various applications in the context of service level agreement.

7.4.4. LRU caches

The work on network caches operating under the standard Least-Recently-Used (LRU) management policy, initiated in 2017 (see 2017 Dante Activity Report), has been completed and published [13]. Under weak statistical assumptions on the content request process, this work establishes the validity of the so-called "Che's approximation" as the cache size and the number of content go to infinity.

7.4.5. Stochastic Multilayer Networks

A stochastic multilayer network is the aggregation of M networks (one per layer) where each is a subgraph of a foundational network G. Each layer network is the result of probabilistically removing links and nodes from G. The resulting network includes any link that appears in at least K layers. This model is an instance of a non-standard site-bond percolation model. Two sets of results are obtained in [28]: first, we derive the probability distribution that the M-layer network is in a given configuration for some particular graph structures (explicit results are provided for a line and an algorithm is provided for a tree), where a configuration is the collective state of all links (each either active or inactive). Next, we show that for appropriate scalings of the node and link

selection processes in a layer, links are asymptotically independent as the number of layers goes to infinity, and follow Poisson distributions. Numerical results are provided to highlight the impact of having several layers on some metrics of interest (including expected size of the cluster a node belongs to in the case of the line). This model finds applications in wireless communication networks with multichannel radios, multiple social networks with overlapping memberships, transportation networks, and, more generally, in any scenario where a common set of nodes can be linked via co-existing means of connectivity.

7.5. Computational Human Dynamics and Temporal Networks

Participants: Márton Karsai, Éric Fleury, Jean-Philippe Magué, Philippe Nain, Jean-Pierre Chevrot.

7.5.1. Correlations and dynamics of consumption patterns in social-economic networks

In [16], we analyse a coupled dataset collecting the mobile phone communications and bank transactions history of a large number of individuals living in a Latin American country [16]. After mapping the social structure and introducing indicators of socioeconomic status, demographic features, and purchasing habits of individuals, we show that typical consumption patterns are strongly correlated with identified socioeconomic classes leading to patterns of stratification in the social structure. In addition, we measure correlations between merchant categories and introduce a correlation network, which emerges with a meaningful community structure. We detect multivariate relations between merchant categories and show correlations in purchasing habits of individuals. Finally, by analysing individual consumption histories, we detect dynamical patterns in purchase behaviour and their correlations with the socioeconomic status, demographic characters and the egocentric social network of individuals. Our work provides novel and detailed insight into the relations between social and consuming behaviour with potential applications in resource allocation, marketing, and recommendation system design.

7.5.2. Mapping temporal-network percolation to weighted, static event graphs

The dynamics of diffusion-like processes on temporal networks are influenced by correlations in the times of contacts. This influence is particularly strong for processes where the spreading agent has a limited lifetime at nodes: disease spreading (recovery time), diffusion of rumors (lifetime of information), and passenger routing (maximum acceptable time between transfers). In [14], we introduce weighted event graphs as a powerful and fast framework for studying connectivity determined by time-respecting paths where the allowed waiting times between contacts have an upper limit. We study percolation on the weighted event graphs and in the underlying temporal networks, with simulated and real-world networks. We show that this type of temporal-network percolation is analogous to directed percolation, and that it can be characterized by multiple order parameters.

7.5.3. Randomized reference models for temporal networks

Many real-world dynamical systems can successfully be analyzed using the temporal network formalism. Empirical temporal networks and dynamic processes that take place in these situations show heterogeneous, non-Markovian, and intrinsically correlated dynamics, making their analysis particularly challenging. Randomized reference models (RRMs) for temporal networks constitute a versatile toolbox for studying such systems. Defined as ensembles of random networks with given features constrained to match those of an input (empirical) network, they may be used to identify statistically significant motifs in empirical temporal networks (i.e. over-represented w.r.t. the null random networks) and to infer the effects of such motifs on dynamical processes unfolding in the network. However, the effects of most randomization procedures on temporal network characteristics remain poorly understood, rendering their use non-trivial and susceptible to misinterpretation. In the work presented in [52], we propose a unified framework for classifying and understanding microcanonical RRMs (MRRMs). We use this framework to propose a canonical naming convention for existing randomization procedures, classify them, and deduce their effects on a range of important temporal network features. We furthermore show that certain classes of compatible MRRMs may be applied in sequential composition to generate more than a hundred new MRRMs from existing ones surveyed in this article. We provide a tutorial for the use of MRRMs to analyze an empirical temporal network and we review applications of MRRMs found

in literature. The taxonomy of MRRMs we have developed provides a reference to ease the use of MRRMs, and the theoretical foundations laid here may further serve as a base for the development of a principled and systematic way to generate and apply randomized reference null models for the study of temporal networks.

7.5.4. Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis

Our usage of language is not solely reliant on cognition but is arguably determined by myriad external factors leading to a global variability of linguistic patterns. This issue, which lies at the core of sociolinguistics and is backed by many small-scale studies on face-to-face communication, is addressed in [29], by constructing a dataset combining the largest French Twitter corpus to date with detailed socioeconomic maps obtained from national census in France. We show how key linguistic variables measured in individual Twitter streams depend on factors like socioeconomic status, location, time, and the social network of individuals. We found that (i) people of higher socioeconomic status, active to a greater degree during the daytime, use a more standard language; (ii) the southern part of the country is more prone to use more standard language than the northern one, while locally the used variety or dialect is determined by the spatial distribution of socioeconomic status; and (iii) individuals connected in the social network are closer linguistically than disconnected ones, even after the effects of status homophily have been removed. Our results inform sociolinguistic theory and may inspire novel learning methods for the inference of socioeconomic status of people from the way they tweet.

7.5.5. Threshold driven contagion on weighted networks

Weighted networks capture the structure of complex systems where interaction strength is meaningful. This information is essential to a large number of processes, such as threshold dynamics, where link weights reflect the amount of influence that neighbours have in determining a node's behaviour. Despite describing numerous cascading phenomena, such as neural firing or social contagion, the modelling of threshold dynamics on weighted networks has been largely overlooked. We fill this gap in [21], by studying a dynamical threshold model over synthetic and real weighted networks with numerical and analytical tools. We show that the time of cascade emergence depends non-monotonously on weight heterogeneities, which accelerate or decelerate the dynamics, and lead to non-trivial parameter spaces for various networks and weight distributions. Our methodology applies to arbitrary binary state processes and link properties, and may prove instrumental in understanding the role of edge heterogeneities in various natural and social phenomena.

7.5.6. Link transmission centrality in large-scale social networks

Understanding the importance of links in transmitting information in a network can provide ways to hinder or postpone ongoing dynamical phenomena like the spreading of epidemic or the diffusion of information. In our work [22], we propose a new measure based on stochastic diffusion processes, the*transmission centrality*, that captures the importance of links by estimating the average number of nodes to whom they transfer information during a global spreading diffusion process. We propose a simple algorithmic solution to compute transmission centrality and to approximate it in very large networks at low computational cost. Finally we apply transmission centrality in the identification of weak ties in three large empirical social networks, showing that this metric outperforms other centrality measures in identifying links that drive spreading processes in a social network.

7.5.7. Prepaid or Postpaid? That Is the Question: Novel Methods of Subscription Type Prediction in Mobile Phone Services

In the paper [41], we investigate the behavioural differences between mobile phone customers with prepaid and postpaid subscriptions. Our study reveals that (a) postpaid customers are more active in terms of service usage and (b) there are strong structural correlations in the mobile phone call network as connections between customers of the same subscription type are much more frequent than those between customers of different subscription types. Based on these observations, we provide methods to detect the subscription type of customers by using information about their personal call statistics, and also their egocentric networks simultaneously. The key of our first approach is to cast this classification problem as a problem of graph labelling, which can be solved by max-flow min-cut algorithms. Our experiments show that, by using both

user attributes and relationships, the proposed graph labelling approach is able to achieve a classification accuracy of $\sim 87\%$, which outperforms by $\sim 7\%$ supervised learning methods using only user attributes. In our second problem, we aim to infer the subscription type of customers of external operators. We propose via approximate methods to solve this problem by using node attributes, and a two-way indirect inference method based on observed homophilic structural correlations. Our results have straightforward applications in behavioural prediction and personal marketing.

7.5.8. Service Adoption Spreading in Online Social Networks

The collective behaviour of people adopting an innovation, product or online service is commonly interpreted as a spreading phenomenon throughout the fabric of society. This process is arguably driven by social influence, social learning and by external effects like media. Observations of such processes date back to the seminal studies by Rogers and Bass, and their mathematical modelling has taken two directions: One paradigm, called simple contagion, identifies adoption spreading with an epidemic process. The other one, named complex contagion, is concerned with behavioural thresholds and successfully explains the emergence of large cascades of adoption resulting in a rapid spreading often seen in empirical data. The observation of real-world adoption processes has become easier lately due to the availability of large digital social network and behavioural datasets. This has allowed simultaneous study of network structures and dynamics of online service adoption, shedding light on the mechanisms and external effects that influence the temporal evolution of behavioural or innovation adoption. These advancements have induced the development of more realistic models of social spreading phenomena, which in turn have provided remarkably good predictions of various empirical adoption processes. In our chapter [39], we review recent data-driven studies addressing real-world service adoption processes. Our studies provide the first detailed empirical evidence of a heterogeneous threshold distribution in adoption. We also describe the modelling of such phenomena with formal methods and data-driven simulations. Our objective is to understand the effects of identified social mechanisms on service adoption spreading, and to provide potential new directions and open questions for future research.

7.5.9. Attention on Weak Ties in Social and Communication Networks

Granovetter's weak tie theory of social networks is built around two central hypotheses. The first states that strong social ties carry the large majority of interaction events; the second maintains that weak social ties, although less active, are often relevant for the exchange of especially important information (e.g., about potential new jobs in Granovetter's work). While several empirical studies have provided support for the first hypothesis, the second has been the object of far less scrutiny. A possible reason is that it involves notions relative to the nature and importance of the information that are hard to quantify and measure, especially in large scale studies. In our work [48], we search for empirical validation of both Granovetter's hypotheses. We find clear empirical support for the first. We also provide empirical evidence and a quantitative interpretation for the second. We show that attention, measured as the fraction of interactions devoted to a particular social connection, is high on weak ties—possibly reflecting the postulated informational purposes of such ties—but also on very strong ties. Data from online social media and mobile communication reveal network-dependent mixtures of these two effects on the basis of a platform's typical usage. Our results establish a clear relationships between attention, importance, and strength of social links, and could lead to improved algorithms to prioritize social media content.

DIANA Project-Team

6. New Results

6.1. Service Transparency

6.1.1. An Intelligent Sampling Framework for Controlled Experimentation and QoE Modeling

Participants: Muhammad Jawad Khokhar, Nawfal Abbasi Saber, Thierry Spetebroot, Chadi Barakat.

For internet applications, measuring, modeling and predicting the quality experienced by end users as a function of network conditions is challenging. A common approach for building application specific Quality of Experience (QoE) models is to rely on controlled experimentation. For accurate QoE modeling, this approach can result in a large number of experiments to carry out because of the multiplicity of the network features, their large span (e.g., bandwidth, delay) and the time needed to setup the experiments themselves. However, most often, the space of network features in which experimentations are carried out shows a high degree of similarity in the training labels of QoE. This similarity, difficult to predict beforehand, amplifies the training cost with little or no improvement in QoE modeling accuracy. So, in this work, funded by ANR BottleNet and IPL BetterNet, we aim to exploit this similarity, and propose a methodology based on active learning, to sample the experimental space intelligently, so that the training cost of experimentation is reduced. We validate our approach for the case of YouTube video streaming QoE modeling from out-of-band network performance measurements, and perform a rigorous analysis of our approach to quantify the gain of active sampling over uniform sampling. We first develop the methodology for an offline case where a pool of scenarios to experiment with is available. Then, we present an online variant that does not require a pool of scenarios, but finds automatically and in an online manner the best scenarios to experiment with. This latter variant outperforms the offline variant both in terms of accuracy and computation complexity. It is published in [22]. The overall methodology and its specification to both the offline and the online cases are published in [15].

6.1.2. A Methodology for Performance Benchmarking of Mobile Networks for Internet Video Streaming

Participants: Muhammad Khokhar, Thierry Spetebroot, Chadi Barakat.

Video streaming is a dominant contributor to the global Internet traffic. Consequently, gauging network performance w.r.t. the video Quality of Experience (QoE) is of paramount importance to both telecom operators and regulators. Modern video streaming systems, e.g. YouTube, have huge catalogs of billions of different videos that vary significantly in content type. Owing to this difference, the QoE of different videos as perceived by end users can vary for the same network Quality of Service (QoS). In this work, funded by ANR BottleNet and IPL BetterNet, we present a methodology for benchmarking performance of mobile operators w.r.t Internet video that considers this variation in QoE. We take a data-driven approach to build a predictive model using supervised machine learning (ML) that takes into account a wide range of videos and network conditions. To that end, we first build and analyze a large catalog of YouTube videos. We then propose and demonstrate a framework of controlled experimentation based on active learning to build the training data for the targeted ML model. Using this model, we then devise YouScore, an estimate of the percentage of YouTube videos that may play out smoothly under a given network condition. Finally, to demonstrate the benchmarking utility of YouScore, we apply it on an open dataset of real user mobile network measurements to compare performance of mobile operators for video streaming. This work is published in [21] and its extension to more sophisticated QoE models that consider other factors than interruptions is ongoing.

6.1.3. On the Cost of Measuring Traffic in a Virtualized Environment

Participants: Karyna Gogunska, Chadi Barakat.

The current trend in application development and deployment is to package applications and services within containers or virtual machines. This results in a blend of virtual and physical resources with complex network interconnection schemas mixing virtual and physical switches along with specific protocols to build virtual networks spanning over several servers. While the complexity of this setup is hidden by private/public cloud management solutions, e.g. OpenStack, this new environment constitutes a challenge when it comes to monitor and debug performance related issues. In this work carried out in collaboration with the Signet team of I3S with the support of the UCN@Sophia Labex, we introduce the problem of measuring traffic in a virtualized environment and focus on one typical scenario, namely virtual machines interconnected with a virtual switch. For this scenario, we assess the cost of continuously measuring the network traffic activity of the machines. Specifically, we seek to estimate the competition that exists to access the physical resources (e.g., CPU) of the physical server between the measurement task and the legacy application activity. This work was published in the IEEE Cloudnet 2018 conference [20] where it was awarded the Best Student Award. The collaboration with I3S is pursued towards a controlled configuration and deployment of measurements tools in a way to limit their impact on the legacy data plane of virtualized environments.

6.1.4. 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 2019. 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) are following the Digital Startup training from Inria/EM Lyon. This training helped formalize and improve the product market fit and the business model. We are also preparing the iLab competition.

The business model of ElectroSmart is to create an affiliation strategy to help companies selling product to reduce EMF exposure to find potential clients. Indeed, ElectroSmart users represent a highly qualified database of people concerned by EMF exposure. This database is invaluable to these companies as it is an emerging market and it is hard for these companies to make efficient marketing campaigns. The benefit for the ElectroSmart users is to have access to negotiated and validated solutions to reduce their EMF exposure. We are currently validating this market. We started our first affiliation campaign in December 2018 with the Spartan company that sells radiation blocking boxers. We already have two more planned campaigns in 2019, with a goal of 5 campaigns in 2019.

6.2. Open Network Architecture

6.2.1. Controller load in SDN networks

Participant: Damien Saucez.

In OpenFlow, a centralized programmable controller installs forwarding rules into switches to implement policies. However, this flexibility comes at the expense of extra overhead in signalling and number of rules to install. The community considered that it was essential to install all rules and strictly respect routing requirements, hence working on making extra fast and large memory switches and controllers. Instead we took an opposite direction and came with a new vision that leverages the SDN concept and considers the network as a black box where tailored rules should be used only for network traffic that really matters while for the rest a good-enough (sub-optimal but cheap) default behaviour should be enough. In the past, we applied this vision to limit the needed memory on network switches in [7]. Lately, we proposed solutions to limit the number of exchanged messages between the switches and the controller. More precisely, in [19], we developed a distributed sampling adaptive algorithm that allows switches to locally decide if they can contact the controller or if instead they should make their own decision locally. Numerical evaluation and emulation in Mininet demonstrate the benefit of the approach. The results were published in IEEE INFOCOM 2018, April 2018.

6.2.2. Resilient Service Function Chains in virtual networks

Participants: Ghada Moualla, Damien Saucez, Thierry Turletti.

Virtualization of network functions has led to the whole new concept of Service Function Chaining (SFC) that aims at building on the fly network services by deploying them in the Cloud. A vast literature proposes techniques to build virtual service chains and map them into physical infrastructure to maximize performance while reducing costs. However, the resiliency of chains is not investigated. However, such service chains are used for critical services like e-health or autonomous transportation systems and thus require high availability. Respecting some availability level is hard in general, but it becomes even harder if the operator of the service is not aware of the physical infrastructure that will support the service, which is the case when SFCs are deployed in multi-tenant data centers. With this work, we propose algorithms to solve the placement of topology-oblivious SFC demands such that placed SFCs respect availability constraints imposed by the tenant. In order to be practically usable, i.e., without knowledge on future demands, we leverage the structural properties of multi-tier data-center topologies such as Fat-Tree or Sine and Leaf topologies to build fast yet efficient online algorithms. We explored two radically different approaches: a deterministic one and a stochastic one and results show that both can be used in very large scale data-centers (i.e., 40k nodes or more) and our simulation results show that the algorithms are able to satisfy as many demands as possible by spreading the load between the replicas and enhancing the network resources utilization [23].

Initial results were published in IEEE International Conference on Cloud Networking 2018, October 2018.

6.2.3. Privacy preserving distributed services

Participants: Damien Saucez, Yevhenii Semenko, Alberto Zirondelli.

Blockchains are expected to help in reducing dependency on centralized platforms (e.g., Uber, Airbnb). With this internship, we have designed a protocol to make a fully distributed, secured, and privacy protecting taxi service – a distributed version of Uber. The analytical study shows that in such system the privacy protection comes with an important overhead in network communications which raises reasonable doubt on the feasibility of actually using fully distributed platforms in an "internet-scale environment" even though our implementation on Android phones shows that it is technically possible to build such systems. This work is done in collaboration with the GREDEG [0], that is evaluating the incentives for users to move to fully distributed platforms that are privacy preserving but that require the users to play an active role in the system.

6.2.4. P4Bricks: Enabling multiprocessing using Linker-based network data plane architecture

Participants: Hardik Soni, Thierry Turletti, Walid Dabbous.

Packet-level programming languages such as P4 usually require to describe all packet processing functionalities for a given programmable network device within a single program. However, this approach monopolizes the device by a single large network application program, which prevents possible addition of new functionalities by other independently written network applications. We propose P4Bricks, a system which aims to deploy and execute multiple independently developed and compiled P4 programs on the same reconfigurable hardware device. P4Bricks is based on a Linker component that merges the programmable parsers/deparsers and restructures the logical pipeline of P4 programs by refactoring, decomposing and scheduling the pipelines' tables. It merges P4 programs according to packet processing semantics (parallel or sequential) specified by the network operator and runs the programs on the stages of the same hardware pipeline, thereby enabling multiprocessing. We present the initial design of our system with an ongoing implementation and study P4 language's fundamental constructs facilitating merging of independently written programs [34], [12].

6.2.5. Applications in ITS Message Dissemination

Participants: Thierry Turletti.

[0] Groupe de Recherche en Droit, Economie, Gestion, a research center related to both the CNRS and the University of Nice-Sophia Antipolis and dealing with economic, managerial and legal aspects. See http://unice.fr/laboratoires/gredeg in French.

We build upon our prior work on D2-ITS, a flexible and extensible framework to dynamically distribute network control to enable message dissemination in Intelligent Transport Systems (ITS), and extend it with handover and load balancing capabilities. More specifically, D2-ITS' new handover feature allows a controller to automatically "delegate" control of a vehicle to another controller as the vehicle moves. Control delegation can also be used as a way to balance load among controllers and ensure that required application quality of service is maintained. We showcase D2-ITS' handover and load-balancing features using the Mininet-Wifi network simulator/emulator. Our preliminary experiments show D2-ITS' ability to seamlessly handover control of vehicles as they move. This work has been presented at the 27th International Conference on Computer Communications and Networks (ICCCN 2018), Jul 2018, Hangzhou, China [17].

6.2.6. 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 IEEE Transactions on Mobile Computing [16]. The part specifying the framework to a chunk-based scenario by accounting for partial storage of videos in vehicles was published in [25].

6.2.7. Cost Optimization of Cloud-RAN Planning and Provisioning for 5G Networks

Participants: Osama Arouk, Thierry Turletti.

We propose a network planning and provisioning framework that optimizes the deployment cost in C-RAN based 5G networks. Our framework is based on a Mixed Integer Quadratically Constrained Programming (MIQCP) model that optimizes "virtualized" 5G service chain deployment cost while performing adequate provisioning to address user demand and performance requirements. We use two realistic scenarios to showcase that our framework can be applied to different types of deployments and discuss the computational cost and scalability of our solution. This work has been presented at the IEEE International Conference on Communications, in May 2018, at Kansas City, MO, United States [18].

6.2.8. Slice Orchestration for Multi-Service Disaggregated Ultra Dense RANs

Participants: Osama Arouk, Thierry Turletti.

Ultra Dense Networks (UDNs) are a natural deployment evolution for handling the tremendous traffic increase related to the emerging 5G services, especially in urban environments. However, the associated infrastructure cost may become prohibitive. The evolving paradigm of network slicing can tackle such a challenge while optimizing the network resource usage, enabling multi-tenancy and facilitating resource sharing and efficient service-oriented communications. Indeed, network slicing in UDN deployments can offer the desired degree of customization in both vanilla Radio Access Network (RAN) designs, but also in the case of disaggregated multi-service RANs. We propose a novel multi-service RAN environment, i.e., RAN runtime, capable to support slice orchestration procedures and to enable flexible customization of slices as per tenant needs. Each network slice can exploit a number of services, which can either be dedicated or shared between multiple slices over a common RAN. The novel architecture we present concentrates on the orchestration and management systems. It interacts with the RAN modules, through the RAN runtime, via a number of new interfaces

enabling a customized dedicated orchestration logic for each slice. We present results for a disaggregated UDN deployment where the RAN runtime is used to support slice-based multi-service chain creation and chain placement, with an auto-scaling mechanism to increase the performance. This work has been published in IEEE Communications Magazine [13].

6.3. Experimental Evaluation

6.3.1. nepi-ng: an efficient experiment control tool in R2lab

Participants: Thierry Parmentelat, Thierry Turletti, Walid Dabbous, Mohamed Naoufal Mahfoudi.

Experimentation is an essential step for realistic evaluation of wireless network protocols. The evaluation methodology entails controllable environment conditions and a rigorous and efficient experiment control and orchestration for a variety of scenarios. Existing experiment control tools such as OMF often lack in efficiency in terms of resource management and rely on abstractions that hide the details about the wireless setup. We propose `nepi-ng`, an efficient experiment control tool that leverages job oriented programming model and efficient single-thread execution of parallel programs using asyncio. `nepi-ng` provides an efficient and modular fine grain synchronization mechanism for networking experiments with light software dependency footprint. This work has been presented at the 12th ACM International Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization (WINTECH) in November 2018 at New Delhi, India [24].

6.3.2. Using nepi-ng for Mesh Networks Experiments

Participants: Thierry Parmentelat, Thierry Turletti, Mohamed Naoufal Mahfoudi, Walid Dabbous.

We describe a demonstration run on R2lab, an open wireless testbed located in an anechoic chamber at Inria Sophia Antipolis. The demonstration consists in easily deploying a Wi-Fi mesh network. The nodes provisioning, configuration and the scenario orchestration and control are automatically done using the `nepi-ng` experiment orchestration tool. A performance comparison of two wireless mesh routing protocols in presence of controlled interference is shown. This demo has been presented at the 12th ACM International Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization (WINTECH) in November 2018 at New Delhi, India [32].

6.3.3. R2Lab Testbed Evaluation for Wireless Mesh Network Experiments

Participants: Farzaneh Pakzad, Thierry Turletti, Thierry Parmentelat Mohamed Naoufal Mahfoudi, Walid Dabbous.

We have provided critical evaluations of new potential testbeds for the evaluation of SDN-based WMNs. We evaluated the R2Lab wireless testbed platform at Inria Sophia Antipolis, France. This testbed has 37 customisable wireless devices in an anechoic chamber for reproducible research in wireless WiFi and 4G/5G networks. Our work presents the first initial evaluation of the testbed for wireless multi-hop experiments, using traditional WMN routing protocols. Our results demonstrate the potential for SDN experiments. We believe this is an important contribution in its own right, since experimental validation is a key research methodology in this context, and trust in the validity of experimental results is absolutely critical.

DIONYSOS Project-Team

7. New Results

7.1. Performance Evaluation

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

Stream Processing Systems. Stream processing systems are today gaining momentum as tools to perform analytics on continuous data streams. Their ability to produce analysis results with sub-second latencies, coupled with their scalability, makes them the preferred choice for many big data companies.

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies.

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

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

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

heads obtained by an item, we count the number of times the coin flipping process of an item has exceeded a given threshold. This threshold is defined by combining theoretical results in leader election and coupon collector problems. Results on simulated data show how impressive is the detection of the top-k items in a large range of distributions [38].

Throughput Prediction in Cellular Networks. Downlink data rates can vary significantly in cellular networks, with a potentially non-negligible effect on the user experience. Content providers address this problem by using different representations (e.g., picture resolution, video resolution and rate) of the same content and switch among these based on measurements collected during the connection. Knowing the achievable data rate before the connection establishment should definitely help content providers to choose the most appropriate representation from the very beginning. We have conducted several large measurement campaigns involving a panel of users connected to a production network in France, to determine whether it is possible to predict the achievable data rate using measurements collected, before establishing the connection to the content provider, on the operator's network and on the mobile node. We establish evidence that it is indeed possible to exploit these measurements to predict, with an acceptable accuracy, the achievable data rate. We thus introduce cooperation strategies between the mobile user, the network operator and the content provider to implement such anticipatory solution [23].

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

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

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

Probabilistic analysis of population protocols. In [20], we studied the well-known problem of dissemination of information in large scale distributed networks through pairwise interactions. This problem, originally called rumor mongering, and then rumor spreading has mainly been investigated in the synchronous model. This model relies on the assumption that all the nodes of the network act in synchrony, that is, at each round of the protocol, each node is allowed to contact a random neighbor. In the paper, we drop this assumption under the argument that it is not realistic in large scale systems. We thus consider the asynchronous variant,

where at random times, nodes successively interact by pairs exchanging their information on the rumor. In a previous paper, we performed a study of the total number of interactions needed for all the nodes of the network to discover the rumor. While most of the existing results involve huge constants that do not allow us to compare different protocols, we provided a thorough analysis of the distribution of this total number of interactions together with its asymptotic behavior. In this paper we extend this discrete time analysis by solving a conjecture proposed previously and we consider the continuous time case, where a Poisson process is associated to each node to determine the instants at which interactions occur. The rumor spreading time is thus more realistic since it is the real time needed for all the nodes of the network to discover the rumor. Once again, as most of the existing results involve huge constants, we provide tight bound and equivalent of the complementary distribution of the rumor spreading time. We also give the exact asymptotic behavior of the complementary distribution of the rumor spreading time around its expected value when the number of nodes tends to infinity.

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

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

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

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

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

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

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

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

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

Modeling loss processes in voice traffic. Markov models of loss incidents happening during packet voice communications are needful for many engineering tasks, namely network dimensioning and automatic quality assessment. Two very simple ones are Bernoulli and 2-state Markov models, but they carry limited information about incurred loss incidents. On the other hand, a general Markov loss model with 2^k states, where k is the window length used for observing the voice packet arrival process, leads to heavy computations and

an excessive lookahead delay. Moreover, legacy Markov loss models concentrate mostly on capturing some physical characteristics of loss incidents, rather than their perceived effects.

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

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

7.2. Machine learning

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

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

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

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

Machine Learning in Quality of Experience assessment. In a series of presentations we have disseminated the main ideas behind a new generation of Quality of Esperience assessing tools in preparation in the team. In the meetings [70] and [69], and also in the plenary [32], we described some of the key features of the tools

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

7.3. Network Economics

Participants: Bruno Tuffin, Patrick Maillé.

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team. The whole problem of network economics, from theory to practice, describing all issues and challenges, is described in our book [83].

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

Roaming. In October 2015, the European parliament has decided to forbid roaming charges among EU mobile phone users, starting June 2017, as a first step toward the unification of the European digital market. We discuss in [79] the impact consequences of such a measure.

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

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

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

Network neutrality. Most of our activity has been devoted to the vivid network neutrality debate, going beyond the traditional for or against neutrality, and trying to tackle it from different angles. We gave a tutorial on this topic [33], with a video available at https://www.youtube.com/watch?v=EaKtzxPHluU .

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

Another current important issue in the current net neutrality debate is that of sponsored data: With wireless sponsored data, a third party, content or service provider, can pay for some of your data traffic so that it is not counted in your plan's monthly cap. This type of behavior is currently under scrutiny, with telecommunication regulators wondering if it could be applied to prevent competitors from entering the market, and what the impact on all telecommunication actors can be. To answer those questions, we design and analyze in [55] a model where a Content Provider (CP) can choose the proportion of data to sponsor and a level of advertisement to get a return on investment, and several Internet Service Providers (ISPs) in competition. We distinguish three scenarios: no sponsoring, the same sponsoring to all users, and a different sponsoring depending on the ISP you have subscribed to. This last possibility may particularly be considered an infringement of the network neutrality principle. We see that sponsoring can be beneficial to users and ISPs depending on the chosen advertisement level. We also discuss the impact of zero-rating where an ISP offers free data to a CP to attract more customers, and vertical integration where a CP and an ISP are the same company.

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

7.4. Monte Carlo

Participants: Bruno Tuffin, Ajit Rai, Gerardo Rubino, Pierre L'Ecuyer.

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types.

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

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

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

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

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

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

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

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

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

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

Random variable generation. Random number generators were invented before there were symbols for writing numbers, and long before mechanical and electronic computers. All major civilizations through the ages found the urge to make random selections, for various reasons. Today, random number generators, particularly on computers, are an important (although often hidden) ingredient in human activity. We study in [18] the lattice structure of random number generators of the MIXMAX family, a class of matrix linear congruential generators that produce a vector of random numbers at each step. These generators were initially proposed and justified as close approximations to certain ergodic dynamical systems having the Kolmogorov K-mixing property, which implies a chaotic (fast-mixing) behavior. But for a K-mixing system, the matrix must have irrational entries, whereas for the MIXMAX it has only integer entries. As a result, the MIXMAX has a lattice structure just like linear congruential and multiple recursive generators. Its matrix entries were also selected in a special way to allow a fast implementation and this has an impact on the lattice structure. We study this lattice structure for vectors of successive and non-successive output values in various dimensions. We show in particular that for coordinates at specific lags not too far apart, in three dimensions, all the nonzero points lie in only two hyperplanes. This is reminiscent of the behavior of lagged-Fibonacci and AWC/SWB generators. And even if we skip the output coordinates involved in this bad structure, other highly structured projections often remain, depending on the choice of parameters.

7.5. Wireless Networks

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

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

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

Even if the support of IoT objects represents a real challenge to the access, as mentioned above, it is nevertheless important to support them effectively in the core network, this is one of the requirements of 5G networks. While this represents an interesting opportunity for operators to grow their business, it will need new mechanisms to scale and manage the envisioned high number of devices and their generated traffic. Particularity, the signaling traffic, which will overload the 5G core Network Function (NF) in charge of authentication and mobility, namely Access and Mobility Management Function (AMF). The objective of [34] is to provide an algorithm based on Control Theory allowing: (i) to equilibrate the load on the AMF instances in order to maintain an optimal response time with limited computing latency; (ii) to scale out or in the AMF instance (using NFV techniques) depending on the network load to save energy and avoid wasting resources. Obtained results via computer system indicate the superiority of our algorithm in ensuring fair load balancing while scaling dynamically with the traffic load.

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

Vehicular networks. According to recent forecasts, constant population growth and urbanization will bring an additional load of 2.9 billion vehicles to road networks by 2050. This will certainly lead to increased air pollution concerns, highly congested roads putting more strain on an already deteriorated infrastructure, and may increase the risk of accidents on the roads as well. Therefore, to face these issues we need not only to promote the usage of smarter and greener means of transportation but also to design advanced solutions that leverage the capabilities of these means along with modern cities' road infrastructure to maximize its utility. To this end, we propose, in [47], an original Cognitive Radio inspired algorithm, named CRITIC, that aims to mimic the principle of Cognitive Radio technology used in wireless networks on road networks. The key idea behind CRITIC is to temporarily grant regular vehicles access to priority (e.g., bus or carpool) lanes whenever they are underutilized in order to reduce road traffic congestion. In [50], we explore novel ways of utilizing inter-vehicle and vehicle to infrastructure communication technology to achieve a safe and efficient lane change manoeuvre for Connected and Autonomous Vehicles (CAVs). The need for such new protocols is due to the risk that every lane change manoeuvre brings to drivers and passengers lives in addition to its

negative impact on congestion level and resulting air pollution, if not performed at the right time and using the appropriate speed. To avoid this risk, we design two new protocols, one is built upon and extends an existing protocol, and it aims to ensure safe and efficient lane change manoeuvre, while the second is an original solution inspired from mutual exclusion concept used in operating systems. This latter complements the former by exclusively granting lane change permission in a way that avoids any risk of collision.

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

7.6. Future networks and architectures

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

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

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

In [63] we set design principles of future distributed edge clouds in order to meet application requirements. We precisely introduce a costless distributed resource allocation algorithm, named *CLOSE*, which considers local information only. We compare via simulations the performance of *CLOSE* against those obtained by using mechanisms proposed in the literature, notably the Tricircle project within OpenStack. It turns out that the proposed distributed algorithm yields better performance while requiring less overhead.

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

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

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

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

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

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

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

DYOGENE Project-Team

6. New Results

6.1. Energy Trade-offs for end-to-end Communications in Urban Vehicular Networks exploiting an Hyperfractal Model

In [33] presented this year at MSWIM DIVANet we show results on the trade-offs between the end-to-end communication delay and energy spent for completing a transmission in vehicular communications in urban settings. This study exploits our innovative model called " hyperfractal " that captures the self-similarity of the topology and vehicle locations in cities. We enrich the model by incorporating roadside infrastructure. We use analytical tools to derive theoretical bounds for the end-to-end communication hop count under two different energy constraints: either total accumulated energy, or maximum energy per node. More precisely, we prove that the hop count is bounded by $O(n(1-\alpha)/(dm-1))$ where $\alpha < 1$ and $m > 2$ is the precise hyperfractal dimension. This proves that for both constraints the energy decreases as we allow to chose among paths of larger length. In fact the asymptotic limit of the energy becomes significantly small when the number of nodes becomes asymptotically large. A lower bound on the network throughput capacity with constraints on path energy is also given. The results are confirmed through exhaustive simulations using different hyperfractal dimensions and path loss coefficients.

6.2. Broadcast Speedup in Vehicular Networks via Information Teleportation

In [32] presented this year at LCN our goal is to increase our understanding of the fundamental communication properties in urban vehicle-to-vehicle mobile networks by exploiting the self-similarity and hierarchical organization of modern cities. We use an innovative model called "hyperfractal" that captures the self-similarities of both the traffic and vehicle locations, and yet avoids the extremes of regularity and randomness. We use analytical tools to derive matching theoretical upper and lower bounds for the information propagation speed in an urban delay tolerant network (i.e., a network that is disconnected at all time, and thus uses a store-carry-and-forward routing model). We prove that the average broadcast time behaves as $n(1-\delta)$ (times a slowly varying function), where δ depends on the precise fractal dimension. Furthermore, we show that the broadcast speedup is due in part to an interesting self-similar phenomenon, that we denote as information teleportation. This phenomenon arises as a consequence of the topology of the vehicle traffic, and triggers an acceleration of the broadcast time. We show that our model fits real cities where open traffic data sets are available. The study presents simulations that confirm the validity of the bounds in multiple realistic settings, including scenarios with variable speed.

6.3. Vehicle-to-Infrastructure Communications Design in Urban Hyperfractals

In [25] presented at SPAWC our goal is to increase the awareness about the communication opportunities that arise in urban vehicle networks when exploiting the self-similarity and hierarchical organization of modern cities. The work uses our innovative model called " hyperfractal " that captures the self-similarity of the urban vehicular networks as well as incorporating roadside infrastructure with its own self-similarity. We use analytical tools to provide achievable trade-offs in operating the roadside units under the constraint of minimum routing path delay while maintaining a reasonably balanced load. The models and results are supported by simulations with different city hyperfractal dimensions in two different routing scenarios: nearest neighbor routing with no collision and minimum delay routing model assuming slotted Aloha, signal to interference ratio (SIR) capture condition, power-path loss, Rayleigh fading.

6.4. Book on Stochastic Geometry Analysis of Cellular Networks

In 2018 we have published a monograph [30] in which we explain the very latest analytic techniques and results from stochastic geometry for modelling the signal-to-interference-plus-noise ratio (SINR) distribution in heterogeneous cellular networks. This book is supposed to help readers to understand the effects of combining different system deployment parameters on key performance indicators such as coverage and capacity, enabling the efficient allocation of simulation resources. In addition to covering results for network models based on the Poisson point process, this book presents recent results for when non-Poisson base station configurations appear Poisson, due to random propagation effects such as fading and shadowing, as well as non-Poisson models for base station configurations, with a focus on determinantal point processes and tractable approximation methods. Theoretical results are illustrated with practical Long-Term Evolution (LTE) applications and compared with real-world deployment results.

6.5. Gibbsian On-Line Distributed Content Caching Strategy for Cellular Networks

In [9], we develop Gibbs sampling based techniques for learning the optimal content placement in a cellular network. A collection of base stations are scattered on the space, each having a cell (possibly overlapping with other cells). Mobile users request for downloads from a finite set of contents according to some popularity distribution. Each base station can store only a strict subset of the contents at a time; if a requested content is not available at any serving base station, it has to be downloaded from the backhaul. Thus, there arises the problem of optimal content placement which can minimize the download rate from the backhaul, or equivalently maximize the cache hit rate. Using similar ideas as Gibbs sampling, we propose simple sequential content update rules that decide whether to store a content at a base station based on the knowledge of contents in neighbouring base stations. The update rule is shown to be asymptotically converging to the optimal content placement for all nodes. Next, we extend the algorithm to address the situation where content popularities and cell topology are initially unknown, but are estimated as new requests arrive to the base stations. Finally, improvement in cache hit rate is demonstrated numerically.

6.6. Location Aware Opportunistic Bandwidth Sharing between Static and Mobile Users with Stochastic Learning in Cellular Networks

In [7], 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.

6.7. Performance analysis of cellular networks with opportunistic scheduling using queueing theory and stochastic geometry

In [38] submitted this year, combining stochastic geometric approach with some classical results from queuing theory,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.

6.8. The Influence of Canyon Shadowing on Device-to-Device Connectivity in Urban Scenario

In [48] submitted this year, 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.

6.9. Determinantal thinning of point processes with network learning applications

In [39] submitted this year, 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.

6.10. Analyzing LoRa long-range, low-power, wide-area networks using stochastic geometry

In [40] submitted this year, 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.

6.11. Statistical learning of geometric characteristics of wireless networks

In [41] to appear in Proc. INFOCOM 2019, 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.

6.12. Ressource allocation in bike sharing systems

Vehicle sharing systems are becoming an urban mode of transportation, and launched in many cities, as Velib' and Autolib' in Paris. Managing such systems is quite difficult. One of the major issues is the availability of the resources: vehicles or free slots. These systems became a hot topic in Operation Research and the importance of stochasticity on the system behavior leads us to propose mathematical stochastic models. The aim is to understand the system behavior and how to manage these systems in order to improve the allocation of both resources to users.

To improve BSS (bike-sharing systems), two types of policies can be deployed: incentives to the users to choose a better station, called *natural* or *green* regulation, or redistribution by trucks, called *active* regulation. In a simple mathematical model, we proved the efficiency of the 2-choice incentive policy for BSS (bike-sharing systems). The drawback of the model is that it ignores the geometry of the system, where the choice is only local. The purpose of this first work is to deal with this policy in real systems.

We use data trip data obtained from JCDecaux and reports on station status collected as open data, to test local choice policy. Indeed we designed and tested a new policy relying on a local small change in user behaviors, by adapting their trips to resource availability around their departure and arrival stations, based on 2-choice policy. Results show that, even with a small user collaboration, the proposed method increases significantly the global balance of the bike sharing system and therefore the user satisfaction. This is done using trip data sets and detecting spatial outliers, stations having a behavior significantly different from their spatial neighbors, in a context where neighbors are heavily correlated. For that we proposed an improved version of the well-known

Moran scatterplot method, using a robust distance metric called Gower similarity. Using this new version of Moran scatterplot, we show that, for the occupancy data set obtained by modifiying trips, the number of spatial outliers drastically decreases. We generalize this study with W. Ghanem and L. Massoulié testing incentive and redistribution policies on a simulator, where the tradeoff between the number of frustrated trips and the penalty for the users can be measured. We propose new versions of these policies including prediction.

6.13. Analyzing the choice of the least loaded queue between two neighboring queues

A model of N queues, with a local choice policy, is studied. Each one-server queue has a Poissonian arrival of customers. When a customer arrives at a queue, he joins the least loaded queue between this queue and the next one, ties solved at random. Service times have exponential distribution. The system is stable if the arrival-to-service rate ratio, also called load, is less than one. When the load tends to zero, we derive the first terms of the expansion in this parameter for the stationary probabilities that a queue has few customers. Then we provide explicit asymptotics, as the load tends to zero, for the stationary probabilities of the queue length. We used the analyticity of the stationary probabilities as a function of the load. It shows the behavior difference between this local choice policy and the 2-choice policy (*supermarket model*).

6.14. Optimal Content Replication and Request Matching in Large Caching Systems

We consider models of content delivery networks in which the servers are constrained by two main resources: memory and bandwidth. In such systems, the throughput crucially depends on how contents are replicated across servers and how the requests of specific contents are matched to servers storing those contents. In this paper, we first formulate the problem of computing the optimal replication policy which if combined with the optimal matching policy maximizes the throughput of the caching system in the stationary regime. It is shown that computing the optimal replication policy for a given system is an NP-hard problem. A greedy replication scheme is proposed and it is shown that the scheme provides a constant factor approximation guarantee. We then propose a simple randomized matching scheme which avoids the problem of interruption in service of the ongoing requests due to re-assignment or repacking of the existing requests in the optimal matching policy. The dynamics of the caching system is analyzed under the combination of proposed replication and matching schemes. We study a limiting regime, where the number of servers and the arrival rates of the contents are scaled proportionally, and show that the proposed policies achieve asymptotic optimality. Extensive simulation results are presented to evaluate the performance of different policies and study the behavior of the caching system under different service time distributions of the requests.

6.15. Statistical thresholds for Tensor PCA

This is a joint work with Aukosh Jagannath and Patrick Lopatto. We study the statistical limits of testing and estimation for a rank one deformation of a Gaussian random tensor. We compute the sharp thresholds for hypothesis testing and estimation by maximum likelihood and show that they are the same. Furthermore, we find that the maximum likelihood estimator achieves the maximal correlation with the planted vector among measurable estimators above the estimation threshold. In this setting, the maximum likelihood estimator exhibits a discontinuous BBP-type transition: below the critical threshold the estimator is orthogonal to the planted vector, but above the critical threshold, it achieves positive correlation which is uniformly bounded away from zero.

6.16. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning

This is a joint work with Andrea Montanari. The Lasso is a popular regression method for high-dimensional problems in which the number of parameters $\theta_1, \cdots, \theta_N$, is larger than the number n of samples: $N > n$. A

useful heuristics relates the statistical properties of the Lasso estimator to that of a simple soft-thresholding denoiser, in a denoising problem in which the parameters $(\theta_i)_{i\leq N}$ are observed in Gaussian noise, with a carefully tuned variance. Earlier work confirmed this picture in the limit $n, N \to \infty$, pointwise in the parameters θ, and in the value of the regularization parameter.

Here, we consider a standard random design model and prove exponential concentration of its empirical distribution around the prediction provided by the Gaussian denoising model. Crucially, our results are uniform with respect to θ belonging to ℓ_q balls, $q \in [0, 1]$, and with respect to the regularization parameter. This allows to derive sharp results for the performances of various data-driven procedures to tune the regularization.

Our proofs make use of Gaussian comparison inequalities, and in particular of a version of Gordon's minimax theorem developed by Thrampoulidis, Oymak, and Hassibi, which controls the optimum value of the Lasso optimization problem. Crucially, we prove a stability property of the minimizer in Wasserstein distance, that allows to characterize properties of the minimizer itself.

6.17. Phase transitions in spiked matrix estimation: information-theoretic analysis

We study here the so-called spiked Wigner and Wishart models, where one observes a low-rank matrix perturbed by some Gaussian noise. These models encompass many classical statistical tasks such as sparse PCA, submatrix localization, community detection or Gaussian mixture clustering. The goal of these notes is to present in a unified manner recent results (as well as new developments) on the information-theoretic limits of these spiked matrix/tensor models. We compute the minimal mean squared error for the estimation of the low-rank signal and compare it to the performance of spectral estimators and message passing algorithms. Phase transition phenomena are observed: depending on the noise level it is either impossible, easy (i.e. using polynomial-time estimators) or hard (information-theoretically possible, but no efficient algorithm is known to succeed) to recover the signal.

6.18. Accelerated decentralized optimization with local updates for smooth and strongly convex objectives

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 $\tau_{\max} = \Omega(\Delta_{\max})$. 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.

6.19. Group synchronization on grids

Group synchronization requires to estimate unknown elements $(\theta_v)_{v\in V}$ of a compact group $\mathbb{G}$ associated to the vertices of a graph $G = (V, E)$, using noisy observations of the group differences associated to the edges. This model is relevant to a variety of applications ranging from structure from motion in computer vision to graph localization and positioning, to certain families of community detection problems.

We focus on the case in which the graph G is the d-dimensional grid. Since the unknowns θ_v are only determined up to a global action of the group, we consider the following weak recovery question. Can we determine the group difference $\theta_u^{-1}\theta_v$ between far apart vertices u, v better than by random guessing? We prove that weak recovery is possible (provided the noise is small enough) for $d \geq 3$ and, for certain finite groups, for $d \geq 2$. Vice-versa, for some continuous groups, we prove that weak recovery is impossible for $d = 2$. Finally, for strong enough noise, weak recovery is always impossible.

6.20. An Impossibility Result for Reconstruction in a Degree-Corrected Planted-Partition Model

We consider the Degree-Corrected Stochastic Block Model (DC-SBM): a random graph on n nodes, having i.i.d. weights $(\phi_u)_{u=1}^n$ (possibly heavy-tailed), partitioned into $q \geq 2$ asymptotically equal-sized clusters. The model parameters are two constants $a, b > 0$ and the finite second moment of the weights $\Phi^{(2)}$. Vertices u and v are connected by an edge with probability $(\phi_u\phi_v/n)a$ when they are in the same class and with probability $(\phi_u\phi_v/n)b$ otherwise. We prove that it is information-theoretically impossible to estimate the clusters in a way positively correlated with the true community structure when $(a-b)2\Phi^{(2)} \leq q(a+b)$. As by-products of our proof we obtain (1) a precise coupling result for local neighbourhoods in DC-SBM's, that we use in a follow up paper [Gulikers et al., 2017] to establish a law of large numbers for local-functionals and (2) that long-range interactions are weak in (power-law) DC-SBM's.

6.21. On the capacity of information processing systems

We propose and analyze a family of information processing systems, where a finite set of experts or servers are employed to extract information about a stream of incoming jobs. Each job is associated with a hidden label drawn from some prior distribution. An inspection by an expert produces a noisy outcome that depends both on the job's hidden label and the type of the expert, and occupies the expert for a finite time duration. A decision maker's task is to dynamically assign inspections so that the resulting outcomes can be used to accurately recover the labels of all jobs, while keeping the system stable. Among our chief motivations are applications in crowd-sourcing, diagnostics, and experiment designs, where one wishes to efficiently learn the nature of a large number of items, using a finite pool of computational resources or human agents. We focus on the capacity of such an information processing system. Given a level of accuracy guarantee, we ask how many experts are needed in order to stabilize the system, and through what inspection architecture. Our main result provides an adaptive inspection policy that is asymptotically optimal in the following sense: the ratio between the required number of experts under our policy and the theoretical optimal converges to one, as the probability of error in label recovery tends to zero.

6.22. Optimal Algorithms for Non-Smooth Distributed Optimization in Networks

In this work, we consider the distributed optimization of non-smooth convex functions using a network of computing units. We investigate this problem under two regularity assumptions: (1) the Lipschitz continuity of the global objective function, and (2) the Lipschitz continuity of local individual functions. Under the local regularity assumption, we provide the first optimal first-order decentralized algorithm called multi-step primal-dual (MSPD) and its corresponding optimal convergence rate. A notable aspect of this result is that, for non-smooth 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 t 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. Under the global regularity assumption, we provide a simple yet efficient algorithm called distributed randomized smoothing (DRS) based on a local smoothing of the objective function, and show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where d is the underlying dimension.

6.23. Zap Meets Momentum: Stochastic Approximation Algorithms with Optimal Convergence Rate

There are two well known Stochastic Approximation techniques that are known to have optimal rate of convergence (measured in terms of asymptotic variance): the Ruppert-Polyak averaging technique, and stochastic Newton-Raphson (SNR)(a matrix gain algorithm that resembles the deterministic Newton-Raphson method). The Zap algorithms, introduced by Devraj and Meyn in 2017, are a version of SNR designed

to behave more closely like their deterministic cousin. It is found that estimates from the Zap Q-learning algorithm converge remarkably quickly, but the per-iteration complexity can be high. In [43], we introduce a new class of stochastic approximation algorithms based on matrix momentum. For a special choice of the matrix momentum and gain sequences, it is found in simulations that the parameter estimates obtained from the algorithm couple with those obtained from the more complex stochastic Newton-Raphson algorithm. Conditions under which coupling is guaranteed are established for a class of linear recursions. Optimal finite-n error bounds are also obtained.

6.24. Ergodic theory for controlled Markov chains with stationary inputs

Consider a stochastic process $\mathbf{X}$ on a finite state space $X = \{1, \cdots, d\}$. It is conditionally Markov, given a real-valued 'input process' ζ. This is assumed to be small, which is modeled through the scaling, $\zeta_t = \varepsilon\zeta^1_t$, $0 \le \varepsilon \le 1$, where ζ^1 is a bounded stationary process. The following conclusions are obtained, subject to smoothness assumptions on the controlled transition matrix and a mixing condition on ζ:

- A stationary version of the process is constructed, that is coupled with a stationary version of the Markov chain $\mathbf{X}^\bullet$ obtained with $\zeta \equiv 0$. The triple $(\mathbf{X}, \mathbf{X}^\bullet, \zeta)$ is a jointly stationary process satisfying $\mathsf{P}\{X(t) \ne X^\bullet(t)\} = O(\varepsilon)$. Moreover, a second-order Taylor-series approximation is obtained:

$$\mathsf{P}\{X(t) = i\} = \mathsf{P}\{X^\bullet(t) = i\} + \varepsilon^2 \varrho(i) + o(\varepsilon^2), \quad 1 \le i \le d,$$

  with an explicit formula for the vector $\varrho \in \mathfrak{R}^d$.

- For any $m \ge 1$ and any function $f : \{1, \cdots, d\} \times \mathfrak{R} \to \mathfrak{R}^m$, the stationary stochastic process $Y(t) = f(X(t), \zeta(t))$ has a power spectral density S_f that admits a second order Taylor series expansion: A function $\mathrm{S}^{(2)}{}_f : [-\pi, \pi] \to C^{m\times m}$ is constructed such that

$$\mathrm{S}_f(\theta) = \mathrm{S}^\bullet_f(\theta) + \varepsilon^2 \mathrm{S}^{(2)}{}_f(\theta) + o(\varepsilon^2), \quad \theta \in [-\pi, \pi].$$

  An explicit formula for the function $\mathrm{S}^{(2)}{}_f$ is obtained, based in part on the bounds in (i).

The results are illustrated using a version of the timing channel of Anantharam and Verdu.

6.25. Ordinary Differential Equation Methods for Markov Decision Processes and Application to Kullback–Leibler Control Cost

A new approach to computation of optimal policies for MDP (Markov decision process) models is introduced in [5], published in SICON this year. The main idea is to solve not one, but an entire family of MDPs, parameterized by a scalar ζ that appears in the one-step reward function. For an MDP with d states, the family of relative value functions $\{h^*_\zeta : \zeta \in \mathbb{R}\}$ is the solution to an ODE, $\frac{d}{d\zeta} h^*_\zeta = \mathcal{V}(h^*_\zeta)$, where the vector field $\mathcal{V} : R^d \to R^d$ has a simple form, based on a matrix inverse. Two general applications are presented: Brockett's quadratic-cost MDP model, and a generalization of the "linearly solvable" MDP framework of Todorov in which the one-step reward function is defined by Kullback–Leibler divergence with respect to nominal dynamics. This technique was introduced by Todorov in 2007, where it was shown under general conditions that the solution to the average-reward optimality equations reduce to a simple eigenvector problem. Since then many authors have sought to apply this technique to control problems and models of bounded rationality in economics. A crucial assumption is that the input process is essentially unconstrained. For example, if the nominal dynamics include randomness from nature (eg, the impact of wind on a moving vehicle), then the optimal control solution does not respect the exogenous nature of this disturbance. In [16] we introduce a technique to solve a more general class of action-constrained MDPs.

6.26. Distributed control design for balancing the grid using flexible loads

Inexpensive energy from the wind and the sun comes with unwanted volatility, such as ramps with the setting sun or a gust of wind. Controllable generators manage supply-demand balance of power today, but this is becoming increasingly costly with increasing penetration of renewable energy. It has been argued since the 1980s that consumers should be put in the loop: "demand response" will help to create needed supply-demand balance. However, consumers use power for a reason and expect that the quality of service (QoS) they receive will lie within reasonable bounds. Moreover, the behavior of some consumers is unpredictable, while the grid operator requires predictable controllable resources to maintain reliability.

The goal of the book chapter [31] is to describe an emerging science for demand dispatch that will create virtual energy storage from flexible loads. By design, the grid-level services from flexible loads will be as controllable and predictable as a generator or fleet of batteries. Strict bounds on QoS will be maintained in all cases. The potential economic impact of these new resources is enormous. California plans to spend billions of dollars on batteries that will provide only a small fraction of the balancing services that can be obtained using demand dispatch. The potential impact on society is enormous: a sustainable energy future is possible with the right mix of infrastructure and control systems.

In [17], presented at IEEE CDC 2018, a natural notion of *energy capacity* is proposed for the special case of thermostatically controlled loads (TCLs). It is shown that this quantity is closely approximated by thermal energy capacity, which is a component of the "leaky battery model" introduced in prior work. Simulation experiments in a distributed control setting show that these energy limits, and accompanying power capacity limits, are reliable indicators of online capacity, even for a heterogeneous population of loads. A feedforward/feedback control scheme is proposed for a large collection of heterogeneous loads. At the local level, control loops are used to create cooperative responses from each load in a given class of homogeneous loads. This simplifies control of the aggregate based on two pieces of information: aggregate power consumption from each class of loads and the *state of charge* surrogate that is a part of the leaky battery model. This information is required at a slow time-scale (say, 5 minute sampling).

In [18], we study the problem of coordination of a collection of on/off thermostatically controlled loads (TCLs) to act as a "virtual battery". Virtual Energy Storage (VES) is provided by the collection by either consuming more (charging) or less (discharging) power than the baseline. VES can be an inexpensive alternative to batteries when a large share of the electricity comes from volatile sources such as solar and wind. Almost all prior work has assumed that the outside weather - which significantly effects a TCLs behavior - is constant. We combine the above distributed load control design with a grid level MPC (model predictive control) that uses predictions of disturbances (weather) over a planning horizon. Additionally, irrespective of the choice of control architecture, there is a fundamental limit to the power and energy capacity of the collection of TCLs. We partially address this issue by scaling the reference signal by a function of the outside air temperature.

6.27. Estimation and control of quality of service in demand dispatch

Flexibility of energy consumption can be harnessed for the purposes of grid-level ancillary services. In particular, through distributed control of a collection of loads, a balancing authority regulation signal can be tracked accurately, while ensuring that the quality of service (QoS) for each load is acceptable on average. Subject to distributed control approaches advocated in recent research, the histogram of QoS is approximately Gaussian, and consequently, each load will eventually receive poor service. In [11], published this year in IEEE Transactions on Smart Grid, statistical techniques are developed to estimate the mean and variance of QoS as a function of the power spectral density of the regulation signal. It is also shown that additional local control can eliminate risk. The histogram of QoS is truncated through this local control, so that strict bounds on service quality are guaranteed. While there is a tradeoff between the grid-level tracking performance (capacity and accuracy) and the bounds imposed on QoS, it is found that the loss of capacity is minor in typical cases.

The previous designs for distributed control of TCLs ensure that the indoor temperature remains within a pre-specified bound, but other QoS metrics, especially the frequency of turning on and off was not limited. In [19], presented at ACM BuildSys 2018, we propose a more advanced control architecture that reduces the

cycling rate of TCLs. We show through simulations that the proposed controller is able to reduce the cycling of individual TCLs compared to the previous designs with little loss in tracking of the grid-supplied reference signal.

6.28. Optimal control of energy storage

Energy storage revenue estimation is essential for analyzing financial feasibility of investment in batteries. In [22], we quantify the cycles of operation considering depth-of-discharge (DoD) of operational cycles and provide an algorithm to calculate equivalent 100% DoD cycles. This facilitates in comparing cycles of different DoDs. The battery life is frequently defined as a combination of cycle and calendar life. We propose a battery capacity degradation model based on the cycle and the calendar life and operational cycles. Using equivalent 100% DoD cycles and revenue generated, we calculate the dollars per cycle revenue of storage performing electricity price based arbitrage and ancillary services for load balancing in real time. Using PJM's (a regional transmission organization in the United States) real data we calculate short term and long term financial potential for the year of 2017. We observe that participating in ancillary services is significantly more beneficial for storage owners compared to participating in energy arbitrage.

Battery life is often described a combination of cycle life and calendar life. In [21], we propose a mechanism to limit the number of cycles of operation over a time horizon in an optimal arbitrage algorithm proposed in our previous work. The cycles of operation have to be tuned based on price volatility to maximize the battery life and arbitrage gains.

In [23], we analyze the effect of real time electricity price (RTP) on the amount of ancillary services required for load balancing in presence of responsive users, information asymmetry and forecast errors in demand and renewable energy sources (RES) generation. We consider a RTP that is determined by the forecasted generation and ramping cost. A community choice aggregator manages the load of all the consumers by setting the price. The consumer's objective is to minimize their overall cost of consumption. Ancillary services are called upon to balance the load in real time. With zero RES in the power network and a high degree of load flexibility, the proposed RTP flattens and the volatility in demand vanishes. However, in presence of RES the volatility in price and demand is reduced up to an extent and ancillary services are required for load balancing. The amount of ancillary services required increases with forecast errors. We also propose a real time algorithm that approximates the optimal consumer behavior under the complete information setting. Extensive numerical simulations are provided using real data from Pecan Street and Elia Belgium.

6.29. Dynamic matching models

The model of First Come First Served infinite bipartite matching was introduced in Caldentey, Kaplan and Weiss, 2009. In this model, there is a sequence of items that are chosen i.i.d. from a finite set $\mathcal{C}$ and an independent sequence of items that are chosen i.i.d. from a finite set $\mathcal{S}$, and a bipartite compatibility graph G between $\mathcal{C}$ and $\mathcal{S}$. Items of the two sequences are matched according to the compatibility graph, and the matching is FCFS, meaning that each item in the one sequence is matched to the earliest compatible unmatched item in the other sequence. In Adan and Weiss, 2012, a Markov chain associated with the matching was analyzed, a condition for stability was derived, and a product form stationary distribution was obtained. In [2], we present several new results that unveil the fundamental structure of the model. First, we provide a pathwise Loynes' type construction which enables to prove the existence of a unique matching for the model defined over all the integers. Second, we prove that the model is dynamically reversible: we define an exchange transformation in which we interchange the positions of each matched pair, and show that the items in the resulting permuted sequences are again independent and i.i.d., and the matching between them is FCFS in reversed time. Third, we obtain product form stationary distributions of several new Markov chains associated with the model. As a by-product, we compute useful performance measures, for instance the link lengths between matched items.

In [51], we propose an explicit construction of the stationary state of Extended Bipartite Matching (EBM) models, as defined in (Busic et. al., 2013). We use a Loynes-type backwards scheme similar in flavor to that in (Moyal et al., 2017), allowing to show the existence and uniqueness of a bi-infinite perfect matching under various conditions, for a large class of matching policies and of bipartite matching structures. The key algebraic element of our construction is the sub-additivity of a suitable stochastic recursive representation of the model, satisfied under most usual matching policies. By doing so, we also derive stability conditions for the system under general stationary ergodic assumptions, subsuming the classical markovian settings.

In [42], we consider holding costs for the items that are waiting to be matched. We model this problem as an MDP (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 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 preliminary numerical experiments, threshold-type policies performs also very well for more general bipartite graphs.

EVA Project-Team

7. New Results

7.1. From SmartMarina to Falco

Participants: Keoma Brun-Laguna, Thomas Watteyne.

SmartMarina project (http://smartmarina.org/) was a technical project in 2017 to study the feasibility of using the wireless technology developed at Inria-EVA for marina management. In 2018, the Wattson Elements company was born, which now commercializes the Falco solution (https://wefalco.fr/).

Figure 1. Screenshot of the Falco promotional video, https://youtu.be/35HdoFLrCf0.

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 2018, the working group published the specification of the 6TiSCH Operation Sublayer (6top) Protocol, RFC8480. Work is also ongoing in the fragment forwarding space, where we are working on how ot efficient forward long IPv6 packets which are fragmented to fit in short IEEE 802.15.4 frames.

7.3. 6TiSCH Security

Participants: Malisa Vucinic, Thomas Watteyne.

The security work of Inria-EVA revolves around 6TiSCH networks and is a continuation of the efforts started during the H2020 ARMOUR project. The work focused on stabilizing the “Minimal Security” solution that has now passed the working group last call in the IETF and is pending finals reviews before being published as an RFC. 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 results in the following action. Although seemingly different, they all contribute to the overall goal of better understand (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. The testbed is depicted in Fig. 2 . You can access its management interface at http://testbed.openwsn.org/.

Figure 2. The OpenTestbed deployed in Inria Paris since July 2018.

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 (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. IoT and Wireless Sensor Networks

More than 50 billions of devices will be connected in 2020. This huge infrastructure of devices, which is managed by highly developed technologies, is called Internet of Things (IoT). The latter provides advanced services, and brings economical and societal benefits. This is the reason why engineers and researchers of both industry and scientific communities are interested in this area. The Internet of Things enables the interconnection of smart physical and virtual objects, managed by highly developed technologies. WSN (Wireless Sensor Network), is an essential part of this paradigm. The WSN uses smart, autonomous and usually limited capacity devices in order to sense and monitor their environment.

7.5.1. Distributed Scheduling for IEEE 802.15.4e TSCH networks

Participants: Yasuyuki Tanaka, Pascale Minet, Thomas Watteyne.

Since the scheduling algorithm is not standardized for IEEE 802.15.4e TSCH networks, many scheduling algorithms have been proposed. Most of them are centralized, few are distributed. Among the distributed scheduling algorithms, 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.

In the study funded by the Inria ADT DASMU (Action de Developement Technologique Distributed Adaptive Scheduling for MUltichannel wireless sensor networks), we focus on a distributed scheduling algorithm that relies on realistic assumptions, does not require complex computation, is valid for any traffic load, is adaptive and compliant with the standardized protocols used in the 6TiSCH working group at IETF.

First results have been obtained and 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. More evaluations are needed to improve the proposal (e.g. less packet drops during transient situations, less overhead) in terms of scheduled cells).

7.5.2. IoT and IEEE 802.15.4e TSCH networks

Participants: Pascale Minet, Ines Khoufi, Zied Soua.

In 2018, we focus on how an IEEE 802.15.4e is autonomously built and how nodes join the network.

To join the TSCH network, a device randomly selects a physical channel used by this network and listens to a beacon advertising this network. Since the physical channel on which the beacon is broadcast changes at each beacon slot due to channel hopping, the joining device will eventually hear a beacon sent by one of its neighbors. Upon receipt of a valid beacon, this device gets synchronized with the TSCH network.

In this study, we focus on the time needed by a node to detect a beacon sent by a TSCH network, as well as on the time needed to build a TSCH network. These times are important for industrial applications where new nodes are inserted progressively, or when failed nodes are replaced. Both times highly depend on the beacon advertisement policy, policy that is not specified in the standard and is under the responsibility of a layer upper than the MAC one. Since beacons are broadcast, they are lost in case of collisions: the vital information they carry is lost. The main problem is how to avoid collisions between two devices that are not neighbors.

That is why we propose the Enhanced Deterministic Beacon Advertising algorithm, called EDBA, that ensures a collision-free advertising of beacons. Since the beacon cells are fairly distributed in the slotframe, the average joining time is minimized. The behavior of a joining node has been modeled by a Markov chain from which the average joining time is computed, taking into account the reliability of wireless links. An intensive performance evaluation based on NS3 simulations allows us to validate this model and conclude on the very good performance of EDBA, even when compared with MBS, considered as the best advertising algorithm in the literature. These results have been published in the Annals of Telecommunications, [10].

7.5.3. UAV-based Data Gathering

Participants: Nadjib Achir (Paris 13), Tounsia Djamah, Paul Muhlethaler, Celia Tazibt (Paris 13).

The recent advances in wireless sensors and Unmanned Aerial Vehicles have created new opportunities for environmental control and low cost aerial data gathering. We propose to use an Unmanned Aerial Vehicle (UAV) for data gathering [36]. Basically, we have proposed a method for UAV path planning based on virtual forces and potential fields. In addition, and more importantly, we present a new approach to compute the attractive forces of the potential field.

We use as our starting point the idea used by Pereira of using a potential field approach. However, we extend this work by considering that each cell in the area apply an attractive force on the drone, not only the deployed sensors. We compared our results with those obtained with Pereira's method and we obtained better performance in terms of data collection time. In other words, for the same period of time our method collect more data. The second advantage of our approach is that it leads to a significant reduction in the distance that the drone must travel.

7.5.4. Towards evaluating Named Data Networking for the IoT: A framework for OMNeT++

Participants: Amar Abane, Samia Bouzefrane (Cnam), Paul Muhlethaler.

Named Data Networking is a promising architecture for emerging Internet applications such as the Internet of Things (IoT). Many studies have already investigated how NDN can be an alternative for IP in future IoT deployments. However, NDN-IoT propositions need accurate evaluation at network level and system level as well. We introduce an NDN framework for OMNeT++ [29]. Designed for low-end devices and gateways of the IoT, the framework is capable of simulating NDN scenarios at the boundary of the network and the system. The framework implementation is presented and used to study a typical aspect of NDN integration in IoT devices.

7.5.5. Evaluation of LORA with stochastic geometry

Participants: Bartek Blaszczyszyn (Dyogen), 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 [39]. 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.

7.5.6. Position Certainty Propagation: A location service for MANETs

Participants: Abdallah Sobehy, Paul Muhlethaler, Eric Renault (Telecom Sud-Paris).

Localization in Mobile Ad-hoc Networks (MANETs) and Wireless Sensor Networks (WSNs) is an issue of great interest, especially in applications such as the IoT and VANETs. We propose a solution that overcomes two limiting characteristics of these types of networks. The first is the high cost of nodes with a location sensor (such as GPS) which we will refer to as anchor nodes. The second is the low computational capability of nodes in the network. The proposed algorithm [28] addresses two issues; self-localization where each non-anchor node should discover its own position, and global localization where a node establishes knowledge of the position of all the nodes in the network. We address the problem as a graph where vertices are nodes in the network and edges indicate connectivity between nodes. The weights of edges represent the Euclidean distance between the nodes. Given a graph with at least three anchor nodes and knowing the maximum communication range for each node, we are able to localize nodes using fairly simple computations in a moderately dense graph.

7.6. Industry 4.0 and Low-Power Wireless Meshed Networks

7.6.1. Deterministic Networking for the Industrial Internet of Things (IIoT)

Participants: Keoma Brun-Laguna, Thomas Watteyne, Pascale Minet.

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 sense, but many challenges remain as industrial application have strong operational requirements. This section of the IoT application is called Industrial IoT (IIoT).

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.

Then come latency and energy-efficiency requirements. As devices are battery-powered, they need to consume as little as possible to be able to operate during years. The next step for the IoT is to target time-critical applications.

Industrial IoT technologies are now adopted by companies over the world, and are now a proven solution. Yet, challenges remain and some of the limits of the technologies are still not fully understood. In his PhD Thesis, Keoma Brun-Laguna addresses TSCH-based Wireless Sensor Networks and studies their latency and lifetime limits under real-world conditions.

We gathered 3M network statistics 32M sensor measurements on 11 datasets with a total of 170,037 mote hours in real-world and testbeds deployments. We assembled what we believed to be the largest dataset available to the networking community.

Based on those datasets and on insights we learned from deploying networks in real-world conditions, we study the limits and trade-offs of TSCH-based Wireless Sensor Networks. We provide methods and tools to estimate the network performances of such networks in various scenarios. We highlight the trade-off between short latency and long network lifetime. We believe we assembled the right tools for protocol designer to build deterministic networking to the Industrial IoT.

7.6.2. Industry 4.0 and IEEE 802.15.4e TSCH networks

Participants: Pascale Minet, Ines Khoufi, Zied Soua.

By the year 2020, it is expected that the number of connected objects will exceed several billions devices. These objects will be present in everyday life for a smarter home and city as well as in future smart factories that will revolutionize the industry organization. This is actually the expected fourth industrial revolution, more known as Industry 4.0. In which, the Internet of Things (IoT) is considered as a key enabler for this major transformation. IoT will allow more intelligent monitoring and self-organizing capabilities than traditional factories. As a consequence, the production process will be more efficient and flexible with products of higher quality.

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. 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 enables to mitigate multipath fading and interferences.

To increase the flexibility and the self-organizing capacities required by Industry 4.0, 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 the flexibility of the schedule ruling all network communications is needed.

In 2018, we show how a TSCH network can adapt to such changes. More precisely, we propose a solution ranging from network construction to data gathering. We show how a TSCH network is autonomously built, supports data gathering and is able to adapt to changes in network topology, traffic and application requirements.

The solution proposed preserves the merits of TSCH network, that can be listed hereafter. The time-slotted multichannel medium access enables parallel transmissions on several channels, leading to shorter latency and higher throughputs. In addition, channel hopping mitigates interference and multipath effects. Furthermore, since transmissions are scheduled, a conflict-free schedule is computed by the network coordinator (i.e. the CPAN). Hence, no collision occurs during data gathering. The absence of collision leads to a higher throughput, because there is no retransmission due to collisions. It also preserves nodes power autonomy.

This simple solution is based on the coexistence of several periodic slotframes. We distinguish three slotframes, which are the Beacon Slotframe, the Data Slotframe and the Shared Slotframe. The network schedule corresponds to the superposition of the three schedules given by each slotframe, where the slotframe with the highest priority wins.

This solution ensures a collision-free dissemination over the whole network. Beacons are broadcast in sequence by increasing depth of devices. This broadcast is also used to disseminate Data Schedules (new schedule or update).

In addition, this solution is adaptive. Topology, traffic or application changes are notified to the CPAN. Depending on the changes notified, the CPAN updates the current schedule or recomputes a new one. Shared slots are used to cope with unexpected events.

We compute the theoretical bounds with regard to key performance indicators and compare them with the values obtained by NS3 simulation. Simulation results confirm the theoretical upper bounds computed for network construction and data gathering. Hence, TSCH networks are able to adapt to traffic or topology changes in a reasonable time which is a strong requirement of Industry 4.0 applications. These results have been presented at the PEMWN 2018 conference in [26]. In some further work, we will study how to improve this delay to support the most demanding applications.

7.7. Machine Learning for an efficient and dynamic management of data centers

7.7.1. Data Analysis in Data Centers

Participants: Eric Renault (Telecom Sud-Paris), Selma Boumerdassi (Cnam), Pascale Minet, Ines Khoufi.

In High Performance Computing (HPC), it is assumed that all machines are homogeneous in terms of CPU and memory capacities, and that the tasks making up the jobs have similar resource requests. It has been shown that this homogeneity relating both to machine capacity and workload, although generally valid for HPC, does no longer apply to data centers. This explains why the publication of data gathered in an operational Google data center over 29 days has aroused great interest among researchers.

It is crucial to have real traces of a Google data center publicly available that are representative of the functioning of real data centers. Our goal is to analyze the data collected and to draw useful conclusions about machines, jobs and tasks as well as resource usage. Our main results have been published in [25], [24] and can be summarized as follows:

- Although 92% of machines have a CPU capacity of 0.5, there are 10 machine configurations in the data center, each configuration is characterized by a pair $(CPU\ capacity, memory\ capacity)$. The most frequent configuration is supported by only 53% of machines.
- Over the 29 days, all the machines in the data center that were removed, were restarted later after an off-period. 50% of these periods have a duration less than or equal to 1000 seconds (i.e. 16.66 minutes), suggesting a maintenance operation.
- The distribution of jobs per category reveals only one job, representing 0.002%, for the Infrastructure, 0.13% of jobs for Monitoring, 9.91% of jobs for Production, 56.30% of jobs for Other, and 33.63% of jobs for Free. 92.05% of jobs have a single task. 95.75% have fewer than 10 tasks. But 12 jobs have 5000 tasks and 114 jobs have around 1000 tasks.
- With regard to resource requests, 0.11% of jobs have a memory request and a CPU request higher than or equal to 10%.
- 94.25% of jobs wait less than 10 seconds before being scheduled. However, some of them wait for more than 1000 seconds. Such large values could be explained by the existence of placement constraints for the jobs, making them harder to place and schedule. 49% of jobs have an execution time less than 100 seconds.

Such results are needed to validate or invalidate some simplifying assumptions that are usually made when reasoning about models, and make the models more accurate for jobs and tasks as well as for available machines. Having validated these models on real data centers, they can then be used for extensive evaluation of placement and scheduling algorithms and more generally for resource allocation (i.e. CPU and memory). These algorithms can then be applied in real data centers.

Another possible use of this data set is to consider it as a learning set in order to predict some feature of the data center, such as the workload of hosts or the next arrival of jobs.

7.7.2. Machine Learning for an Energy-Efficient Management of Data Centers

Participants: Ruben Milocco (University Of Camahue, Argentina), 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. The problem of managing Data Centers (DC) and clouds optimally, in the sense that the demand is met with a minimal energy cost, remains a major issue. In this research, we evaluate the maximum energy saving that can be obtained in DCs by means of a proactive management of resources. The proposed management is based on models that predict resource requests.

Diverse approaches to obtain predictive models of DCs have been studied recently. Among the most popular methods with the comparatively lowest prediction errors are the predictive models of the ARMAX family. Hence, we study the predictive model given by the ARMAX family. We compare its performance with that of the Last Value (LV) model which predicts that the next value will be equal to the current one. To the best of our knowledge, there are no studies relating to the performance bounds that can be achieved using these models. In this research, we study the limits of the improvement in terms of energy cost that can be obtained using proactive strategies for DC management based on predictive models.

Using the Google dataset collected over a period of 29 days and made publicly available, we evaluate the largest benefit that can be obtained with those two predictors.

7.8. Protocols and Models for Wireless Networks - Application to VANETs

7.8.1. 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 [35]. 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. To predict the position of vehicles in this context, we adopt the machine learning methodology. As a pre-requisite, the vehicles know their positions and the vehicles send their positions in the packets. The positioning system can thus perform a training sequence and build a model. The system is then able to handle a prediction request. In this request, a vehicle without external positioning will request its position from the neighboring RSUs. The RSUs which receive this request message from the vehicle will know the power at which the message was received and will study the positioning request using the training set. In this study, we use and compare three widely recognized techniques : K Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest. We study these techniques in various configurations and discuss their respective advantages and drawbacks. Our results show that these three techniques provide very good results in terms of position predictions when the error on the transmission power is small.

7.8.2. Predicting transmission success with Machine-Learning and Support Vector Machine in VANETs

Participants: Samia Bouzefrane (Cnam), Soumya Banerjee (Birla Institute Of Technology, Mesra), Paul Mühlethaler, Mamoudou Sangare.

We study the use of the Support Vector Machine technique to estimate the probability of the reception of a given transmission in a Vehicular Ad hoc NETwork (VANET). The transmission takes place between a vehicle and a RoadSide Unit (RSU) at a given distance and with a given transmission rate. The RSU computes the statistics of the receptions and is able to compute the percentage of successful transmissions versus the distance between the vehicle and the RSU and the transmission rate. Starting from this statistic, a Support Vector Machine (SVM) scheme can produce a model. Then, given a transmission rate and a distance between the vehicle and the RSU, the SVM technique can estimate the probability of a succcessful reception. This probability can be used to build an adaptive technique which optimizes the expected throughput between the vehicle and the RSU. Instead of using transmission values of a real experiment, we use the results of an analytical model of CSMA that is customized for 1D VANETs. The model we adopt to perform this task uses a Matern selection process to mimic the transmission in a CSMA IEEE 802.11p VANET. With this model we obtain a closed formula for the probability of successful transmissions. Thus with these results we can train an SVM model and predict other values for other couples : distance, transmission rate. The numerical results we obtain show that SVM seems very suitable to predict the reception probability in a VANET.

7.8.3. TDMA scheduling strategies for vehicular ad hoc networks: from a distributed to a centralized approach

Participants: Mohammed Hadded, Anis Laouiti (Telecom Sud-Paris, Paul Mühlethaler.

We focus on vehicular safety applications based on the Dedicated Short Range Communication (DSRC) standard. We propose a new mechanism to alleviate channel congestion by reducing the beacons load while maintaining an accurate awareness level. Our scheme is based on the collective perception concept which consists in sharing perceived status information collected by vehicles equipped with different types of sensors (radars, lidars, cameras, etc.). To achieve our goal, we propose two main schemes [30]. The first one consists in implementing the collective perception capability on vehicles and adding a new category of status messages to share locally collected sensor data in order to reduce channels load and enhance vehicles' awareness. The second scheme concerns the accuracy level of the received information from the collective perception enabled vehicles by fixing a prior error threshold on the position. The method proposed is validated by simulations and

the results obtained are compared to those of an application based on the traditional beaconing scheme of the IEEE802.11p standard. The simulations show that the proposed scheme is able to significantly reduce the load on the control channel incurred by the beacons and the packet error ratio for different network densities and built-in sensors characteristics.

7.8.4. A Collaborative Environment Perception Approach for Vehicular Ad hoc Networks

Participants: Sadia Ingrachen, Nadjib Achir (Paris 13), Paul Mühlethaler, Tounsia Djamah (Paris 13), Amine Berqia (Paris 13).

We focus on vehicular safety applications based on the Dedicated Short Range Communication (DSRC) standard. We propose a new mechanism to alleviate channel congestion by reducing the beacons load while maintaining an accurate awareness level. Our scheme is based on the collective perception concept which consists in sharing perceived status information collected by vehicles equipped with different types of sensors (radars, lidars, cameras, etc.). To achieve our goal, we propose two main schemes [31]. The first one consists in implementing the collective perception capability on vehicles and adding a new category of status messages to share locally collected sensor data in order to reduce channels load and enhance vehicles' awareness. The second scheme concerns the accuracy level of the received information from the collective perception enabled vehicles by fixing a prior error threshold on the position. The method proposed is validated by simulations and the results obtained are compared to those of an application based on the traditional beaconing scheme of the IEEE802.11p standard. The simulations show that the proposed scheme is able to significantly reduce the load on the control channel incurred by the beacons and the packet error ratio for different network densities and built-in sensors characteristics.

FUN Project-Team

7. New Results

7.1. Performance Evaluation, Security, Safety and Verification

Participants: Antoine Gallais, Nathalie Mitton, Allan Blanchard.

7.1.1. Performance Evaluation and validation methodology

Envisioned communication densities in Internet of Things applications are increasing continuously. Because these wireless devices are often battery powered, we need specific energy efficient (low-power) solutions. Moreover, these smart objects use low-cost hardware with possibly weak links, leading to a lossy network. Once deployed, these low-power lossy networks (LLNs) are intended to collect the expected measurements, handle transient faults, topology changes, etc. Consequently, validation and verification during the protocol development are a matter of prime importance. A large range of theoretical or practical tools are available for performance evaluation. A theoretical analysis may demonstrate that the performance guarantees are respected, while simulations or experiments aim on estimating the behavior of a set of protocols within real-world scenarios. In [16], we review the various parameters that should be taken into account during such a performance evaluation. Our primary purpose is to provide a tutorial that specifies guidelines for conducting performance evaluation campaigns of network protocols in LLNs. We detail the general approach adopted in order to evaluate the performance of layer 2 and 3 protocols in LLNs. Furthermore, we also specify the methodology that should be adopted during the performance evaluation, while reviewing the numerous models and tools that are available to the research community.

7.1.2. Correlated failures

Current practices of fault-tolerant network design ignore the fact that most network infrastructure faults are localized or spatially correlated (i.e., confined to geo-graphic regions). Network operators require new tools to mitigate the impact of such region-based faults on their infrastructures. Utilizing the support from the U.S. Department of Defense, and by consolidating a wide range of theories and solutions developed in the last few years, [14] designs RAPTOR, an advanced Network Planning and Management Tool that facilitates the design and provisioning of robust and resilient networks.The tool provides multi-faceted network design, evaluation, and simulation capabilities for network planners. Future extensions of the tool currently being worked upon not only expand the tool's capabilities, but also extend these capabilities to heterogeneous interdependent networks such as communication, power, water, and satellite networks.

7.1.3. Contiki verification

Internet of Things (IoT) applications are becoming increasingly critical and require formal verification. Our recent work presented formal verification of the linked list module of Contiki, an OS for IoT. It relies on a parallel view of a linked list via a companion ghost array and uses an inductive predicate to link both views. In this work, a few interactively proved lemmas allow for the automatic verification of the list functions specifications, expressed in the acsl specification language and proved with the Frama-C/Wp tool. In a broader verification context, especially as long as the whole system is not yet formally verified, it would be very useful to use runtime verification , in particular, to test client modules that use the list module. It is not possible with the current specifications, which include an inductive predicate and axiomatically defined functions. In [27], an early-idea paper we show how to define a provably equivalent non-inductive predicate and a provably equivalent non-axiomatic function that belong to the executable subset e-acsl of acsl and can be transformed into executable C code. Finally, we propose an extension of Frama-C to handle both axiomatic specifications for deductive verification and executable specifications for runtime verification.

In [23], [47], we target Contiki, a widely used open-source OS for IoT, and present a verification case study of one of its most critical modules: that of linked lists. Its API and list representation differ from the classical linked list implementations, and are particularly challenging for deductive verification. The proposed verification technique relies on a parallel view of a list through a companion ghost array. This approach makes it possible to perform most proofs automatically using the Frama-C/WP tool, only a small number of auxiliary lemmas being proved interactively in the Coq proof assistant. We present an elegant segment-based reasoning over the companion array developed for the proof. Finally, we validate the proposed specification by proving a few functions manipulating lists.

With the wide expansion of multiprocessor architectures, the analysis and reasoning for programs under weak memory models has become an important concern. [13] presents MMFilter, an original constraint solver for generating program behaviors respecting a particular memory model. It is implemented in Prolog using CHR (Constraint Handling Rules). The CHR formalism provides a convenient generic solution for specifying memory models. It benefits from the existing optimized implementations of CHR and can be easily extended to new models. We present MMFilter design, illustrate the encoding of memory model constraints in CHR and discuss the benefits and limitations of the proposed technique.

7.2. Alternative communication paradigms

Participants: Antonio Costanzo, Valeria Loscri.

Nowadays, the always growing of connected objects and the strong demand to downsizing the devices in order to make the Internet of Things (IoT) paradigm more pervasive and ubiquitous, has motivated academic and industry people to investigate from one side mechanisms able to adapt quickly to the rapid external changes and to the quality of Services (QoS) parameters defined by the users and imposed by the adoption of new services and from another side, the investigation of portion of spectrum that have not been considered till this moment such as Terahertz band.

Nowadays, the always growing of connected objects and the strong demand to downsizing the devices in order to make the Internet of Things (IoT) paradigm more pervasive and ubiquitous, has motivated academic and industry people to investigate from one side mechanisms able to adapt quickly to the rapid external changes and to the quality of Services (QoS) parameters defined by the users and imposed by the adoption of new services and from another side, the investigation of portion of spectrum that have not been considered till this moment such as Terahertz band. In order to be able to realize a paradigm shift towards the Internet of Everything concept, a downsizing of devices is imposed allowing new applications as *in-vivo* diagnosis and monitoring. In order to be effective at this level it is imperative to analyze the new context, by highlighting the unique features to concretely realize the IoE paradigm. In this context, we have studied quantum particles called phonons, quasi-particles derived from vibrations of atoms in solids. Phonons have been envisaged as enabler of information transfer and their special characteristics have been exploited in [17]. Phonons have been also considered for a quantum channel in [26]. Another interesting approach for enabling the nano communication paradigm is represented by molecular communication. In particular, a main issue that is important to face is the coexistence between an artificial molecular communication and a biological system as explained in [40]. Alternative communication paradigms have attracted a lot of attention in the last a few years, not only by academic researchers but also by industry. Research on optical communication and in particular the possible exploitation of Visible Light communication with a twofold objective, to illuminate and to communicate has been object of an increasing interest. In this directions, we have proposed context-aware VLC systems in[39], [38] and [24]. The context is different in respect to the “traditional” wireless communication, since the external environment can change fastly and abruptly. Based on this primary observation, our main objective is to make theVLC system aware of the external noise and try to make it as robust as possible in respect of it.

7.3. Self-Organization

Participants: Antoine Gallais, Nathalie Mitton, Valeria Loscri, Farouk Mezghani, Anjalalaina Jean Cristanel Razafimandimby.

7.3.1. Stable parent selection

The Industrial Internet of Things consists in the use of low power lossy networks to enable next industrial applications. To work properly, the network has to provide strict guarantees concerning the delay and the reliability. IEEE 802.15.4-TSCH proposes time synchronized and slow channel hopping medium access control to cope with these requirements. It relies on a strict schedule of the transmissions, spread over orthogonal radio channels, to set up a resilient wireless infrastructure. A routing protocol (e.g. RPL) has then to construct energy-efficient routes on top of this link-layer topology (as investigated in the 6TiSCH IETF working group). Most of existing solutions rely on tree-based topologies, where each node has to select one or multiple parents to forward its traffic to the destination. Unfortunately, the links to the routing parents exhibit time-varying characteristics, due to e.g. obstacles, and external interference, thus leading to oscillations and increased required control of the routing topology. Moreover, the network has to provision enough resources (i.e., time, channel) to cope with those variations, while still being reactive to node/link failures. We investigated the stability of 6TiSCH networks, and especially the impact on routing parent selection. We identified moments of instability due to oscillations in the radio conditions caused by external interference and obstacles, in two indoor testbeds with different channel conditions. We identified the causes of instabilities, and proposed solutions for each of the layers in the 6TiSCH stack. First, at the MAC layer, we demonstrated that a rearrangement of shared cells in the slotframe reduces the probability of collisions for control packets, paving the way to a faster negotiation during topology reconfigurations. Next, we eased the schedule consistency management between two nodes (renegociated from scratch in the current standard, upon detection of a schedule inconsistency). Finally, at the routing layer, we exploited the existing correlation between the broadcast packet reception rate and the unicast link quality to create a two-step parent selection that favors stable parents. We finally obtained a network that converged faster and that reacted accurately during moments of instabilities. Results are available in [46], [42].

7.3.2. Bayesian communications

The amount of data that are generated in IoT devices is huge and the most of time data are highly correlated, by making useless the forwarding of all the raw data generated. Bearing that in mind, we have designated and implemented an effective mechanism to reduce the amount of data sent in the network in [45]. Results are encouraging since there is a size effect of less interfering in the communication system with an important impact on battery consumption for wireless devices that are energy constrained.

7.3.3. Multi-technology self-organization

Opportunistic communications present a promising solution for disaster network recovery in emergency situations such as hurricanes, earthquakes, and floods, where infrastructure might be destroyed. Some recent works in the literature have proposed opportunistic-based disaster recovery solutions, but they have omitted the consideration of mobile devices that come with different network technologies and various initial energy levels. [19], [30] present COPE, an energy-aware Cooperative OPportunistic alErt diffusion scheme for trapped survivors to use during disaster scenarios to report their position and ease their rescue operation. It aims to maintain mobile devices functional for as long as possible for maximum network coverage until reaching proximate rescuers. COPE deals with mobile devices that come with an assortment of networks and aims to perform systematic network interface selection. Furthermore, it considers mobile devices with various energy levels and allows low-energy nodes to hold their charge for longer time with the support of high-energy nodes. A proof-of-concept implementation has been performed to study the doability and efficiency of COPE, and to highlight the lessons learned. Following-up with these results, we performed several experimentations and could benchmark smartphone performances with regards to their multi-communications interfaces. Testing experiments have been carried out to measure the performance of smartphones in terms of energy consumption, clock synchronization and transmission range. We believe that such experimental results can support technological choices for rescue operations but also for many other applications relying on smartphone performances. Results are available in [30].

7.3.4. Heterogeneous Self-organizing (smart) Things

In the panorama of the Internet of Things, one main important issue is the management of heterogeneous objects, that need to communicate in order to exchange information and to interact in order to be able to synergically accomplish complex tasks and for providing services to final users. In this context, the thesis [10] has tried to face the main challenges related to complex heterogeneous systems, where objects are able to self-organize to each other and are equipped with some kind of intelligence in order to dynamically react to the environment changes. Several tools have been exploited ranging from artificial neural networks to genetic algorithms and different solutions have been proposed to make these systems dynamic and responding to the $*$self properties.

7.4. Smart Grids

Participants: Nathalie Mitton, Jad Nassar.

The Smart Grid (SG) aims to transform the current electric grid into a "smarter" network where the integration of renewable energy resources, energy efficiency and fault tolerance are the main benefits. This is done by interconnecting every energy source, storage point or central control point with connected devices, where heterogeneous SG applications and signaling messages will have different requirements in terms of reliability, latency and priority. Hence, data routing and prioritization are the main challenges in such networks.

So far, RPL (Routing Protocol for Low-Power and Lossy networks) protocol is widely used on Smart Grids for distributing commands over the grid. RPL assures traffic differentiation at the network layer in wireless sensor networks through the logical subdivision of the network in multiple instances, each one relying on a specific Objective Function. However, RPL is not optimized for Smart Grids, as its main objective functions and their associated metric does not allow Quality of Service differentiation.

In order to overcome this, we propose $OFQS$ an objective function [20] with a multi-objective metric that considers the delay and the remaining energy in the battery nodes alongside with the dynamic quality of the communication links. Our function automatically adapts to the number of instances (traffic classes) providing a Quality of Service differentiation based on the different Smart Grid applications requirements. We tested our approach on a real sensor testbed. The experimental results show that our proposal provides a lower packet delivery latency and a higher packet delivery ratio while extending the lifetime of the network compared to solutions in the literature.

The management of communication is an issue in WSN-based Smart Grid: billions of messages with different sizes and priorities are sent across the network. Data aggregation is a potential solution to reduce loads on the communication links, thus achieving a better utilization of the wireless channel and reducing energy consumption. On the other hand, SG applications require different Quality of Service (QoS) priorities. Delays caused by data aggregation must then be controlled in order to achieve a proper communication. In [33], [34], we propose a work in progress, that consists of a QoS efficient data aggregation algorithm with two aggregation functions for the different traffics in a SG network. We expect to reduce the energy consumption while respecting the data delivery delays for the different SG applications.

In order to reduce the amount of data sent over the network, and thus reduce energy consumption, data prediction is another potent solution of data reduction. It consists on predicting the values sensed by sensor nodes within certain error threshold, and resides both at the sensors and at the sink. The raw data is sent only if the desired accuracy is not satisfied, thereby reducing data transmission. We focus on time series estimation with Least Mean Square (LMS) for data prediction in WSN, in a Smart Grid context, where several applications with different data types and Quality of Service (QoS) requirements will exist on the same network. LMS proved its simplicity and robustness for a wide variety of applications, but the parameters selection (step size and filter length) can directly affect its global performance, choosing the right ones is then crucial. Having no clear and robust method on how to optimize these parameters for a variety of applications, we propose in [44] a modification of the original LMS that consists of training the filter for a certain time with the data itself in order to customize the aforementioned parameters. We consider different types of real data traces for the photo voltaic cells monitoring. Our simulation results provide a better data prediction while minimizing the mean square error compared to an existing solution in literature.

All these solutions have also been detailed in [12].

7.5. Connected Cars

Participants: Nathalie Mitton, Valeria Loscri, Joao Batista Pinto Neto.

7.5.1. Geolocalisation

Connected car technology promises to drastically reduce the number of accidents involving vehicles. Nevertheless, this technology requires the vehicle precise location to work. The adoption of Global Positioning System (GPS) as a navigation device imposes limitations to geolocation information under non-line-of-sight conditions. [22] introduces the Time Series Dead Reckoning System (TedriS) as a solution for dead reckoning navigation when the GPS fails. TedriS uses Time Series Regression Models (TSRM) and the data from the rear wheel speed sensor of the vehicle to estimate the absolute position. The process to estimate the position is carried out in two phases: training and predicting. In the training phase, a novel technique applies TSRM and stores the relationship between the GPS and the rear wheel speed data; then in the predicting phase, this relationship is used. We analyze TedriS using traces collected at the campus of Federal University of Rio de Janeiro (UFRJ), Brazil, and with indoor experiments with a robot. Results show an accuracy compatible with dead-reckoning navigation state-of-art systems.

7.5.2. Data forwarding

Intelligent inter-vehicle communication is a key research field in the context of vehicular networks that applies in real-life applications (e.g., management of accidents, intelligent fuel consumption, smart traffic jams, etc.). Considering different roles of nodes based on their "social aptitude" to relay information could provide a social component in the vehicular structure that can be useful in getting a clear prediction of the topological evolution in time and space proving to be very effective in managing intelligent data forwarding. In [36], we characterize a vehicular network as a graph using the link layer connectivity level and we classify nodes on the basis of specific attributes characterizing their "social aptitude" to forward data. Two forwarding approaches are presented, based on different socialites that allow to (i) select the most social node (i.e., a social hub) or (ii) choose among various social nodes.

7.5.3. Internet of vehicles

Internet, in its most recent evolution, is going to be the playground where a multitude of heterogeneous interconnected "things" autonomously exchange information to accomplish some tasks or to provide a service. Recently, the idea of giving to those smart devices the capability to organize themselves according to a social structure, gave birth to the so-called paradigm of the Social Internet of Things. The expected benefits of SIoT range from the enhanced effectiveness, scalability and speed of the navigability of the network of interconnected objects, to the provision of a level of trustworthiness that can be established by averaging the social relationships among things that are "friends". Bearing in mind the beneficial effects of social components in IoT, we consider a social structure in a vehicular context i.e., Social Internet of Vehicles (SIoV). In SIoV, smart vehicles build social relationships with other social objects they might come into contact, with the intent of creating an overlay social network to be exploited for information search and dissemination for vehicular applications. In [43], we aim to investigate the social behavior of vehicles in SIoV and how it is affected by mobility patterns. Specifically, through the analysis of simulated traffic traces, we distinguish friendly and acquaintance vehicles based on the encounter time and connection maintenance.

7.6. Robots and drones

Participants: Nathalie Mitton, Valeria Loscri, Farouk Mezghani, Anjalalaina Jean Cristanel Razafimandimby.

Internet of Robotic Things (IoRT) is a new concept introduced for the first time by ABI Research. Unlike the Internet of Things (IoT), IoRT provides an active sensorization and is considered as the new evolution of IoT. In this context, we propose a Neuro-Dominating Set algorithm (NDS) [21] to efficiently deploy a team of mobile wireless robots in an IoRT scenario, in order to reach a desired inter-robot distance, while maintaining global connectivity in the whole network. We use the term Neuro-Dominating Set to describe our approach, since it is inspired by both neural network and dominating set principles. With NDS algorithm, a robot adopts different behaviors according whether it is a dominating or a dominated robot. Our main goal is to show and demonstrate the beneficial effect of using different behaviors in the IoRT concept. The obtained results show that the proposed method outperforms an existing related technique (i.e., the Virtual Angular Force approach) and the neural network based approach presented in our previous work. As an objective, we aim to decrease the overall traveled distance and keep a low energy consumption level, while maintaining network connectivity and an acceptable convergence time.

Routing a fleet of robots in a known surface is a complex problem. It consists in the determination of the exact trajectory each robot has to follow to collect information. This is what we propose in [32] with the objective is to maximize the exploration of the given surface. To ensure that the robots can execute the mission in a collaborative manner, connectivity constraints are considered. These constraints guarantee that robots can communicate among each other and share the collected information. Moreover, the trajectories of the robots need to respect autonomy constraints.

When a disaster strikes, the telecommunications infrastructure gets damaged making rescue operations more challenging. Connecting first responders through flying base stations (i.e. drone mounted LTE (Long-Term Evolution) femtocell base station) presents a promising alternative to support infrastructure failure during disasters. The drone can travel the area and communicate with ground mobile devices, such as smartphones, and serves as flying data link to share information between survivors and rescuers. Problem statement. We would like to submit the following open problem to the community. Given the position of the ground mobile devices to serve, the problem presented here is about the dynamic drone path planning. As the drone autonomy is very limited and due to the high cost of drone mounted base station, the goal of this problem is to determine the best energy-efficient and minimum-time path to travel the area as fast as possible while still remaining in range of each survivor long enough to assure full servicing. This is the problem stated in [31].

7.7. MAC mechanisms

Participant: Nathalie Mitton.

In the era of the Internet of Things (IoT), the number of connected devices is growing dramatically. Often, connected objects use Industrial, Scientific and Medical (ISM) radio bands for communication. These kinds of bands are available without license, which facilitates development and implementation of new connected objects. However, it also leads to an increased level of interference in these bands. Interferences not only negatively affect the Quality of Service, but also cause energy losses, which is especially unfavorable for the energy constrained Wireless Sensor Networks (WSN). In [25], we develop an explicit formula of outage probability in a distributed wireless sensor network (WSN), assuming the MAC layer protocol being a slotted-ALOHA. And adopting a Markovian approach, we develop a model that analyses the performance of the slotted-ALOHA in order to improve these performances, in particular, by adding a preliminary stage of channel reservation, we show that this modification is important to have a high performance distributed wireless sensor network.

Several wild animal species are endangered by poaching. As a solution, deploying wireless sensors on animals able to send regular messages and also alert messages has been envisaged recently by several authorities and foundations. In that context, we have proposed WildMAC [35], a multichannel, multihop wireless communication protocol for these specific wireless sensor networks that have to collect data from unknown large areas with different QoS requirements. WildMAC is a TDMA based MAC protocol that leverages long range communication properties to propose an efficient data collection mean. Its performance evaluation shows it meets QoS requirements. To size the different parameters of WildMAC, we relied on the results of the study of [25].

7.8. RFID

Participants: Nathalie Mitton, Abdoul Aziz Mbacke, Ibrahim Amadou.

While RFID technology is gaining increased attention from industrial community deploying different RFID-based applications, it still suffers from reading collisions. As such, many proposals were made by the scientific community to try and alleviate that issue using different techniques either centralized or distributed, monochannel or multichannels, TDMA or CSMA. However, the wide range of solutions and their diversity make it hard to have a clear and fair overview of the different works. [18] surveys the most relevant and recent known state-of-the-art anti-collision for RFID protocols. It provides a classification and performance evaluation taking into consideration different criteria as well as a guide to choose the best protocol for given applications depending on their constraints or requirements but also in regard to their deployment environments.

Among all these approaches, [29], [28] propose new reader anti-collision schemes and data-priority aware data collection in a multi-hop RFID data collection protocol. [28] examines the implementation of two applications: for industrial IoT and for smart cities, respectively. Both applications, in regards to their requirements and configuration, challenge the operation of a RFID sensing solution combined with a dynamic wireless data gathering over multihops. They require the use of both mobile and fixed readers to cover the extent of deployment area and a quick retrieval of tag information. We propose a distributed crosslayer solution for improving the efficiency of the RFID system in terms of collision and throughput but also its proficiency in terms of tag information routing towards one or multiple sinks. Simulation results show that we can achieve high level of throughput while maintaining a low level of collision and a fairness of reader medium access above 95% in situations where readers can be fix and mobile, while tag information is routed with a data rate of 97% at worst and reliable delays for considered applications. [29] proposes cross-layer solutions meant for both scheduling of readers' activity to avoid collisions, and a multihop routing towards base stations, to gather read tag data. This routing is performed with a data priority aware mechanism allowing end-to-end delay reduction of urgent data packets delivery up to 13% faster compared to standard ones. Using fuzzy logic, we combine several observed metrics to reduce the load of forwarding nodes and improve latency as well as data rate. We validate our proposal running simulations on industrial and urban scenarios.

All these solutions have also been detailed in [11].

7.9. Smart Cities

Smart cities are a key factor in the consumption of materials and resources. As populations grow and resources become scarcer, the efficient usage of these limited goods becomes more important. Building on and integrating with a huge amount of data, the cities of the future are becoming a realization today. There are millions of sensors in place already, monitoring various things in metropolises. In the near future, these sensors will multiply until they can monitor everything from streetlights and trashcans to road conditions and energy consumption. In this context, effective strategies or solutions for refining data sets can play a key role. In [37], we propose a scheme in which passive RFID is shown as an interesting alternative and complement to WSN to alleviate the cost of some Smart City applications.

Also, in Smart Cities, crowd sensing may help to identify the current speed for each street, the congested areas, etc. In this context, map matching techniques are required to map a sequence of GPS waypoints into a set of streets on a common map. Unfortunately, most map matching approaches are probabilistic. In [41], we propose rather an unambiguous algorithm, able to identify all the possible paths that match a given sequence of waypoints. We need an unambiguous identification for each waypoints set. For instance, the actual speed should be assigned to the correct set of streets, without error. To identify all the possible streets, we construct the set of candidates iteratively. We identify all the edge candidates around each waypoint, and reconstruct all the possible sub-routes that connect them. We then verify a set of constraints, to eliminate impossible routes. The road segments common to all computed routes form an unambiguous match. We evaluate the matching ratio of our technique on real city maps (London, Paris and Luxembourg). We also validate our approach with a real GPS trace in Seattle.

In parallel, we proposed a MOOC in the framework of the IPL CityLab project (See Section 9.2.1), whose working documents are available online [48].

GANG Project-Team

7. New Results

7.1. Graph and Combinatorial Algorithms

7.1.1. Random Walks with Multiple Step Lengths

In nature, search processes that use randomly oriented steps of different lengths have been observed at both the microscopic and the macroscopic scales. Physicists have analyzed in depth two such processes on grid topologies: *Intermittent Search*, which uses two step lengths, and *Lévy Walk*, which uses many. Taking a computational perspective, in [26] we consider the number of distinct step lengths k as a *complexity measure* of the considered process. Our goal is to understand what is the optimal achievable time needed to cover the whole terrain, for any given value of k. Attention is restricted to dimension one, since on higher dimensions, the simple random walk already displays a quasi linear cover time.

We say X is a *k-intermittent search* on the one dimensional n-node cycle if there exists a probability distribution $\mathbf{p} = (p_i)_{i=1}^{k}$, and integers $L_1, L_2, ..., L_k$, such that on each step X makes a jump $\pm L_i$ with probability p_i, where the direction of the jump ($+$ or $-$) is chosen independently with probability $1/2$. When performing a jump of length L_i, the process consumes time L_i, and is only considered to visit the last point reached by the jump (and not any other intermediate nodes). This assumption is consistent with biological evidence, in which entities do not search while moving ballistically. We provide upper and lower bounds for the cover time achievable by k-intermittent searches for any integer k. In particular, we prove that in order to reduce the cover time $\Theta(n^2)$ of a simple random walk to $\widetilde{\Theta}(n)$, roughly $\frac{\log n}{\log \log n}$ step lengths are both necessary and sufficient, and we provide an example where the lengths form an exponential sequence.

In addition, inspired by the notion of intermittent search, we introduce the *Walk or Probe* problem, which can be defined with respect to arbitrary graphs. Here, it is assumed that querying (probing) a node takes significantly more time than moving to a random neighbor. Hence, to efficiently probe all nodes, the goal is to balance the time spent walking randomly and the time spent probing. We provide preliminary results for connected graphs and regular graphs.

7.1.2. Searching a Tree with Permanently Noisy Advice

In [16], we consider a search problem on trees using unreliable guiding instructions. Specifically, an agent starts a search at the root of a tree aiming to find a treasure hidden at one of the nodes by an adversary. Each visited node holds information, called *advice*, regarding the most promising neighbor to continue the search. However, the memory holding this information may be unreliable. Modeling this scenario, we focus on a probabilistic setting. That is, the advice at a node is a pointer to one of its neighbors. With probability q each node is *faulty*, independently of other nodes, in which case its advice points at an arbitrary neighbor, chosen uniformly at random. Otherwise, the node is *sound* and points at the correct neighbor. Crucially, the advice is *permanent*, in the sense that querying a node several times would yield the same answer. We evaluate efficiency by two measures: The *move complexity* denotes the expected number of edge traversals, and the *query complexity* denotes the expected number of queries.

Let Δ denote the maximal degree. Roughly speaking, the main message of this paper is that a phase transition occurs when the *noise parameterq* is roughly $1/\sqrt{\Delta}$. More precisely, we prove that above the threshold, every search algorithm has query complexity (and move complexity) which is both exponential in the depth d of the treasure and polynomial in the number of nodes n. Conversely, below the threshold, there exists an algorithm with move complexity $O(d\sqrt{\Delta})$, and an algorithm with query complexity $O(\sqrt{\Delta}\log\Delta\log^2 n)$. Moreover, for the case of regular trees, we obtain an algorithm with query complexity $O(\sqrt{\Delta}\log n\log\log n)$. For q that is below but close to the threshold, the bound for the move complexity is tight, and the bounds for the query complexity are not far from the lower bound of $\Omega(\sqrt{\Delta}\log_\Delta n)$.

In addition, we also consider a *semi-adversarial* variant, in which faulty nodes are still chosen at random, but an adversary chooses (beforehand) the advice of such nodes. For this variant, the threshold for efficient moving algorithms happens when the noise parameter is roughly $1/\Delta$. In fact, above this threshold a simple protocol that follows each advice with a fixed probability already achieves optimal move complexity.

7.1.3. Patterns on 3 vertices

In [31] we deal with graph classes characterization and recognition. A popular way to characterize a graph class is to list a minimal set of forbidden induced subgraphs. Unfortunately this strategy usually does not lead to an efficient recognition algorithm. On the other hand, many graph classes can be efficiently recognized by techniques based on some interesting orderings of the nodes, such as the ones given by traversals.

We study specifically graph classes that have an ordering avoiding some ordered structures. More precisely, we consider what we call *patterns on three nodes*, and the recognition complexity of the associated classes. In this domain, there are two key previous works. Damashke started the study of the classes defined by forbidden patterns, a set that contains interval, chordal and bipartite graphs among others. On the algorithmic side, Hell, Mohar and Rafiey proved that any class defined by a set of forbidden patterns can be recognized in polynomial time. We improve on these two works, by characterizing systematically all the classes defined sets of forbidden patterns (on three nodes), and proving that among the 23 different classes (up to complementation) that we find, 21 can actually be recognized in linear time.

Beyond this result, we consider that this type of characterization is very useful, leads to a rich structure of classes, and generates a lot of open questions worth investigating.

7.1.4. The Dependent Doors Problem: An Investigation into Sequential Decisions without Feedback

In [13], we introduce the *dependent doors problem* as an abstraction for situations in which one must perform a sequence of dependent decisions, without receiving feedback information on the effectiveness of previously made actions. Informally, the problem considers a set of d doors that are initially closed, and the aim is to open all of them as fast as possible. To open a door, the algorithm knocks on it and it might open or not according to some probability distribution. This distribution may depend on which other doors are currently open, as well as on which other doors were open during each of the previous knocks on that door. The algorithm aims to minimize the expected time until all doors open. Crucially, it must act at any time without knowing whether or which other doors have already opened. In this work, we focus on scenarios where dependencies between doors are both positively correlated and acyclic.

The fundamental distribution of a door describes the probability it opens in the best of conditions (with respect to other doors being open or closed). We show that if in two configurations of d doors corresponding doors share the same fundamental distribution, then these configurations have the same optimal running time up to a universal constant, no matter what are the dependencies between doors and what are the distributions. We also identify algorithms that are optimal up to a universal constant factor. For the case in which all doors share the same fundamental distribution we additionally provide a simpler algorithm, and a formula to calculate its running time. We furthermore analyse the price of lacking feedback for several configurations governed by standard fundamental distributions. In particular, we show that the price is logarithmic in d for memoryless doors, but can potentially grow to be linear in d for other distributions.

We then turn our attention to investigate precise bounds. Even for the case of two doors, identifying the optimal sequence is an intriguing combinatorial question. Here, we study the case of two cascading memoryless doors. That is, the first door opens on each knock independently with probability p_1. The second door can only open if the first door is open, in which case it will open on each knock independently with probability p_2. We solve this problem almost completely by identifying algorithms that are optimal up to an additive term of 1.

7.1.5. Finding maximum cliques in disk and unit ball graphs

In an *intersection graph*, the vertices are geometric objects with an edge between any pair of intersecting objects. Intersection graphs have been studied for many different families of objects due to their practical

applications and their rich structural properties. Among the most studied ones are *disk graphs*, which are intersection graphs of closed disks in the plane, and their special case, *unit disk graphs*, where all the radii are equal. Their applications range from sensor networks to map labeling, and many standard optimization problems have been studied on disk graphs. Most of the hard optimization and decision problems remain NP-hard on disk graphs and even unit disk graphs. For instance, disk graphs contain planar graphs on which several of those problems are intractable.

The complexity of MAXIMUM CLIQUE on general disk graphs is a notorious open question in computational geometry. On the one hand, no polynomial-time algorithm is known, even when the geometric representation is given. On the other hand, the NP-hardness of the problem has not been established, even when only the graph is given as input.

Recently, Bonnet *et al.* showed that the disjoint union of two odd cycles is not the complement of a disk graph. From this result, they obtained a subexponential algorithm running in time $2^{\widetilde{O}(n^{2/3})}$ for MAXIMUM CLIQUE on disk graphs, based on a win-win approach. They also got a QPTAS by calling a PTAS for MAXIMUM INDEPENDENT SET on graphs with sublinear odd cycle packing number due to Bock *et al.*, or branching on a low-degree vertex.

In [17], our main contributions are twofold. The first is a randomized EPTAS (Efficient Polynomial-Time Approximation Scheme, that is, a PTAS in time $f(\varepsilon)n^{O(1)}$) for MAXIMUM INDEPENDENT SET on graphs of $\mathcal{X}(d, \beta, 1)$. The class $\mathcal{X}(d, \beta, 1)$ denotes the class of graphs whose neighborhood hypergraph has VC-dimension at most d, independence number at least βn, and no disjoint union of two odd cycles as an induced subgraph. Using the forbidden induced subgraph result of Bonnet *et al.*, it is then easy to reduce MAXIMUM CLIQUE on disk graphs to MAXIMUM INDEPENDENT SET on $\mathcal{X}(4, \beta, 1)$ for some constant β. We therefore obtain a randomized EPTAS (and a PTAS) for MAXIMUM CLIQUE on disk graphs, settling almost [0] completely the approximability of this problem.

The second contribution is to show the same forbidden induced subgraph for unit ball graphs as the one obtained for disk graphs : their complement cannot have a disjoint union of two odd cycles as an induced subgraph. The proofs are radically different and the classes are incomparable. So the fact that the same obstruction applies for disk graphs and unit ball graphs might be somewhat accidental. And again we therefore obtain a randomized EPTAS in time $2^{\widetilde{O}(1/\varepsilon^3)}n^{O(1)}$ for MAXIMUM CLIQUE on unit ball graphs, even without the geometric representation.

Before that result, the best approximation factor was 2.553, due to Afshani and Chan. In particular, even getting a 2-approximation algorithm (as for disk graphs) was open.

Finally we show that such an approximation scheme, even in subexponential time, is unlikely for ball graphs (that is, 3-dimensional disk graphs with arbitrary radii), and unit 4-dimensional disk graphs. Our lower bounds also imply NP-hardness. To the best of our knowledge, the NP-hardness of MAXIMUM CLIQUE on unit d-dimensional disk graphs was only known when d is superconstant ($d = \Omega(\log n)$).

7.1.6. δ-hyperbolicity

In [19], we show that the eccentricities (and thus the centrality indices) 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 $c\delta$, i.e., after a linear time preprocessing, for every vertex v of G one can compute in $O(1)$ time an estimate $\widehat{e}(v)$ of its eccentricity $ecc_G(v)$ such that $ecc_G(v) \leq \widehat{e}(v) \leq ecc_G(v) + c\delta$ for a small constant c. We prove that every δ-hyperbolic graph G has a shortest path tree, constructible in linear time, such that for every vertex v of G, $ecc_G(v) \leq ecc_T(v) \leq ecc_G(v) + c\delta$. 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 $c'\delta$ 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

[0] The NP-hardness, ruling out a 1-approximation, is still to show.

for approximating centrality 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.

7.1.7. Graph searches and geometric convexities in graphs

In an attempt to understand graph searching on cocomparability graphs has been so successful, one quickly notices that the orderings produced by these traversals are precisely words of some antimatroids or convex geometries. The notion of antimatroids and convex geometries have appeared in the literature under various settings; in this work, we focus on the graph searching setting, where we discuss some known geometries on cocomparability graphs, and then present new structural properties on AT-free graphs in the hope of exploring whether the algorithms on cocomparability graphs can be lifted to this larger graph class. A first version of this work in collaboration with Feodor Dragan and Lalla Mouatadib was presented at ICGT Lyon, july 2018.

7.2. Distributed Computing

7.2.1. On the Limits of Noise in Distributed Computing

Biological systems can share and collectively process information to yield emergent effects, despite inherent noise in communication. While man-made systems often employ intricate structural solutions to overcome noise, the structure of many biological systems is more amorphous. It is not well understood how communication noise may affect the computational repertoire of such groups. To approach this question we consider in [9], [15] the basic collective task of rumor spreading, in which information from few knowledgeable sources must reliably flow into the rest of the population. We study the effect of communication noise on the ability of groups that lack stable structures to efficiently solve this task. We present an impossibility result which strongly restricts reliable rumor spreading in such groups. Namely, we prove that, in the presence of even moderate levels of noise that affect all facets of the communication, no scheme can significantly outperform the trivial one in which agents have to wait until directly interacting with the sources—a process which requires linear time in the population size. Our results imply that in order to achieve efficient rumor spread a system must exhibit either some degree of structural stability or, alternatively, some facet of the communication which is immune to noise. We then corroborate this claim by providing new analyses of experimental data regarding recruitment in Cataglyphis niger desert ants. Finally, in light of our theoretical results, we discuss strategies to overcome noise in other biological systems.

7.2.2. Minimizing message size in stochastic communication patterns: fast self-stabilizing protocols with 3 bits

In [8], we consider the basic PULL model of communication, in which in each round, each agent extracts information from few randomly chosen agents. We seek to identify the smallest amount of information revealed in each interaction (message size) that nevertheless allows for efficient and robust computations of fundamental information dissemination tasks. We focus on the *Majority Bit Dissemination* problem that considers a population of n agents, with a designated subset of *source agents*. Each source agent holds an *input bit* and each agent holds an *output bit*. The goal is to let all agents converge their output bits on the most frequent input bit of the sources (the *majority bit*). Note that the particular case of a single source agent corresponds to the classical problem of *Broadcast* (also termed *Rumor Spreading*). We concentrate on the severe fault-tolerant context of *self-stabilization*, in which a correct configuration must be reached eventually, despite all agents starting the execution with arbitrary initial states. In particular, the specification of who is a source and what is its initial input bit may be set by an adversary.

We first design a general compiler which can essentially transform any self-stabilizing algorithm with a certain property (called "the *the bitwise-independence property*") that uses ℓ-bits messages to one that uses only $\log \ell$-bits messages, while paying only a small penalty in the running time. By applying this compiler recursively we then obtain a self-stabilizing *Clock Synchronization* protocol, in which agents synchronize their clocks modulo some given integer T, within $\widetilde{O}(\log n \log T)$ rounds w.h.p., and using messages that contain 3 bits only. We then employ the new Clock Synchronization tool to obtain a self-stabilizing Majority Bit Dissemination protocol which converges in $\widetilde{O}(\log n)$ time, w.h.p., on every initial configuration, provided that the ratio of

sources supporting the minority opinion is bounded away from half. Moreover, this protocol also uses only 3 bits per interaction.

7.2.3. Intense Competition can Drive Selfish Explorers to Optimize Coverage

In [30], we consider a game-theoretic setting in which selfish individuals compete over resources of varying quality. The motivating example is a group of animals that disperse over patches of food of different abundances. In such scenarios, individuals are biased towards selecting the higher quality patches, while, at the same time, aiming to avoid costly collisions or overlaps. Our goal is to investigate the impact of collision costs on the parallel coverage of resources by the whole group.

Consider M *sites*, where a site x has *value* $f(x)$. We think of $f(x)$ as the reward associated with site x, and assume that if a single individual visits x exclusively, it receives this exact reward. Typically, we assume that if $\ell > 1$ individuals visit x then each receives at most $f(x)/\ell$. In particular, when competition costs are high, each individual might receive an amount strictly less than $f(x)/\ell$, which could even be negative. Conversely, modeling cooperation at a site, we also consider cases where each one gets more than $f(x)/\ell$. There are k identical *players* that compete over the rewards. They independently act in parallel, in a one-shot scenario, each specifying a single site to visit, without knowing which sites are explored by others. The group performance is evaluated by the expected *coverage*, defined as the sum of $f(x)$ over all sites that are explored by at least one player. Since we assume that players cannot coordinate before choosing their site we focus on symmetric strategies.

The main takeaway message of this paper is that the optimal symmetric coverage is expected to emerge when collision costs are relatively high, so that the following "Judgment of Solomon" type of rule holds: If a single player explores a site x then it gains its full reward $f(x)$, but if several players explore it, then neither one receives any reward. Under this policy, it turns out that there exists a unique symmetric Nash Equilibrium strategy, which is, in fact, evolutionary stable. Moreover, this strategy yields the best possible coverage among all symmetric strategies. Viewing the coverage measure as the social welfare, this policy thus enjoys a *(Symmetric) Price of Anarchy* of precisely 1, whereas, in fact, any other congestion policy has a price strictly greater than 1.

Our model falls within the scope of mechanism design, and more precisely in the area of incentivizing exploration. It finds relevance in evolutionary ecology, and further connects to studies on Bayesian parallel search algorithms.

7.2.4. Universal Protocols for Information Dissemination Using Emergent Signals

In [23], we consider a population of n agents which communicate with each other in a decentralized manner, through random pairwise interactions. One or more agents in the population may act as authoritative sources of information, and the objective of the remaining agents is to obtain information from or about these source agents. We study two basic tasks: broadcasting, in which the agents are to learn the bit-state of an authoritative source which is present in the population, and source detection, in which the agents are required to decide if at least one source agent is present in the population or not.

We focus on designing protocols which meet two natural conditions: (1) universality, i.e., independence of population size, and (2) rapid convergence to a correct global state after a reconfiguration, such as a change in the state of a source agent. Our main positive result is to show that both of these constraints can be met. For both the broadcasting problem and the source detection problem, we obtain solutions with a convergence time of $O(\log^2 n)$ rounds, w.h.p., from any starting configuration. The solution to broadcasting is exact, which means that all agents reach the state broadcast by the source, while the solution to source detection admits one-sided error on a ε-fraction of the population (which is unavoidable for this problem). Both protocols are easy to implement in practice and have a compact formulation.

Our protocols exploit the properties of self-organizing oscillatory dynamics. On the hardness side, our main structural insight is to prove that any protocol which meets the constraints of universality and of rapid convergence after reconfiguration must display a form of non-stationary behavior (of which oscillatory dynamics are an example). We also observe that the periodicity of the oscillatory behavior of the protocol,

when present, must necessarily depend on the number $^{\#}X$ of source agents present in the population. For instance, our protocols inherently rely on the emergence of a signal passing through the population, whose period is $\Theta(\log \frac{n}{\#X})$ rounds for most starting configurations. The design of clocks with tunable frequency may be of independent interest, notably in modeling biological networks.

7.2.5. Ergodic Effects in Token Circulation

In [25], we consider a dynamical process in a network which distributes all particles (tokens) located at a node among its neighbors, in a round-robin manner.

We show that in the recurrent state of this dynamics (i.e., disregarding a polynomially long initialization phase of the system), the number of particles located on a given edge, averaged over an interval of time, is tightly concentrated around the average particle density in the system. Formally, for a system of k particles in a graph of m edges, during any interval of length T, this time-averaged value is $k/m \pm \widetilde{O}(1/T)$, whenever $gcd(m,k) = \widetilde{O}(1)$ (and so, e.g., whenever m is a prime number). To achieve these bounds, we link the behavior of the studied dynamics to ergodic properties of traversals based on Eulerian circuits on a symmetric directed graph. These results are proved through sum set methods and are likely to be of independent interest.

As a corollary, we also obtain bounds on the *idleness* of the studied dynamics, i.e., on the longest possible time between two consecutive appearances of a token on an edge, taken over all edges. Designing trajectories for k tokens in a way which minimizes idleness is fundamental to the study of the patrolling problem in networks. Our results immediately imply a bound of $\widetilde{O}(m/k)$ on the idleness of the studied process, showing that it is a distributed $\widetilde{O}(1)$-competitive solution to the patrolling task, for all of the covered cases. Our work also provides some further insights that may be interesting in load-balancing applications.

7.2.6. Improved Analysis of Deterministic Load-Balancing Schemes

In [7], we consider the problem of deterministic load balancing of tokens in the discrete model. A set of n processors is connected into a d-regular undirected network. In every time step, each processor exchanges some of its tokens with each of its neighbors in the network. The goal is to minimize the discrepancy between the number of tokens on the most-loaded and the least-loaded processor as quickly as possible.

Rabani et al. (1998) present a general technique for the analysis of a wide class of discrete load balancing algorithms. Their approach is to characterize the deviation between the actual loads of a discrete balancing algorithm with the distribution generated by a related Markov chain. The Markov chain can also be regarded as the underlying model of a continuous diffusion algorithm. Rabani et al. showed that after time $T = O(\log(Kn)/\mu)$, any algorithm of their class achieves a discrepancy of $O(d \log n/\mu)$, where μ is the spectral gap of the transition matrix of the graph, and K is the initial load discrepancy in the system.

In this work we identify some natural additional conditions on deterministic balancing algorithms, resulting in a class of algorithms reaching a smaller discrepancy. This class contains well-known algorithms, eg., the Rotor-Router. Specifically, we introduce the notion of cumulatively fair load-balancing algorithms where in any interval of consecutive time steps, the total number of tokens sent out over an edge by a node is the same (up to constants) for all adjacent edges. We prove that algorithms which are cumulatively fair and where every node retains a sufficient part of its load in each step, achieve a discrepancy of $O(\min\{d\sqrt{\log n/\mu}, d\sqrt{n}\})$ in time $O(T)$. We also show that in general neither of these assumptions may be omitted without increasing discrepancy. We then show by a combinatorial potential reduction argument that any cumulatively fair scheme satisfying some additional assumptions achieves a discrepancy of $O(d)$ almost as quickly as the continuous diffusion process. This positive result applies to some of the simplest and most natural discrete load balancing schemes.

7.2.7. The assignment problem

In the allocation problem, asynchronous processors must partition a set of items so that each processor leave knowing all items exclusively allocated to it. In [21], we introduce a new variant of the allocation problem called the assignment problem, in which processors might leave having only partial knowledge of their assigned items. The missing items in a processor's assignment must eventually be announced by other processors.

While allocation has consensus power 2, we show that the assignment problem is solvable read-write wait-free when k processors compete for at least $2k-1$ items. Moreover, we propose a long-lived read-write wait-free assignment algorithm which is fair, allocating no more than 2 items per processor, and in which a slow processor may delay the assignment of at most n items, where n is the number of processors.

The assignment problem and its read-write solution may be of practical interest for implementing resource allocators and work queues, which are pervasive concurrent programming patterns, as well as stream-processing systems.

7.2.8. A Characterization of t-Resilient Colorless Task Anonymous Solvability

One of the central questions in distributed computability is characterizing the tasks that are solvable in a given system model. In the anonymous case, where processes have no identifiers and communicate through multi-writer/multi-reader registers, there is a recent topological characterization (Yanagisawa 2017) of the colorless tasks that are solvable when any number of asynchronous processes may crash. In [22], we consider the case where at most t asynchronous processes may crash, where $1 \leq t < n$. We prove that a colorless task is t-resilient solvable anonymously if and only if it is t-resilient solvable non-anonymously. We obtain our results through various reductions and simulations that explore how to extend techniques for non-anonymous computation to anonymous one.

7.2.9. Implementing Snapshot Objects on Top of Crash-Prone Asynchronous Message-Passing Systems

In asynchronous crash-prone read/write shared-memory systems there is the notion of a snapshot object, which simulates the behavior of an array of single-writer/multi-reader (SWMR) shared registers that can be read atomically. Processes in the system can access the object invoking (any number of times) two operations, denoted write() and snapshot(). A process invokes write() to update the value of its register in the array. When it invokes snapshot(), the process obtains the values of all registers, as if it read them simultaneously. It is known that a snapshot object can be implemented on top of SWMR registers, tolerating any number of process failures. Snapshot objects provide a level of abstraction higher than individual SWMR registers, and they simplify the design of applications. Building a snapshot object on an asynchronous crash-prone message-passing system has similar benefits. The object can be implemented by using the known simulations of a SWMR shared memory on top of an asynchronous message-passing system (if less than half the processes can crash), and then build a snapshot object on top of the simulated SWMR memory. [10] presents an algorithm that implements a snapshot object directly on top of the message-passing system, without building an intermediate layer of a SWMR shared memory. To the authors knowledge, the proposed algorithm is the first providing such a direct construction. The algorithm is more efficient than the indirect solution, yet relatively simple.

7.2.10. Distributed decision

We have carried out our study of distributed decision, either for its potential application to the design of fault-tolerant distributed algorithm, or for the purpose of designing a complexity/computanility theory for distributed network computing.

In the framework of *distributed network computing*, it is known that not all Turing-decidable predicates on labeled networks can be decided *locally* whenever the computing entities are Turing machines (TM), and this holds even if nodes are running *non-deterministic* Turing machines (NTM). In contrast, we show in [6] that every Turing-decidable predicate on labeled networks can be decided locally if nodes are running *alternating* Turing machines (ATM). More specifically, we show that, for every such predicate, there is a local algorithm for ATMs, with at most two alternations, that decides whether the actual labeled network satisfies that predicate. To this aim, we define a hierarchy of classes of decision tasks, where the lowest level contains tasks solvable with TMs, the first level those solvable with NTMs, and the level $k > 1$ contains those tasks solvable with ATMs with $k-1$ alternations. We characterize the entire hierarchy, and show that it collapses in the second level. In addition, we show separation results between the classes of network predicates that are locally decidable with TMs, NTMs, and ATMs, and we establish the existence of completeness results for

each of these classes, using novel notions of *local reduction*. We complete these results by a study of the local decision hierarchy when certificates are bounded to be of logarithmic size.

Distributed proofs are mechanisms enabling the nodes of a network to collectively and efficiently check the correctness of Boolean predicates on the structure of the network (e.g. having a specific diameter), or on data structures distributed over the nodes (e.g. a spanning tree). In [24], we consider well known mechanisms consisting of two components: a *prover* that assigns a *certificate* to each node, and a distributed algorithm called *verifier* that is in charge of verifying the distributed proof formed by the collection of all certificates. We show that many network predicates have distributed proofs offering a high level of redundancy, explicitly or implicitly. We use this remarkable property of distributed proofs to establish perfect tradeoffs between the *size of the certificate* stored at every node, and the *number of rounds* of the verification protocol.

The role of unique node identifiers in network computing is well understood as far as *symmetry breaking* is concerned. However, the unique identifiers also *leak information* about the computing environment—in particular, they provide some nodes with information related to the size of the network. It was recently proved that in the context of *local decision*, there are some decision problems that cannot be solved without unique identifiers, but unique identifiers leak a *sufficient* amount of information such that the problem becomes solvable (PODC 2013). In [11], we give a complete picture of what is the *minimal* amount of information that we need to leak from the environment to the nodes in order to solve local decision problems. Our key results are related to *scalar oracles* f that, for any given n, provide a multiset $f(n)$ of n labels; then the adversary assigns the labels to the n nodes in the network. This is a direct generalisation of the usual assumption of unique node identifiers. We give a complete characterisation of the *weakest oracle* that leaks at least as much information as the unique identifiers. Our main result is the following dichotomy: we classify scalar oracles as *large* and *small*, depending on their asymptotic behaviour, and show that (1) any large oracle is at least as powerful as the unique identifiers in the context of local decision problems, while (2) for any small oracle there are local decision problems that still benefit from unique identifiers.

7.3. Models and Algorithms for Networks

7.3.1. Revisiting Radius, Diameter, and all Eccentricity Computation in Graphs through Certificates

In [28], we introduce notions of certificates allowing to bound eccentricities in a graph. In particular , we revisit radius (minimum eccentricity) and diameter (maximum eccentricity) computation and explain the efficiency of practical radius and diameter algorithms by the existence of small certificates for radius and diameter plus few additional properties. We show how such computation is related to covering a graph with certain balls or complementary of balls. We introduce several new algorithmic techniques related to eccentricity computation and propose algorithms for radius, diameter and all eccentricities with theoretical guarantees with respect to certain graph parameters. This is complemented by experimental results on various real-world graphs showing that these parameters appear to be low in practice. We also obtain refined results in the case where the input graph has low doubling dimension, has low hyperbolicity, or is chordal.

7.3.2. Efficient Loop Detection in Forwarding Networks and Representing Atoms in a Field of Sets

In [29], we consider the problem of detecting loops in a forwarding network which is known to be NP-complete when general rules such as wildcard expressions are used. Yet, network analyzer tools such as Netplumber (Kazemian et al., NSDI'13) or Veriflow (Khurshid et al., NSDI'13) efficiently solve this problem in networks with thousands of forwarding rules. In this paper, we complement such experimental validation of practical heuristics with the first provably efficient algorithm in the context of general rules. Our main tool is a canonical representation of the atoms (i.e. the minimal non-empty sets) of the field of sets generated by a collection of sets. This tool is particularly suited when the intersection of two sets can be efficiently computed and represented. In the case of forwarding networks, each forwarding rule is associated with the set of packet headers it matches. The atoms then correspond to classes of headers with same behavior in the network. We

propose an algorithm for atom computation and provide the first polynomial time algorithm for loop detection in terms of number of classes (which can be exponential in general). This contrasts with previous methods that can be exponential, even in simple cases with linear number of classes. Second, we introduce a notion of network dimension captured by the overlapping degree of forwarding rules. The values of this measure appear to be very low in practice and constant overlapping degree ensures polynomial number of header classes. Forwarding loop detection is thus polynomial in forwarding networks with constant overlapping degree.

7.3.3. Exact Distance Oracles Using Hopsets

In [33], we consider for fixed $h \geq 2$ 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 while incurring only a quadratic increase in the query decoding time, and actually offer a speedup 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 hopsets achieving polylog approximation both for space and query time compared to the optimal 3-hopset oracle given the space bound.

7.3.4. Game Theory in Networks

Two notable contributions to game theory applied to networks are worth being mentioned.

In [14], we show that the Preferential Attachment rule naturally emerges in the context of evolutionary network formation, as the *unique* Nash equilibrium of a simple social network game. To demonstrate this result, we start from the fact that each node of a social network aims at maximizing its degree in the future, as this degree is representing its social capital in the "society" formed by the nodes and their connections. We show that, to maximize the node degree in the future, the unique Nash equilibrium consists in playing the Preferential Attachment rule when each node connects to the network. This result provides additional formal support to the commonly used Preferential Attachment model, initially designed to capture the "rich get richer" aphorism. In the process of establishing our result, we expose new connections between Preferential Attachment, random walks, and Young's Lattice.

In [20], we notice that distributed tasks such as constructing a maximal independent set (MIS) in a network, or properly coloring the nodes or the edges of a network with reasonably few colors, are known to admit efficient distributed randomized algorithms. Those algorithms essentially proceed according to some simple generic rules, by letting each node choosing a temptative value at random, and checking whether this choice is consistent with the choices of the nodes in its vicinity. If this is the case, then the node outputs the chosen value, else it repeats the same process. However, although such algorithms are, with high probability, running in a polylogarithmic number of rounds, they are *not robust* against actions performed by rational but selfish nodes. Indeed, such nodes may prefer specific individual outputs over others, e.g., because the formers suit better with some individual constraints. For instance, a node may prefer not being placed in a MIS as it is not willing to serve as a relay node. Similarly, a node may prefer not being assigned some radio frequencies (i.e., colors) as these frequencies would interfere with other devices running at that node. We show that the probability distribution governing the choices of the output values in the generic algorithm can be tuned such that no nodes will rationally deviate from this distribution. More formally, and more generally, we prove that the large class of so-called LCL tasks, including MIS and coloring, admit simple "Luby's style" algorithms where the probability distribution governing the individual choices of the output values forms a Nash equilibrium. In fact, we establish the existence of a stronger form of equilibria, called symmetric trembling-hand perfect equilibria for those games.

INFINE-POST Team

5. New Results

5.1. IoT Scripting Over-The-Air

Participants: Emmanuel Baccelli, Francisco Acosta.

A large part of the Internet of Things (IoT) will consist of interconnecting low-end devices, whose characteristics include very small memory capacity (a few kBytes) and limited energy consumption (1000 times less than a RaspberryPi). IoT use-cases require the orchestration of different pieces of logic running concurrently on low-end IoT devices and elsewhere on the network (e.g. in the cloud) and communicating with one another. In a number of use-cases, the logic that needs to run on low-end IoT devices is not known upfront, before deploying the device(s). For instance, some part of the logic (e.g. pre-processing of some data) may need to be transferred on demand, from the cloud to the device, for privacy or performance reasons. Another example is the fine-tuning of some parameters of the logic running on some device, which can only be done after the deployment (e.g. the sensitivity of a distributed alarm system on-site). In such context, this paper presents a generic approach to host, run and update IoT application logic on heterogeneous low-end devices, using over-the-air scripting and small containers. Based on RIOT and Javascript, we provide a proof- of-concept implementation of this approach for a building automation IoT scenario, as well as a preliminary evaluation of this implementation running on common off-the-shelf low-end IoT hardware. Our evaluation shows the prototype runs on common off-the-shelf low-end IoT hardware with as little as 32kB of memory. Recent prior work in this domain also proposed Actinium, an approach using small, distributed runtime containers on computers proxying for low-end IoT devices, accessible as Web resources, and hosting JavaScript logic. Compared to Actinium, we eliminate the need for Web resource proxying, as runtime containers are running directly on the low-end IoT devices.

This work was published and presented at the IEEE Percom 2018 conference as ”Scripting Over-The-Air: Towards Containers on Low-end Devices in the Internet of Things”.

5.2. Information-centric IoT Robotics

Participants: Loic Dauphin, Cedric Adjih, Emmanuel Baccelli.

As IoT emerges, minibots (miniature robots) have appeared on the market. A large community emerged, designing do-it- yourself minibots, and cheap, re-programmable minibots with communication capabilities are now available. For in- stance, small wheeled robots such as the Zooid are based on a small microcontroller (8kB RAM, 64kB ROM) and communicating with a low-power radio in the 2.4 GHz ISM band. Other examples are cheap drones such as the Cheerson CX-10, which has similar hardware characteristics, and which costs under 15$. Simple robotic arms and legged robots are also available, such as the MetaBot. A current trend bases software embedded in minibots on open source frameworks. The Robot Operating System (ROS) is a software framework for robot application development which has become a de facto standard for most areas in robotics. Other open source robotics frameworks include software suite tailored for drones, some of which provide compatibility with ROS. In fact, we observe that minibots have a number of charac- teristics in common with low-end devices found in the Internet of Things (IoT). Compared to low-end IoT devices, minibots are based on similar hardware and their software follows similar trends. For instance, an IoT-enabled actuator based on a System-on-Chip (SoC) embarking a small microcontroller, and a radio communicating with a remote server, is very similar to a simple radio-controlled robot. Low-end IoT devices use similar radio modules, and software embedded in IoT devices is more and more based on a variety of open source, lightweight operating systems such as RIOT, FreeRTOS and NuttX, among others. Similarly, as for IoT embedded systems, the network com- ponent of minibots represents by itself in important part of the software (in terms of features, code/memory size, and performance). In fact, a wide variety of radio modules and communication protocols are used on minibots. The protocols used by micro-robots

for (internal or external) communication range from direct motor control (pulse width modulation PWM, pulse position modulation PPM, or PCM), to serial/bus protocols, and high level protocols such as Real-time Publish- Subscribe Protocol (RTPS). In this work we thus to explored the potential of bundling open source robotics software frameworks with IoT software and network architectures, to program and control minibots. To do so, we extend our recent work by designing ROS-ready technology for a minibot based on RIOT and ROS2. We focus primarily on software and networking aspects, targeting ultra-lightweight robots based on a reprogrammable SoC with a microcontroller running at approximately 50 MHz, with 10kB RAM, 100kB Flash, and a low-power radio. Using an information-centric networking paradigm extending NDN, we design and implement the communication primitives required by RIOT-ROS2. Our prototype is able to maintain full compatibility between ROS nodes running on the minibot(s) and ROS nodes running elsewhere on the network without the use of a bridge. We show that RIOT-ROS2 fits on low-end robotics hardware such as a System-on-Chip with an ARM Cortex-M0+ microcontroller. On the software and network performance evaluation side, we illustrate that the latency incurred with our ICN approach is completely acceptable for minibot control, even on constrained radio, based on micro-benchmarks.

This work was published and presented at the IEEE PEMWN 2018 conference as "RIOT-ROS2: Low-Cost Robots in IoT Controlled via Information-Centric Networking".

5.3. Human Mobility completion of Sparse Call Detail Records

Participants: Guangshuo Chen, Aline Carneiro Viana, Marco Fiore [CNR - IEIIT (Italy)], Carlos Sarraute [Grandata Labs].

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.

These works have been published as invited papers at the ACM CHANTS 2016 workshop (in conjunction with ACM MobiCom 2016), at the IEEE DAWM workshop (in conjunction with IEEE Percom 2017) and at Computer Communication Elsevier journal in 2018. Another journal version (also registered as TR: hal-01675570) is in revision at the EPJ Data Science Journal.

5.4. Adaptive sampling frequency of human mobility

Participants: Panagiota Katsikouli, Aline Carneiro Viana, Marco Fiore [CNR - IEIIT (Italy)], Diego Madariaga.

The problem we address here is the design of a location sampling system for smartphones and handheld devices that reduces the energy consumed by the continuous activation of the GPS, it reduces the space required to store recorded locations, while reliably capturing the movements of the tracked user. The applications here are related to a number of fields relevant to ubiquitous computing, such as energy-efficient mobile computing, location-based service operations, active probing of subscribers' positions in mobile networks and trajectory data compression.

To this end, we propose an adaptive sampling system without the use of any assisting sensors for the activation of GPS, such as accelerometer, or GSM information. Our system captures the mobility of a user with high accuracy and reliably adjusts the sampling frequency depending on the user's movement. During high mobility, our system densely samples the locations of the tracked user, but at a rate at most the usual rate found today in most applications (e.g., 1 sample per minute). During low mobility, we sample sparsely at much lower rate than usual. As a result, the recorded trace contains much less samples than it would contain if we sampled with the fixed pre-defined sampling rate, requiring less storage space and less energy to activate the GPS.

Our first quest for a response led to the discovery of *(i)* seemingly universal spectral properties of human mobility, and *(ii)* a linear scaling law of the localization error with respect to the sampling interval. Our findings were based on the analysis of fine-grained GPS trajectories of 119 users worldwide. This work was published at the IEEE Globecom 2017 international conference.

We have improved the published sampling approach by incorporating human behavioral features at the sampling decisions to make it more adaptive. This is an on-going work with Panagiota Katsikouli, who spent 5 months in our team working as an internship and is currently doing a Post-Doc at the AGORA Inria team, and Diego Madariaga who spent 3 months in our team working as an internship and is going to start a PhD in co-tutelle with Aline C. Viana. Diego has implemented an Android application to sample mobility data of users according to our adaptive system described here above. The application is currently under deployment and 8 volunteers are running it in their smartphones. The collected data will allow us validating the correctness and performance of our adaptive sampling system. A patent discussion is also on-going with Inria, currently performing a marked/business study.

5.5. Inference of human personality from mobile phones datasets

Participants: Adriano Di Luzio, Aline Carneiro Viana, Julinda Stefa, Katia Jaffres-Runser [INPT-ENSEEIHT - IRIT (Toulouse University)], Alessandro Mei [Sapienza University (Italy) - Dept. of Computer Science].

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. We take a ground-breaking step in (i) deeply capturing human habits in terms of movements, visits, wireless connectivity as well as some routinary actions from a crowdsourced mobility dataset and in (ii) better understanding the relationship between personality traits and individual behavior. We do so by leveraging a dataset collecting very detailed routines of individuals originating from different countries located in 2 different continents, who answered the Big Five Inventory and allowed continuous collection of data from their smartphones for research purposes for 3 years. We use this dataset to engineer a set of human-adapted features that capture three aspects of human behavior: Temporal Mobility (e.g. time at home/work or commuting), Spatial Mobility (e.g. number of most frequent places, maximum distance from home), and the Context of Use (battery charging habits, wireless hotspots availabilities). Then, we use the features that have a statistically significant correlation with the OCEAN traits to predict the personality of a test-set portion of our dataset through cross validation.

Our results attest an accurate prediction of users' personality traits when a 5-level granularity is used per trait. This brings a much higher precision to our predicted results, when compared to the usual 3-level literature granularity. In addition, our prediction methodology carefully takes advantage of engineered features that (1)

are more human-adapted and consequently, allow better capturing individuals' habits in terms of movements, visits, connectivity, context, as well as actions (note that contrarily to the literature, neither calls behavior nor data content is leveraged in our analysis), and (2) are designed having in mind the differences and particularities among the Big5 traits of personality. Thus, this work has the potential to impact the way we characterise unique behaviors of individuals as well as quantify how human personality influences lives and actions. Our results show (1) a significant correlation of most of the traits with a small set of mobility-related features and (2) that we are able to predict the individuals' Big5 traits with considerable accuracy (e.g., prediction of the 5 levels of Openness trait shows an F1 score of 0.77), which is significantly outperforming a benchmark approach, when only considering a set of only 3 of our human-adapted features. Finally, we discuss the ethical concerns of our work, its privacy implications, and ways to tradeoff privacy and benefits.

This is an on-going work with Adriano di Luzio, who spent 4 months in our team working as an internship, Julinda Stefa, an invited research visitor at Infine, and two other researchers: Katia Jaffres-Runser and Alessandro Mei. A paper describing this work is under submission at ACM Mobihoc 2018, but a technical report is also registered under the name hal-01954733.

5.6. Data offloading decision via mobile crowdsensing

Participants: Emanuel Lima, Aline Carneiro Viana, Ana Aguiar [FEUP (Portugal) - Dept. of Electrical and Computer Engineering], Paulo Carvalho [FEUP (Portugal) - Dept. of Electrical and Computer Engineering].

According to Cisco forecasts [0], 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 intend to extend these works by studying human mobility from the perspective of mobile data offloading. This brings two major differences compared to the related work. First, high temporal resolution of positioning datasets is needed. In the majority of the related work, important locations have a temporal dimension representing the time spent by a user in that location, which confers its degree of importance. This time is usually in the order of several minutes which is suitable for the case of detecting important locations but not for a mobile data offloading scenario. Here, according to the amount of data traffic that needs to be offloaded, locations with a visiting temporal resolution of few seconds may be enough for data offloading. Thus, we expect to discover additional offloading opportunities, which were not visible with a coarser temporal resolution. Second, while important locations are usually limited in size, offloading locations can have any arbitrary shape and size.

In this work, offloading regions are defined as spatially aggregated locations where users have mobility suitable to offload. The main contribution of this work are: (a) the identification of offloading regions on an individual basis through unsupervised learning; (b) the characterization of these regions in terms of availability, sojourn, and transition time based on their relevance; (c) the study of the impact of the users mobility on the design of mobile offloading systems. This work was published at ACM CHANTS 2018.

We now working on the extension of this work, which will incorporate the mobility prediction of the users. Such prediction is essential to the design of the decision offloading strategy. Such strategy will be used to allow a mobile phone of a user deciding if offload or not her traffic, i.e., when, where (in which offloading region) and how (if the traffic will be offloadied to one or more Access Points). This is an on-going work with the the PhD Emanuel Lima, who spent 4 months as an intern in our team, and his advisors.

[0] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

5.7. Infering friends in the crowd in Device-to-Device communication

Participants: Rafael Costa, Aline Carneiro Viana, Leobino Sampaio [UFBA (Brazil) - Institute of Mathematics], Artur Ziviani [National Laboratory for Scientific Computing (Brazil)].

The next generation of mobile phone networks (5G) will have to deal with spectrum bottleneck and other major challenges to serve more users with high-demanding requirements. Among those are higher scalability and data rates, lower latencies and energy consumption plus reliable ubiquitous connectivity. Thus, there is a need for a better spectrum reuse and data offloading in cellular networks while meeting user expectations. According to literature, one of the 10 key enabling technologies for 5G is device-to-device (D2D) communications, an approach based on direct user involvement. Nowadays, mobile devices are attached to human daily life activities, and therefore communication architectures using context and human behavior information are promising for the future. User-centric communication arose as an alternative to increase capillarity and to offload data traffic in cellular networks through opportunistic connections among users. Although having the user as main concern, solutions in the user-centric communication/networking area still do not see the user as an individual, but as a network active element. Hence, these solutions tend to only consider user features that can be measured from the network point of view, ignoring the ones that are intrinsic from human activity (e.g., daily routines, personality traits, etc). In this work, we plan to investigate how human-aspects and behavior can be useful to leverage future device-to-device communication.

This is the PhD thesis subject of Rafael Costa, aiming the design of a methodology to select next-hops in a D2D communication that will be human-aware: i.e., that will consider not only available physical 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. A tutorial paper is under submission to a journal (a TR is in hal-01675445) and a 4h-tutorial was presented at the SBRC 2018 conference [0] (the biggest conference on Computer and Network Science in Brazil).

The next step is then the design of forwarding strategies for data offloading through Device-to-Device (D2D) communication, 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.

5.8. Urban Computing Leveraging Location-Based Social Network Data: a Survey

Participants: Thiago H. Silva [UTFPR (Brazil) - Dept. of Computer Science], Aline Carneiro Viana, Antonio Loureiro.

Urban computing is an interdisciplinary area in which urban issues are studied using state-of-the-art computing technologies. This area is at the intersection of a variety of disciplines: sociology, urban planning, civil engineering, computer science, and economics, to name a few. More than half of the world's population today live in cities and, consequently, there is enormous pressure on providing the proper infrastructure to cities, such as transport, housing, water, and energy. To understand and partly tackle these issues, urban computing combines various data sources such as those coming from Internet of Things (IoT) devices; statistical data about cities and its population (e.g., the Census); and data from Location-Based Social Networks (LBSN), sometimes also termed as location-based social media. One fundamental difference between data from LBSNs and data from other sources is that the former offers unprecedented geographic and temporal resolutions: it reflects individual user actions (fine-grained temporal resolution) at the scale of entire world-class cities (global geographic resolution).

Urban computing with LBSN data has its particularities. For instance, users who share data in Foursquare, a popular LBSN, usually have the goal of showing to their friends where they are while also providing personalized recommendations of places they visit. Nevertheless, when correctly analyzed for knowledge extraction, this data can be used to better understand city dynamics and related social, economic, and cultural aspects. To achieve this purpose, new approaches and techniques are commonly needed to explore that data properly.

[0] http://www.sbrc2018.ufscar.br/minicurso-1-mc-1/

In order to better study such needs, we have published at ACM Computing Survey Journal (the ACM journal with highest impact factor) a survey that provides an extensive discussion of the related literature, focusing on major findings and applications. Although its richness concerning knowledge provision, LBSN data presents several challenges, requiring extra attention to its manipulation and usability, which drives future research opportunities in the field of urban computing using LBSN data. Our work is complementary to two existing surveys in the area of urban computing (i.e., by Jiang et al. and by Zheng et al.) since they only mention briefly few studies that explore LBSN data, neglecting key challenges that revolve around LBSNs. We hope that taken together, our effort and these existing ones, provide a broad perspective of urban computing studies and its development through the lens of different data-driven approaches.

5.9. Identifying how places impact each other by means of user mobility

Participants: Lucas Santos, Pedro Olmo [UFMG (Brazil) - Dept. of Computer Science], 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. In other words, how important is this POI for its neighborhood? We proposed thus a framework to measure POI influence by means of LBSN data. First, we modeled the problem using mobility graph approach where each POI is a node and the transitions of users among POI is a weighted vertex. Also, we treat the users' check-in records among POI as a measure of uncertainty, and their strength can be measured by entropy, which enabled to measure direct influence. Second, using same graph, we propose another influence measure taking account the POI importance for its one-hop vicinity in terms of incoming human transition. In addition, this mobility graph can be viewed as a collaborative filtering. We use this collaborative filter for compute the G-causality and evaluate if the transitions among POI has a causal relation and consequently, the influence among POI. Moreover, to the best of our knowledge, we are the first study which investigated POI influence by means of human mobility using LBSN data source.

This work is being prepared for a submission to an international conference.

Neo Project-Team

7. New Results

7.1. Stochastic Modeling

Participants: Eitan Altman, Konstantin Avrachenkov, Mandar Datar, Swapnil Dhamal, Alain Jean-Marie, Albert Sunny.

7.1.1. Markov chains with restart/jumps

In [7], K. Avrachenkov together with A. Piunovskiy and Y. Zhang (Univ. of Liverpool, UK) consider a discrete-time Markov process with restart. At each step the process either with a positive probability restarts from a given distribution, or with the complementary probability continues according to a Markov transition kernel. The main contribution of this work is an explicit expression for the expectation of the hitting time (to a given target set) of the process with restart. The formula is convenient when considering the problem of optimization of the expected hitting time with respect to the restart probability. The results with are illustrated with two examples in uncountable and countable state spaces and with an application to network centrality.

Then, in [19], K. Avrachenkov and I. Bogdanov (HSE, Russia) study the relaxation time in the random walk with jumps. The random walk with jumps combines random walk based sampling with uniform node sampling and improves the performance of network analysis and learning tasks. They derive various conditions under which the relaxation time decreases with the introduction of jumps.

7.1.2. Markov modeling of Lasers

A. Jean-Marie has continued the investigation of Markov models of Lasers at several levels of physical accuracy, in conjunction with F. Philippe, L. Chusseau and A. Vallet (Univ. Montpellier and CNRS). In [17], a Markov model of relatively low complexity, the "Canonical Markov Model" (CMM), is built on the basis of a time-scale decomposition of physical phenomena. This simplified model is validated by comparison with a "microscopic Markov model" previously existing. Thanks to its smaller state space, simulations with the CMM are orders of magnitude faster, and numerical investigation of stationary and transient features become possible. As an example, the focus is put in [17], [39] on the Laser "threshold", a phenomenon related to sojourn of the CMM in states where no light is emitted. Simulations and numerical solutions reveal the existence of a bi-modal distribution for the particles for a certain range of parameters, thereby predicting a certain instability of the Laser for these values. Investigations continue with a quantification of the intensity of "flashes" through the computation of hitting times in the CMM.

7.1.3. The marmoteCore platform

The development of `marmoteCore` (see Section 6.1) has been pursued by A. Jean-Marie. The software library is now being used in NEO's research projects such as [17] or queuing models supporting the analysis of Green Data Centers. `marmoteCore` provides the classes necessary to represent the state space of Markov models, from the elementary bricks that are interval or rectangular domains, simplices, or binary sequences. From there, the user easily programs the construction of probability transition matrices or infinitesimal generators. Structural analysis methods allow to identify recurrent and transient classes, and to compute the period of the model. Numerous methods allow the Monte Carlo simulation of the chain, the computation of transient and stationary distributions, as well as hitting times. In conjunction with E. Hyon (Univ. Paris-Nanterre), extensions of the core of the software are being programmed for Markov Decision Processes and Stochastic Games.

7.1.4. Blockchain mining

S. Dhamal, T. Chahed (Telecom SudParis), W. Ben-Ameur (Telecom SudParis), E. Altman, A. Sunny, and S. Poojary (UAPV, the Univ. of Avignon) have studied a stochastic game framework for distributed computing settings such as blockchain mining in [42]. A continuous-time Markov chain model, where players arrive and depart according to a stochastic process, is proposed, and their investment strategies are determined based on the state of the system. Two scenarios are analyzed, based on whether the rate of problem getting solved is dependent on or independent of the computational power invested by the players. The equilibrium strategies are shown to follow a threshold policy when this rate is proportional to the total invested power, while the players are shown to invest proportionally to the reward-cost ratio when this rate is independent of the invested power. The effects of arrival and departure rates on the players' utilities are quantified using simulations.

The paper extends the game theoretic modeling and analysis of the static case (fixed number of miners) done in [18] by E. Altman in collaboration with A. Reiffers-Masson (IISc, India), D. Sadoc Menasché (UFRJ), M. Datar and S. Dhamal, and C. Touati (Inria Grenoble Rhône-Alpes).

7.2. Queueing Theory

Participants: Sara Alouf, Konstantin Avrachenkov, Alain Jean-Marie, Dimitra Politaki.

7.2.1. Multiclass processor sharing and random order scheduling policies

In [2], K. Avrachenkov and T. Bodas (LAAS-CNRS) consider a single server system serving a multiclass population. Some popular scheduling policies for such system are the discriminatory processor sharing (DPS), discriminatory random order service (DROS), generalized processor sharing (GPS) and weighted fair queueing (WFQ). In this work, the authors propose two classes of policies, namely MPS (Multi-class Processor Sharing) and MROS (Multi-class Random Order Service), that generalize the four policies mentioned above. For the special case when the multi-class population arrive according to Poisson processes and have independent and exponential service requirement with parameter μ, they show that the tail of the sojourn time distribution for a class i customer in a system with the MPS policy is a constant multiple of the tail of the waiting time distribution of a class i customer in a system with the MROS policy. This result implies that for a class i customer, the tail of the sojourn time distribution in a system with the DPS (GPS) scheduling policy is a constant multiple of the tail of the waiting time distribution in a system with the DROS (respectively WFQ) policy.

7.2.2. The marmoteCore-Q tool

Using the `marmoteCore` platform, a tool called `marmoteCore-Q` has been developed by D. Politaki under the supervision of S. Alouf and A. Jean-Marie for the simulation of a family of queueing models based on the general BMAP/PH/c queue with impatience and resubmissions. There exist many special cases of this queue for which analytical results are known. Examples are: the M/M/1 queue and its finite capacity version, the M/M/c/K queue, the M/PH/1 and M/PH/∞ queues, the M^X/M/1 and M^X/M/∞ queues. Such examples are used to validate the implementation of the `marmoteCore-Q` tool.

7.3. Random Graph and Matrix Models

Participants: Konstantin Avrachenkov, Maximilien Dreveton.

In [5], K. Avrachenkov, together with A. Kadavankandy (CentraleSupélec) and N. Litvak (Univ. of Twente, The Netherlands), analyse a mean-field model of Personalized PageRank on the Erdös-Rényi random graph containing a denser planted Erdös-Rényi subgraph. They investigate the regimes where the values of Personalized PageRank concentrate around the mean-field value. They also study the optimization of the damping factor, the only parameter in Personalized PageRank. Their theoretical results help to understand the applicability of Personalized PageRank and its limitations for local graph clustering.

7.4. Data Analysis and Learning

Participant: Konstantin Avrachenkov.

7.4.1. Unsupervised learning

In [6], K. Avrachenkov, together with A. Kondratev, V. Mazalov (Petrozavodsk State Univ., Russia) and D. Rubanov (Amadeus), applied game-theoretic methods for community detection in networks. The traditional methods for detecting community structure are based on selecting dense subgraphs inside the network. Here the authors propose to use the methods of cooperative game theory that highlight not only the link density but also the mechanisms of cluster formation. Specifically, they suggest two approaches from cooperative game theory: the first approach is based on the Myerson value, whereas the second approach is based on hedonic games. Both approaches allow to detect clusters with various resolutions. However, the tuning of the resolution parameter in the hedonic games approach is particularly intuitive. Furthermore, the modularity-based approach and its generalizations as well as ratio cut and normalized cut methods can be viewed as particular cases of the hedonic games. Finally, for approaches based on potential hedonic games a very efficient computational scheme using Gibbs sampling is suggested.

7.4.2. Semi-supervised learning

Graph Semi-supervised learning (gSSL) aims to classify data exploiting two initial inputs: firstly, the data are structured in a network whose edges convey information on the proximity, in a wide sense, of two data points (e.g. correlation or spatial proximity) and, second, there is a partial information on some nodes, which have previously been labelled. Thus, the classification problem is usually a balance between two terms: one diffusing the information from the labelled points to the unlabelled ones through the network and another one that constrains the solution to be similar, on the labelled nodes, to the given labels. In practice, popular SSL methods as Standard Laplacian (SL), Normalized Laplacian (NL) or PageRank (PR), exploit those operators defined on graphs to spread the labels and, from a random walk perspective, the classification of a given point is given the maximum of the expected number of visits from one class. Anomalous diffusion can alter the way a graph is "explored" and, therefore, it can alter classification performance. In a nushell, Lévy flights/walks are a way to create superdiffusive regimes: the customary rule for their ignition is to allow the walkers to perform non-local jumps, whose length is distributed according to a fat-tailed probability density function with diverging second moment. Mathematically speaking, there have been several attempts to convert the Lévy flight phenomenon on networks and, in the context of gSSL, K. Avrachenkov in conjunction with S. De Nigris, E. Bautista, P. Abry and P. Gonçalves, settled in [38] for the use of fractional operators. In this SSL context, the authors cast those operators in the SSL problem in each different incarnation (SL, PR and NL) and investigated the beneficial effect of such a procedure for classification.

In [13], K. Avrachenkov, together with A. Kadavankandy (CentraleSupélec), L. Cottatellucci (EURECOM) and R. Sundaresan (IISc, India), tackle the problem of hidden community detection. We consider Belief Propagation (BP) applied to the problem of detecting a hidden Erdös-Rényi (ER) graph embedded in a larger and sparser ER graph, in the presence of side-information. We derive two related algorithms based on BP to perform subgraph detection in the presence of two kinds of side-information. The first variant of side-information consists of a set of nodes, called cues, known to be from the subgraph. The second variant of side-information consists of a set of nodes that are cues with a given probability. It was shown in past works that BP without side-information fails to detect the subgraph correctly when a so-called effective signal-to-noise ratio (SNR) parameter falls below a threshold. In contrast, in the presence of non-trivial side-information, we show that the BP algorithm achieves asymptotically zero error for any value of a suitably defined phase-transition parameter. We validate our results on synthetic datasets and a few real world networks.

7.4.3. Supervised learning

Graphlets are defined as k-node connected induced subgraph patterns. For instance, for an undirected graph, 3-node graphlets include closed triangles and open triangles. The number of each graphlet, called graphlet count, is a signature which characterizes the local network structure of a given graph. Graphlet count plays a prominent role in network analysis of many fields, most notably bioinformatics and social science. However, computing exact graphlet count is inherently difficult and computational expensive because the number of graphlets grows exponentially large as the graph size and/or graphlet size grow. To deal with this difficulty, many sampling methods were proposed to estimate graphlet count with bounded error. Nevertheless, these

methods require large number of samples to be statistically reliable, which is still computationally demanding. Intuitively, learning from historic graphs can make estimation more accurate and avoid many repetitive counting to reduce computational cost. Based on this idea, in [29] K. Avrachenkov, together with X. Liu, J. Chen and J. Lui (CUHK, Hong Kong), propose a convolutional neural network (CNN) framework and two preprocessing techniques to estimate graphlet count. Extensive experiments on two types of random graphs and real world biochemistry graphs show that their framework can offer substantial speedup on estimating graphlet count of new graphs with high accuracy.

7.5. Game Theory

Participants: Eitan Altman, Swapnil Dhamal.

7.5.1. Resource allocation polytope games

S. Dhamal, W. Ben-Ameur, T. Chahed (both from Telecom SudParis), and E. Altman have studied two-player resource allocation polytope games in [24]. The strategy of a player is considered to be restricted by the strategy of the other player, with common coupled constraints. In the context of such games, novel notions of independent optimal strategy profile and common contiguous set are introduced. Necessary and sufficient conditions are derived for the game to have a unique pure strategy Nash equilibrium. Given an instance of the game, an efficient algorithm is presented to compute the price of anarchy. Under reasonable conditions, the price of stability is shown to be 1. A paradox is shown that higher budgets may lead to worse outcomes.

7.6. Applications in Telecommunications

Participants: Zaid Allybokus, Sara Alouf, Eitan Altman, Konstantin Avrachenkov, Swapnil Dhamal, Alain Jean-Marie, Giovanni Neglia, Dimitra Politaki.

7.6.1. Caching

A fundamental brick of the information-centric architectures proposed for Internet evolution is in-network caching, i.e. the possibility for the routers to store locally the contents and directly serve future requests. This has raised a new interest in the performance of networks of caches. Since 2012, there has been a significant research activity in NEO on this topic. Our work raised the attention of researchers at Akamai Technologies (the world leader in Content Delivery Networks). In real caching systems the hit rate is often limited by the speed at which contents can be retrieved by the Hard-Disk Drive (HDD) (this is the so-called *spurious misses*' problem). Akamai researchers asked us to design an algorithm to solve this problem. In [43] G. Neglia and D. Tsigkari, together with D. Carra (Univ. of Verona, Italy), M. Feng, V. Janardhan (Akamai Technologies, USA), and P. Michiardi (EURECOM) have proposed a simple randomized caching policy that makes optimal use of the RAM to minimize the load on the HDD and then the number of spurious misses. Moreover, experiments in Akamai CDN have shown that our policy reduces the HDD load by an additional 10% in comparison to the (highly optimized) baseline policy currently employed by Akamai. In [15] a subset of the same authors (G. Neglia, D. Carra, P. Michardi) have shown that the same approach can be adapted to minimize any miss cost function as far as the cost is additive over the misses.

More recently, we moved to consider the problem of caches' coordination in a dense cellular network scenario, where caches are deployed at base stations (BSs) and a user can potentially retrieve the content from multiple BSs. In this setting, the optimal content placement problem is NP-hard even when the goal is simply to maximize the hit ratio. Most of the existing literature has proposed heuristics assuming that content popularities are static and known, but in reality their estimation can be very difficult at the scale of the geographical area covered by a BS. In [14] E. Leonardi (Politecnico di Torino, Italy) and G. Neglia have introduced a class of simple and fully distributed caching policies, which require neither direct communication among BSs, nor a priori knowledge of content popularity (strongly deviating from the assumptions of existing literature). They have shown that optimal coordination can be achieved by applying minor changes to existing policies and piggybacking an additional information bit to each content request. How to achieve coordination for more complex performance metrics (e.g. the retrieval time or fairness) is still an open research problem that is now the PhD subject of G. Iecker, co-supervised by G. Neglia and T. Spyropoulos (EURECOM).

7.6.2. Modeling and workload characterization of data center clusters

There are many challenges faced when modeling computing clusters. In such systems, jobs to be executed are submitted by users. These jobs may generate a large number of tasks. Some tasks may be executed more than once while other may abandon before execution. D. Politaki, S. Alouf, F. Hermenier (Nutanix), and A. Jean-Marie have developed a multi-server queueing system with abandonments and resubmissions to model computing clusters. To capture the correlations observed in real workload submissions, a Batch Markov Arrival Process is considered. The service time is assumed to have a phase-type distribution. This model has not been analyzed in the literature. The distributions of the interarrivals and the service times found in the Google Cluster Data have been characterized and compared with fitted distributions. The authors findings support the model assumptions. Ongoing work investigates the approaches that can be adopted to overcome the technical challenges found in the performance evaluation of the computing clusters. In particular, the developed tool `marmoteCore-Q` (see §7.2.2) will be used.

To understand the essential characteristics of a computing cluster for modelling purposes, the same authors have looked into two datasets consisting of job scheduler logs. The first dataset comes from a Google cluster and is publicly available (https://github.com/google/cluster-data). The second dataset has been collected from the internal computing cluster of Inria Sophia-Antipolis Méditerrannée. After a preliminary analysis and sanitizing of each dataset, a numerical analysis is performed to characterize the different stochastic processes taking place in the computing cluster. In particular, the authors characterize the impatience process, the re-submission process, the arrival process (batch sizes and correlations) and the service time, considering the impact of the scheduling class and of the execution type.

7.6.3. Software Defined Networks (SDN)

The performance of computer networks relies on how bandwidth is shared among different flows. Fair resource allocation is a challenging problem particularly when the flows evolve over time. To address this issue, bandwidth sharing techniques that quickly react to the traffic fluctuations are of interest, especially in large scale settings with hundreds of nodes and thousands of flows. In this context, K. Avrachenkov and Z. Allybokus, together with J. Leguay (Huawei Research) and L. Maggi (Nokia Bell Labs), in [1] propose a distributed algorithm based on the Alternating Direction Method of Multipliers (ADMM) that tackles the multi-path fair resource allocation problem in a distributed SDN control architecture. Their ADMM-based algorithm continuously generates a sequence of resource allocation solutions converging to the fair allocation while always remaining feasible, a property that standard primal-dual decomposition methods often lack. Thanks to the distribution of all computer intensive operations, they demonstrate that large instances can be handled at scale.

7.6.4. Impulsive control of G-AIMD dynamics

Motivated by various applications from Internet congestion control to power control in smart grids and electric vehicle charging, in [20] K. Avrachenkov together with A. Piunovskiy and Y. Zhang (Univ. of Liverpool, UK) study Generalized Additive Increase Multiplicative Decrease (G-AIMD) dynamics under impulsive control in continuous time with the time average alpha-fairness criterion. They first show that the control under relaxed constraints can be described by a threshold. Then, they propose a Whittle-type index heuristic for the hard constraint problem. They prove that in the homogeneous case the index policy is asymptotically optimal when the number of users is large.

7.6.5. Application of Machine Learning to optimal resource allocation in cellular networks

In [9], E. Altman in collaboration with A. Chattopadhyay and B. Błaszczyszyn (from Inria DYOGENE team) consider location-dependent opportunistic bandwidth sharing between static and mobile downlink users in a cellular network. 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 Markov decision process (MDP) where the per-step reward is a linear combination of instantaneous data volumes received by static and mobile users. The transition structure of this MDP is not known in general. They thus propose a learning algorithms based on

stochastic approximation with one and with two time scales. The results are extended to address the issue of fair bandwidth sharing between the two classes of users.

To optimize routing of flows in datacenters, SDN controllers receive a packet-in message whenever a new flow appears in the network. Unfortunately, flow arrival rates can peak to millions per second, impairing the ability of controllers to treat them on time. Flow scheduling copes with this by segmenting the traffic between elephant and mice flows and by treating elephant flows in priority, as they disrupt short lived TCP flows and create bottlenecks. In [21], E. Altman in collaboration with F. De Pellegrini (UAPV), L. Maggi (Huawei), A. Massaro (FBK Trento), D. Saucez (Inria DIANA team) and J. Leguay (Huawei Research) propose a stochastic approximation based learning algorithm called SOFIA and able to perform optimal online flow segmentation. Extensive numerical experiments characterize the performance of SOFIA.

7.6.6. Forecast Scheduling

With the age of big data and with geo-localisation measurements available, the precision in predicting the mobility of users increases, and hence also that of the prediction of channel conditions. In [35], E. Altman in collaboration with H. Zaaraoui, S. Jema, Z. Altman (Orange Labs) and T. Jimenez (UAPV) propose a convex optimization approach to Forecast Scheduling which makes use of current and future predicted channel conditions to obtain an optimal alpha fair schedule. They further extend the model in [34] to take into account different types of random events such as arrival and departure of users and uncertainties in the mobile trajectories. Simulation results illustrate the significant performance gain achieved by the Forecast Scheduling algorithms in the presence of random events.

7.6.7. Fairness in allocation to users with different time constraints

E. Altman and S. Ramanath (IIT Bombay, India) study in [31] how to allocate resources fairly when different users have different time constraints for using the resources. They formulate this as a Markov Decision Process (MDP) for a two user case and provide a Dynamic Program (DP) solution. Simulation results in an LTE framework are provided to support the theoretical claims.

7.7. Applications in Social Networks

Participants: Eitan Altman, Konstantin Avrachenkov, Swapnil Dhamal, Giovanni Neglia.

7.7.1. Fairness in Online Social Network Timelines

Facebook News Feed personalization algorithm has a significant impact, on a daily basis, on the lifestyle, mood and opinion of millions of Internet users. Nonetheless, the behavior of such algorithm lacks transparency, motivating measurements, modeling and analysis in order to understand and improve its properties. E. Altman and G. Neglia, together with other researchers from THANES team (E. Hargreaves and D. Menasché from UFRJ, A. Reiffers-Masson from IIsc, and E. Altman) and with the journalist C. Agosti (Univ. of Amsterdam), have proposed a reproducible methodology encompassing measurements, an analytical model and a fairness-based News Feed design. The model leverages the versatility and analytical tractability of time-to-live (TTL) counters to capture the visibility and occupancy of publishers over a News Feed. Measurements from 2018 Italian political election are used to parameterize and to validate the expressive power of the proposed model. Then, we have conducted a what-if analysis to assess the visibility and occupancy bias incurred by users against a baseline derived from the model. Our results indicate that a significant bias exists and it is more prominent at the top position of the News Feed. In addition, we have found that the bias is non-negligible even for users that are deliberately set as neutral with respect to their political views, motivating the proposal of a novel and more transparent fairness-based News Feed design. This is a very recent research direction, but it has already led to 4 publications [36], [27], [28], [12] with a *best paper award* for [36].

7.7.2. Sampling online social networks

In the framework of network sampling, random walk (RW) based estimation techniques provide many pragmatic solutions while uncovering the unknown network as little as possible. Despite several theoretical advances in this area, RW based sampling techniques usually make a strong assumption that the samples are in stationary regime, and hence are impelled to leave out the samples collected during the burn-in period. In [4] K. Avrachenkov, together with V.S. Borkar (IIT Bomaby, India), A. Kadavankandy (CentraleSupélec) and J.K. Sreedharan (Purdue Univ., USA), propose two sampling schemes without burn-in time constraint to estimate the average of an arbitrary function defined on the network nodes, for example, the average age of users in a social network. The central idea of the algorithms lies in exploiting regeneration of RWs at revisits to an aggregated super-node or to a set of nodes, and in strategies to enhance the frequency of such regenerations either by contracting the graph or by making the hitting set larger. Our first algorithm, which is based on reinforcement learning (RL), uses stochastic approximation to derive an estimator. This method can be seen as intermediate between purely stochastic Markov chain Monte Carlo iterations and deterministic relative value iterations. The second algorithm, which we call the Ratio with Tours (RT)-estimator, is a modified form of respondent-driven sampling (RDS) that accommodates the idea of regeneration. We study the methods via simulations on real networks. We observe that the trajectories of RL-estimator are much more stable than those of standard random walk based estimation procedures, and its error performance is comparable to that of respondent-driven sampling (RDS) which has a smaller asymptotic variance than many other estimators. Simulation studies also show that the mean squared error of RT-estimator decays much faster than that of RDS with time. The newly developed RW based estimators (RL- and RT-estimators) allow to avoid burn-in period, provide better control of stability along the sample path, and overall reduce the estimation time.

7.7.3. Crawling ephemeral content

In [3], K. Avrachenkov and V.S. Borkar (IIT Bombay, India) consider the task of scheduling a crawler to retrieve from several sites their ephemeral content. This is content, such as news or posts at social network groups, for which a user typically loses interest after some days or hours. Thus development of a timely crawling policy for ephemeral information sources is very important. The authors first formulate this problem as an optimal control problem with average reward. The reward can be measured in terms of the number of clicks or relevant search requests. The problem in its exact formulation suffers from the curse of dimensionality and quickly becomes intractable even with a moderate number of information sources. Fortunately, this problem admits a Whittle index, a celebrated heuristics which leads to problem decomposition and to a very simple and efficient crawling policy. The authors derive the Whittle index for a simple deterministic model and provide its theoretical justification. They also outline an extension to a fully stochastic model.

7.7.4. Posting behavior

In [32], E. Altman in collaboration with A. Reiffers-Masson (IISc, India), Y. Hayel and G. Marrel (UAPV) consider a "generalized" fractional program in order to solve a popularity optimization problem in which a source of contents controls the topics of her contents and the rate with which posts are sent to a time line. The objective of the source is to maximize its overall popularity in an Online Social Network (OSN). The authors propose an efficient algorithm that converges to the optimal solution of the Popularity maximization problem.

7.7.5. Recommendation system for OSNs

When a user interested in a service/item, visits an online web-portal, it provides description of its interest through initial search keywords. The system recommends items based on these keywords. The user is satisfied if it finds the item of its choice and the system benefits, otherwise the user explores an item from the list. In [33], E. Altman in collaboration with K. Veeraruna, S. Memon, M. Hanawal and R. Devanand (IEOR IIT Bombay, India), develop algorithms that efficiently utilize user responses to recommended items and find the item of user's interest quickly. The authors first derive optimal policies in the continuous Euclidean space and adapt the same to the space of discrete items.

7.7.6. Opinion dynamics

S. Dhamal, W. Ben-Ameur, T. Chahed (both from Telecom SudParis), and E. Altman have studied the problem of optimally investing in nodes of a social network, wherein two camps attempt to maximize adoption of their respective opinions by the population. In [11], 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. In [23], the possibility of campaigning in multiple phases is explored, where the final opinion of a node in a phase acts as its initial bias for the next phase. A further intricate setting where a camp's influence on a node also depends on the node's initial bias, is analyzed in [22]. Extensive simulations are conducted on real-world social networks for all the considered settings.

7.7.7. Information diffusion under practical models

S. Dhamal has studied the effectiveness of adaptive seeding in multiple phases under the independent cascade model of information diffusion, in [25]. The effect on the mean and standard deviation of the extent of diffusion is observed, with an explanation of how adaptive seeding reduces uncertainty in diffusion. The other aspects studied are: how the number of phases impacts the effectiveness of diffusion, how the diffusion progresses phase-by-phase, and how to optimally split the total seeding budget across phases. Another study [26] generalizes the linear threshold model to account for multiple product features, and presents an integrated framework 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.

7.8. Applications to Energy

Participant: Giovanni Neglia.

7.8.1. Smart grids

Balancing energy demand and production is becoming a more and more challenging task for energy utilities because of the larger penetration of renewable energies, more difficult to predict and control. While the traditional solution is to dynamically adapt energy production to follow the time-varying demand, a new trend is to drive the demand itself. We have first considered the direct control of inelastic home appliances, whose energy consumption cannot be shaped, but simply deferred. Our solution does not suppose any particular intelligence at the appliances, the actuators are rather smart plugs, simple devices with communication capabilities that can be inserted between appliances' plugs and power sockets and are able to interrupt/reactivate power flow. During previous years we have considered both closed-loop and open-loop control of such devices in order to satisfy a probabilistic bound on the aggregated power consumption. Recently, G. Neglia, together with L. Giarré (Univ. di Modena e Reggio Emilia, Italy), I. Tinnirello and G. Di Bella (Univ. di Palermo, Italy) have considered a mixed approach [16]. They have been able to quantify the trade-off between the amount of controlled power and delays experienced by the users to evaluate to which scale this solution should be deployed.

We have also looked at Demand-Response (DR) programs, whereby users of an electricity network are encouraged by economic incentives to re-arrange their consumption in order to reduce production costs. Several recent works proposed DR mechanisms relying on a macroscopic description of the population that does not model individual choices of users. In [8], G. Neglia, together with A. Benegiamo (EURECOM/Inria) and P. Loiseau (EURECOM) has shown that these macroscopic models hide important assumptions that can jeopardize the mechanisms' implementation (such as the ability to make personalized offers and to perfectly estimate the demand that is moved from a timeslot to another). Then, starting from a microscopic description that explicitly models each user's decision, they have introduced new DR mechanisms with various assumptions on the provider's capabilities. Contrarily to previous studies, they have found that 1) the resulting

optimization problems are complex and can be solved numerically only through heuristics, 2) the savings from DR mechanisms are significantly lower than those suggested by previous studies.

POEMS-POST Team

7. New Results

7.1. New schemes for time-domain simulations

7.1.1. Solving the Isotropic Linear Elastodynamics Equations Using Potentials

Participant: Patrick Joly.

This work is done in collaboration with Sébastien Impériale (EPI M3DISIM) and Jorge Albella and Jeronimo Rodríguez from the University of Santiago de Compostela.

We pursue our research on the numerical solution of 2D elastodynamic equations in piecewise homogeneous media using the decomposition of the displacement fields into the sum of the gradient and the rotational (respectively) of two scalar potentials potentials. This allows us to obtain an automatic decomposition of the wave field into the sum of pressure and shear waves (respectively). The approach is expected to be efficient when the velocity of shear waves is much smaller than the velocity of pressure waves, since one can adapt the discretization to each type of waves. This appears as a challenge for finite element methods , the most delicate issue being the treatment of boundary and transmission conditions, where the two potentials are coupled..

A stable (mixed) variational formulation of the evolution problem based on a clever choice of Lagrange multipliers has been proposed as well as various finite element approximations which have been successfully implemented. The analysis of the continuous problem has been published in a long paper in the journal of Scientific computing.The numerical analysis of the discretized problem is in progress.

7.1.2. Time domain Half-Space Matching method

Participants: Sonia Fliss, Hajer Methenni.

This work is done in the framework of the PhD of Hajer Methenni (funded by CEA-LIST) and in collaboration with Sebastien Imperiale (EPI M3DISIM) and Alexandre Imperiale (CEA-LIST).

The objective of this work is to propose a numerical method to solve the elastodynamics equations in a locally perturbed unbounded anisotropic media. Let us mention that all the classical methods to restrict the computation around the perturbations are unstable in anisotropic elastic media (PMLs for instance) or really costly (Integral equations). The idea is to extend the method already developed for the corresponding time harmonic problem, called the Halfspace Matching Method. We have considered, for now, the 2D scalar wave equation but the method is constructed in order to be applied to the elastodynamic problem. The method consists in coupling several representations of the solution in half-planes surrounding the defect with a FE representation in a bounded domain including the defect. In order to ensure the stability of the method, we first semi-discretize in time the equations and apply the method to the semi-discrete problem. Thus, for each time step, by ensuring that all the representations of the solution match, in particular in the intersection of the half-planes, we end up, at each time step, with a system of equations which couples, via integral operators, the solution at this time step in the bounded domain and its traces on the edge of the half-planes, the right hand side being a convolution operator involving the solution at the previous time steps. The method has been implemented and validated with Xlife++.

We are now looking to make the method more efficient by implementing methods of acceleration. Finally, we will also seek to develop another version of the method based on the Convolution quadrature.

7.1.3. Time domain modelling for wave propagation in fractal trees

Participants: Patrick Joly, Maryna Kachanovska.

In order to simulate wave propagation in fractal trees (see section 7.4.3), which have infinite structure, it is necessary to be able to truncate the computations to a finite subtree. This was done using Dirichlet-to-Neumann (DtN) operators in our previous work in collaboration with A. Semin (TU Darmstadt). In this case a DtN operator is a convolution operator, whose kernel is not known in a closed form. Based on the results of this previous work, in 2017 we had proposed two methods for approximating these convolution operators:

- constructing an exact DtN operator for a semi-discretized system (in the spirit of convolution quadrature methods).
- truncating meromorphic expansion for the symbol (Fourier transform of the convolution kernel) of the DtN operator, which allows to approximate the DtN operator by local operators.

This year we have performed a complete convergence and stability analysis of these methods, based on the energy techniques.

In particular, for the convolution quadrature methods, we were able to obtain all the estimates using time-domain analysis, by avoiding passage to the Laplace domain.

As for the method based on the meromorphic expansion of the symbol of the DtN operator, we have shown that the error induced by truncating the expansion to L terms can be controlled by a remainder of a series, which, in particular, depends on the eigenvalues of the weighted Laplacian on the fractal trees. To obtain an explicit dependence of the error on L, we have computed Weyl bounds for the eigenvalues, based on a refinement of the ideas of [Kigami, Lapidus, Comm. Math. Phys. 158 (1993)].

Additionally, we have addressed some computational aspects of the two methods, in particular, efficient evaluation of the symbol of the DtN operator (we have an algorithm that allows to evaluate it at the frequency ω in $O(\log^k |\omega|)$ time), as well as a method for efficient computation of the poles of the symbol (based on Möbius transform and polynomial interpolation).

7.2. Integral equations and boundary element methods (BEMs)

7.2.1. Accelerated and adapted BEMs for wave propagation

Participants: Faisal Amlani, Stéphanie Chaillat.

This work is done in collaboration with Adrien Loseille (EPI Gamma3).

We extend to high-order curved elements a recently introduced metric-based anisotropic mesh adaptation strategy for accelerated boundary element methods (e.g. Fast Multipole(FM-) BEM) applied to exterior boundary value problems. This method derives from an adaptation framework for volumetric finite element methods and is based on an iterative procedure that completely remeshes at each refinement step and that leads to a strategy that is independent of discretization technique (e.g., collocation or Galerkin) and integral representation (e.g., single- or double-layer). In effect, it results in a truly anisotropic adaptation that alters the size, shape and orientation of each element according to an optimal metric based on a numerically recovered Hessian of the boundary solution. The algorithm is principally characterized by its ability to recover optimal convergence rates for both flat and curved discretizations (e.g. P_0-, P_1- or P_2-elements) of a geometry containing singularities such as corners and edges. This is especially powerful for realistic geometries that include engineering detail (whose solutions often entail severe singular behavior).

Additionally, we address — by way of introducing hierarchical ($\mathcal{H}$-) matrix preconditioning applied to fast multipole methods via a Flexible GMRES (FGMRES) routine — the computational difficulties that arise when resolving highly anisotropic (and hence highly ill-conditioned) linear systems. The new technique, which uses a very coarse $\mathcal{H}$-matrix system (constructed rapidly via high-performance parallelization) to precondition the full Fast Multipole Method system, drastically reduces the overall computation time as well as the iterative solve time, further improving the tractability of addressing even larger and more complex geometries by FM-BEM.

7.2.2. Preconditioned $\mathcal{H}$-matrix based BEMs for wave propagation

Participants: Stéphanie Chaillat, Patrick Ciarlet, Félix Kpadonou.

We are interested with fast boundary element methods (BEMs) for the solution of acoustic and elastodynamic problems.

The discretisation of the boundary integral equations, using BEM, yields to a linear system, with a fully-populated matrix. Standard methods to solve this system are prohibitive in terms of memory requirements and solution time. Thus one is rapidly limited in terms of complexity of problems that can be solved. The $\mathcal{H}$-matrix based BEMs is commonly used to address these limitations. It is a purely algebraic approach.

The starting point is that the BEM matrix can be partitioned into some blocks which can either be of low or full rank. Memory can be saved by using low-rank revealing technique such as the Adaptive Cross Approximation. We have already study the efficiency of this approach for wave propagation problems. The purpose being the applications to large scale problems, we are now interested in an efficient implementation of the solver in a high performance computing setting. Thus, a bottleneck, with an hierarchical matrix data-sparse representation, is the management of the memory and its (prior) estimation for array allocations.

The first part of our work has been devoted to the proposition of an a priori estimation of the ranks of the blocks in the hierarchical matrix. Afterwards, we have implemented a parallel construction of the $\mathcal{H}$-matrix representation and H-matrix vector product (basic operation in any iterative solver), using a multi-threading OpenMP parallelization. The solution is then computed through the GMRES iterative solver. A crucial point is then the solution time of that solver and the number of iterations as the problem complexity increases. We have developed a two-level, nested outer-inner, iterative solver strategy. The inner solver preconditioned the outer. The preconditioner is a coarse data-sparse representation of the BEM system matrix.

7.2.3. Coupling integral equations and high-frequency methods

Participants: Marc Bonnet, Marc Lenoir, Eric Lunéville, Laure Pesudo.

This theme concerns wave propagation phenomena which involve two different space scales, namely, on the one hand, a medium scale associated with lengths of the same order of magnitude as the wavelength (medium-frequency regime) and on the other hand, a long scale related to lengths which are large compared to the wavelength (high-frequency regime). Integral equation methods are known to be well suited for the former, whereas high-frequency methods such as geometric optics are generally used for the latter. Because of the presence of both scales, both kinds of simulation methods are simultaneously needed but these techniques do not lend themselves easily to coupling.

The scattering of an acoustic wave by two sound-hard obstacles: a large obstacle subject to high-frequency regime relatively to the wavelength and a small one subject to medium-frequency regime has been investigated by Marc Lenoir, Eric Lunéville and Laure Pesudo. The technique proposed in this case consists in an iterative method which allows to decouple the two obstacles and to use Geometric Optics or Physical Optics for the large obstacle and Boundary Element Method for the small obstacle. This approach has been validated on various situations using the XLife++ library developed in the lab. When the obstacles are not sticked, even if they are very close, the iterative method coupling BEM and some high-frequency methods (ray approximation or Kirchoff approximation) works very well. When the obstacle are sticked, the "natural" iterative method is no longer convergent. We are currently looking for some improved methods to deal with these cases that have a practical interest.

7.2.4. The eddy current model as a low-frequency, high-conductivity asymptotic form of the Maxwell transmission problem

Participant: Marc Bonnet.

In this work, done in collaboration with Edouard Demaldent (CEA LIST), we study the relationship between the Maxwell and eddy current (EC) models for three-dimensional configurations involving highly-conducting bounded bodies in air and sources placed remotely from those bodies. Such configurations typically occur in the numerical simulation of eddy current non destructive testing (ECT). The underlying Maxwell transmission problem is formulated using boundary integral formulations of PMCHWT type. In this context, we derive and rigorously justify an asymptotic expansion of the Maxwell integral problem with respect to the non-dimensional parameter $\gamma := \sqrt{\omega\varepsilon_0/\sigma}$. The EC integral problem is shown to constitute the limiting form of the

Maxwell integral problem as $\gamma \to 0$, i.e. as its low-frequency and high-conductivity limit. Estimates in γ are obtained for the solution remainders (in terms of the surface currents, which are the primary unknowns of the PMCHWT problem, and the electromagnetic fields) and the impedance variation measured at the extremities of the excitating coil. In particular, the leading and remainder orders in γ of the surface currents are found to depend on the current component (electric or magnetic, charge-free or not). Three-dimensional illustrative numerical simulations corroborate these theoretical findings.

7.2.5. Modelling the fluid-structure coupling caused by a far-field underwater explosion

Participants: Marc Bonnet, Stéphanie Chaillat, Damien Mavaleix-Marchessoux.

This work, funded by Naval Group and a CIFRE PhD grant, addresses the computational modelling of the mechanical effect on ships of remote underwater explosions. We aim at a comprehensive modelling approach that accounts for the effect of the initial (fast) wave impinging the ship as well as that of later, slower, water motions. Both fluid motion regimes are treated by boundary element methods (respectively for the wave and potential flow models), while the structure is modelled using finite elements. To cater for large and geometrically complex structures, the BEM-FEM interface requires large numbers of DOFs, which entails the use of a fast BEM solver. Accordingly, the wave-like fluid motions are to be computed by means of the convolution quadrature method (CQM) implemented in the in-house fast BEM code COFFEE. This work is in progress (the thesis having started in Dec. 2017). Work accomplished so far has mainly consisted in (a) thoroughly examinating the physical modelling issues, (b) formulating the mathematical and computational model that takes relevant physical features into account, and (c) implementing and assessing the CQM under conditions similar to those of the aimed application.

7.3. Domain decomposition methods

7.3.1. Transparent boundary conditions with overlap in unbounded anisotropic media

Participants: Anne-Sophie Bonnet Ben-Dhia, Sonia Fliss, Yohanes Tjandrawidjaja.

This work is done in the framework of the PhD of Yohanes Tjandrawidjaja (funded by CEA-LIST), in collaboration with Vahan Baronian (CEA). This follows the PhD of Antoine Tonnoir (now Assistant Professor at Insa of Rouen) who developed a new approach, the Half-Space Matching Method, to solve scattering problems in 2D unbounded anisotropic media. The objective is to extend the method to a 3D plate of finite width.

In 2D, our approach consists in coupling several plane-waves representations of the solution in half-spaces surrounding the defect with a FE computation of the solution around the defect. The difficulty is to ensure that all these representations match, in particular in the infinite intersections of the half-spaces. It leads to a formulation which couples, via integral operators, the solution in a bounded domain including the defect and some traces of the solution on the edges of the half-planes. We have proven that, in presence of dissipation, this system is a Fredholm equation of the second kind, in an L2 functional framework. The truncation of the Fourier integrals and the finite element approximation of the corresponding numerical method have been also analyzed.

The method has been extended to the 3D case, for an application to non-destructive testing. The objective is to simulate the interaction of Lamb waves with a defect in an anisotropic elastic plate. The additional complexity compared to the 2D case lies in the representations which are obtained semi-analytically by decomposition on Lamb modes. In addition, the system of equations couples the FE representation in the bounded perturbed domain with not only the displacement, but also the normal stress of the solution on the infinite bands limiting the half-plates. A first numerical result has been obtained in the isotropic case.

The perspectives now concern the efficiency of the method (which could be improved by replacing the direct inversion by a preconditioned iterative inversion with an efficient product matrix-vector), the analysis of the method in the case without dissipation and the analysis of the method in the elastic case.

7.3.2. Coupling BEMs in overlapping domains when a global Green's function is not available

Participants: Anne-Sophie Bonnet Ben-Dhia, Stéphanie Chaillat, Sonia Fliss, Yohanes Tjandrawidjaja.

We consider in this work problems for which the Green's function is not available, so that classical Boundary Integral equation methods are not applicable. Let us mention for instance the junction of two different stratified media (tapered optical fibers in integrated optics or junction of two topographic elastic surfaces in geophysics).

To this end, we propose a generalization of the Half-Space Matching method (see section 7.3.1).

In this work, by replacing the Fourier representations by integral representations, we are able to replace the half-spaces by more general unbounded overlapping sub-domains. We choose the sub-domains in such a way that an explicit Green's function is available for each subdomain. For instance, for the configuration described above (figure 1 a), it suffices to introduce two infinite sub domains, each of them containing only one stratification (figures 1 c and 1 d) and a bounded domain containing the junction (figure 1 b). The formulation couples the solution in the bounded domain with the single and double layer potentials on each boundary of the sub-domains. The approximation relies on a FE discretisation of the volume unknown and a truncation and a discretization of the boundary/surface unknowns.

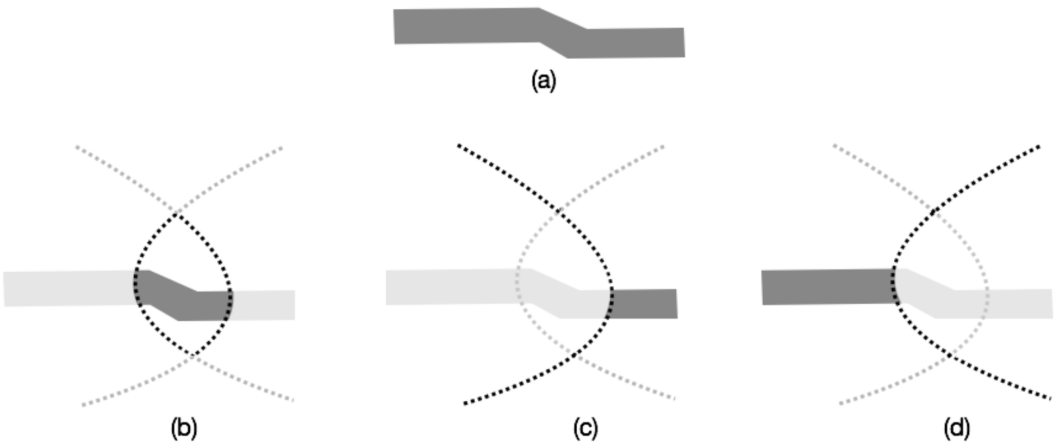

Figure 1. Coupling BEMs in overlapping domains

A study concerning the choice of the discretisation parameters and the shape of the infinite lines have to be done. The theoretical analysis of the method raises challenging open questions: for instance, a first uniqueness result has been derived, which requires the definition of a variational formulation on a Rie. Finally, we want to apply the method to the scattering by a step, i.e. the junction of two semi infinite-planes joined together by a step.

7.3.3. Domain decomposition method for acoustics with uniform exponential rate of convergence using non-local impedance operators

Participants: Patrick Joly, Francis Collino, Émile Parolin.

This work is done in the framework of the PhD of Émile Parolin (funded by ANR NonlocalDD), in collaboration with X. Claeys (EPI Alpines & LJLL).

We continued the work on non-overlapping domain decomposition methods with non-local transmission conditions for time-harmonic wave propagation. The analysis of such methods is conducted by writing them as a relaxed Jacobi algorithm. In the absence of junctions points, the continuous algorithm converges exponentially fast under suitable assumptions on the impedance operators. These assumptions cannot be satisfied using local operators and rely in practice on singular integral operators. The progress achieved is as follows.

- In the context of acoustic wave propagation, we established a new result on the robustness of the algorithm with respect to the mesh size. We have proven that for Lagrange finite element approximations the exponential rate of convergence of the algorithm is independent of the discretization parameter, hence does not deteriorate when the mesh is refined. The proof relies on the Scott-Zhang interpolator and led to the submission.
- We have been working on the extension to 3D time harmonic Maxwell's equations. The main difficulty is to design well adapted operators taking into account the specificity of the corresponding trace spaces. An adequate operator must behave like a pseudo-differential operator with opposite order on the 'curl part' and 'grad part' of a tangential field (this is related to the Helmholtz decomposition of tangential fields). Guided by potential theory for elliptic operators, we proposed two classes of suitable operators. The first one is based on Bessel potentials (fractional powers of the shifted Laplacian) and the second one relies on Riesz potentials. We have shown that the proposed operators satisfy the desired properties in the case of a sphere using modal analysis techniques. We have also been working on the design of the finite element approximation of these operators.

7.3.4. Domain decomposition method with cross-point treatment for high-frequency acoustic scattering

Participant: Axel Modave.

This work is done in collaboration with X. Antoine (IECL & EPI SPHINX) and C. Geuzaine (Université de Liège).

Solving high-frequency time-harmonic scattering problems using FE techniques is challenging, as such problems lead to very large, complex and indefinite linear systems. Optimized Schwarz domain decomposition methods (DDMs) are currently a very promising approach, where subproblems of smaller sizes are solved in parallel using direct solvers, and are combined in an iterative procedure. It is well-known that the convergence rate of these methods strongly depends on the transmission condition enforced on the interfaces between the subdomains.

Local transmission conditions based on high-order absorbing boundary conditions (HABCs) have proved well suited. They represent a good compromise between basic impedance conditions (which lead to suboptimal convergence) and the exact Dirichlet-to-Neumann (DtN) map related to the complementary of the subdomain (which is expensive to compute). However, a direct application of this approach for domain decomposition configurations with cross-points, where more than two subdomains meet, does not provide satisfactory results.

We work on improved DDMs that efficiently addresses configurations with cross-points. Noting that these points actually are corners for the subdomains, our strategy consists in incorporating a corner treatment developed for HABCs (see section 7.7.1) into the DDM procedure. We propose a cross-point treatment for HABC-based DDMs in settings with cross-points and right angles. The method is implemented and successfully tested for two-dimensional examples. The analysis of this method is currently in progress. Extensions to more complicated settings (e.g. 3D, with non-right angles, other physical waves) will be investigated in the future.

7.4. Wave propagation in complex media

7.4.1. Enriched Homogenization in presence of boundaries or interfaces

Participants: Clement Beneteau, Sonia Fliss.

This work is done in the framework of the PhD of Clement Beneteau and is done in collaboration with X. Claeys (Sorbonne & EPI Alpines).

This work is motivated by the fact that classical homogenization theory poorly takes into account interfaces or boundaries. It is particularly unfortunate when one is interested in phenomena arising at the interfaces or the boundaries of the periodic media (the propagation of plasmonic waves at the surface of metamaterials for instance). To overcome this limitation, we have constructed an effective model which is enriched near the interfaces and/or the boundaries. For now, we have treated and analysed the case of simple geometries: for instance a half-plane with Dirichlet or Neumann boundary conditions or a plane interface between two periodic half spaces. We have derived a high order approximate model which consists in replacing the periodic media by an effective one but the boundary/transmission conditions are not classical. The obtained conditions involve Laplace- Beltrami operators at the interface and requires to solve cell problems in periodicity cell (as in classical homogenization) and in infinite strips (to take into account the phenomena near the boundary/interface). We establish well posedness for the approximate model and error estimates which justify that this new model is more accurate. From a numerical point of view, the only difficulty comes from the problems set in infinite strips. The method has been implemented using Xlife++.

This approach has been extended to the long time homogenisation of the wave equation. It is well known that the classical effective homogenized wave equation does not capture the long time dispersive effects of the waves in the periodic media. Since the works of Santosa and Symes in the 90's, several effective equations (involving differential operators of order at least 4) that capture these dispersive effects have been proposed, but only in infinite media. In presence of boundaries or interfaces, the question of boundary/transmission conditions for these effective equations was never treated. We have first results in that direction.

7.4.2. Transmission conditions between homogeneous medium and periodic cavities

Participant: Jean-François Mercier.

In collaboration with A. Maurel (Langevin Institute), J. J. Marigo (LMS) and K. Pham (Imsia).

We have developed a model for resonant arrays of Helmholtz cavities, thanks to a two scale asymptotic analysis. The model combines volumic homogenization to replace the cavity region by a homogeneous anisotropic slab and interface homogenization to replace the region of the necks by transmission conditions. The coefficients entering in the effective wave equation are simply related to the fraction of air in the periodic cell of the array. Those involved in the jump conditions encapsulate the effects of the neck geometry.

In parallel, this effective model has been exploited to study the resonance of the Helmholtz resonators with a focus on the influence of the neck shape. The homogenization makes a parameter B to appear which determines unambiguously the resonance frequency of any neck. As expected, this parameter depends on the length and on the minimum opening of the neck, and it is shown to depend also on the surface of air inside the neck. Once these three geometrical parameters are known, B has an additional but weak dependence on the neck shape, with explicit bounds.

7.4.3. Mathematical analysis of wave propagation in fractal trees

Participants: Patrick Joly, Maryna Kachanovska.

We have continued our work (in collaboration with A. Semin (TU Darmstadt)) on wave propagation in fractal trees which model human lungs. One of the major results of this year is a complete analysis of such models. In particular, provided Sobolev spaces H^1_μ, L^2_μ (which generalize weighted Sobolev spaces on an interval to the case of fractal trees) we clarified the following questions for a range of parameters of the trees not covered by the previous theory: existence of traces of H^1_μ-functions on fractal trees; approximation of H^1_μ-functions by compactly supported functions; compact embedding of H^1_μ into L^2_μ.

7.4.4. Hyperbolic Metamaterials in Frequency Domain: Free Space

Participants: Patrick Ciarlet, Maryna Kachanovska.

In this project we consider the wave propagation in 2D hyperbolic metamaterials [Poddubny et al., Nature Photonics, 2013], which are modelled by Maxwell equations with a diagonal frequency-dependent tensor of dielectric permittivity ε and scalar frequency-independent magnetic permeability. In the time domain, the corresponding models are well-posed and stable. Surprisingly, in some regimes in the frequency domain, when the signs of the diagonal entries of ε do not coincide, the problem becomes hyperbolic (and hence the name). The main goal of this project is to justify the well-posedness of such models in the frequency domain, first of all starting with the case of the free space. We have obtained partial results in this direction: radiation condition, which ensures the well-posedness of the problem, mapping properties of the resolvent (with refined estimates on the propagation of singularities in these models). We are currently working on the limiting absorption and limiting amplitude principles.

7.5. Spectral theory and modal approaches for waveguides

7.5.1. Scattering solutions in an unbounded strip governed by a plate model

Participants: Laurent Bourgeois, Sonia Fliss.

Together with Lucas Chesnel (EPI DEFI), we have initiated a new work on a particular waveguide which consists of a thin strip governed by a Kirchhoff-Love bilaplacian model. The aim is to build some radiation conditions and prove well-posedness of scattering problems for that simple model and for two kinds of boundary conditions: the strip is either simply supported or clamped. In the first case, we have shown that using a Dirichlet-to-Neumann operator enables us to prove fredholmness. Such approach is not possible in the second case, for which a completely different angle of attack is chosen: a Kondratiev approach involving weighted Sobolev spaces and detached asymptotics.

7.5.2. Modal analysis of electromagnetic dispersive media

Participants: Christophe Hazard, Sandrine Paolantoni.

We investigate the spectral effects of an interface between vacuum and a negative material (NM), that is, a dispersive material whose electric permittivity and/or magnetic permeability become negative in some frequency range. Our first work in this context concerns an elementary situation, namely, a two-dimensional scalar model (derived from the complete Maxwell's equations) which involves the simplest existing model of NM, referred to as the non-dissipative Drude model (for which negativity occurs at low frequencies). By considering a polygonal cavity, we have shown that the presence of the Drude material gives rise to various components of an essential spectrum corresponding to various unusual resonance phenomena: first, a low frequency bulk resonance (accumulation at the zero frequency of positive eigenvalues whose associated eigenvectors are confined in the Drude material); then, a surface resonance for one particular critical frequency (at which the so-called surface plasmons occurs, that is, localized highly oscillating vibrations at the interface between the Drude material and the vacuum); finally, corner resonances in a critical frequency interval (here, localized highly oscillating vibrations occur near any corner of the interface, interpreted as a "black hole" phenomenon). An article which presents these results has been submitted. Most recent works were devoted to the numerical simulation of these resonance phenomena in the context of the code XLiFE++ developped in the lab.

7.5.3. Formulation of invisibility in waveguides as an eigenvalue problem

Participant: Anne-Sophie Bonnet-Ben Dhia.

This work is done in collaboration with Lucas Chesnel from EPI DEFI and Vincent Pagneux from Laboratoire d'Acoustique de l'Université du Maine.

We consider an infinite acoustic waveguide (with a bounded cross-section) which is locally perturbed. At some exceptional frequencies and for particular incident waves, it may occur that all the energy of the incident wave is transmitted, the only effect in reflection being a superposition of evanescent modes in the vicinity of the perturbation. We have proposed an approach for which these reflection-less frequencies appear directly as eigenvalues of a new problem. This problem is very similar to the formulation of the scattering problem using Perfectly Matched Layers, except a slight modification in the PML. Precisely, we use two conjugated dilation parameters, α in the outlet and $\overline{\alpha}$ in the inlet, in order to select outgoing waves in the outlet and ingoing waves in the inlet. In fact, we show that the real eigenfrequencies that are obtained correspond either to trapped modes or to reflection-less modes. In addition to this real spectrum, we find intrinsic complex frequencies, which also contain information about the quality of the transmission through the waveguide. Mathematically, the non-selfadjoint eigenvalue problem with conjugated PMLs has strange properties: the discreteness of the point spectrum is not stable by compact perturbations and pathological examples can be exhibited.

7.6. Inverse problems

7.6.1. Linear Sampling Method with realistic data in waveguides

Participants: Laurent Bourgeois, Arnaud Recoquillay.

Our activities in the field of inverse scattering in waveguides with the help of sampling methods has now a quite long history. Very recently, we have focused on elastodynamics and realistic data, that is surface data in the time domain. This has been the subject of the PhD of Arnaud Recoquillay. It was motivated by Non Destructive Testing activities for tubular structures and was the object of a partnership with CEA List (Vahan Baronian).

Our strategy consists in transforming the time domain problem into a multi-frequency problem by the Fourier transform. This allows us to take full advantage of the established efficiency of modal frequency-domain sampling methods. In particular, we have shown how to optimize the number of sources/receivers and the distance between them in order to obtain the best possible imaging results.

Our main achievement is an experimental validation of such approach in the presence of real data: the measurements were carried at CEA on steel plates with the help of piezoelectric sensors. The identification results are encouraging and pave the way of a future integration of sampling methods in real NDT activities.

7.6.2. The "exterior approach" to solve inverse obstacle problems

Participants: Laurent Bourgeois, Arnaud Recoquillay, Dmitry Ponomarev.

This work is done in collaboration with Jérémi Dardé (IMT Toulouse).

We consider some inverse obstacle problems in acoustics by using a single incident wave, either in the frequency or in the time domain. When so few data are available, a Linear Sampling type method cannot be applied. In order to solve those kinds of problem, we propose an "exterior approach", coupling a mixed formulation of quasi-reversibility and a simple level set method. In such iterative approach, for a given defect D, we update the solution u with the help of a mixed formulation of quasi-reversibility while for a given solution u, we update the defect D with the help of a level set method based on a Poisson problem. We have studied two cases. The first case concerns the waveguide geometry in the frequency domain. The second case concerns a bounded spatial set in the time domain when data are given in a finite time interval. This last case is challenging because it raises the (open) question of the minimal final time which is required to ensure uniqueness of the obstacle from the lateral Cauchy data.

7.6.3. Inverse acoustic scattering using high-order small-inclusion expansion of misfit function

Participant: Marc Bonnet.

This work concerns an extension of the topological derivative concept for 3D inverse acoustic scattering problems involving the identification of penetrable obstacles, whereby the featured data-misfit cost function J is expanded in powers of the characteristic radius a of a single small inhomogeneity. The $O(a^6)$ approximation of J is derived and justified for a single obstacle of given location, shape and material properties embedded in a 3D acoustic medium of arbitrary shape, and the generalization to multiple small obstacles is outlined. Simpler and more explicit expressions are obtained when the scatterer is centrally-symmetric or spherical. An approximate and computationally light global search procedure, where the location and size of the unknown object are estimated by minimizing the $O(a^6)$ approximation over a search grid, is proposed and demonstrated on numerical experiments, where the identification from known acoustic pressure on the surface of a penetrable scatterer embedded in a acoustic semi-infinite medium, and whose shape may differ from that of the trial obstacle assumed in the expansion of J, is considered. measurements configuration situated far enough from the probing region.

7.6.4. Microstructural topological sensitivities of the second-order macroscopic model for waves in periodic media

Participant: Marc Bonnet.

This work is done in collaboration with Bojan Guzina (University of Minnesota, USA) and Rémi Cornaggia (IRMAR, Rennes).

We consider scalar waves in periodic media through the lens of a second-order effective i.e. macroscopic description, and we aim to compute the sensitivities of the relevant effective parameters due to topological perturbations of a microscopic unit cell. Specifically, our analysis focuses on the tensorial coefficients in the governing mean-field equation – including both the leading order (i.e. quasi-static) terms, and their second-order counterparts. The results demonstrate that the sought sensitivities are computable in terms of (i) three unit-cell solutions used to formulate the unperturbed macroscopic model; (ii) two adoint-field solutions driven by the mass density variation inside the unperturbed unit cell; and (iii) the usual polarization tensor, appearing in the related studies of non-periodic media, that synthesizes the geometric and constitutive features of a point-like perturbation. The proposed developments may be useful toward (a) the design of periodic media to manipulate macroscopic waves via the microstructure-generated effects of dispersion and anisotropy, and (b) sub-wavelength sensing of periodic defects or perturbations.

7.6.5. Analysis of topological derivative as a tool for qualitative identification

Participant: Marc Bonnet.

This work is a collaboration with Fioralba Cakoni (Rutgers University, USA).

The concept of topological derivative has proved effective as a qualitative inversion tool for a wave-based identification of finite-sized objects. Although for the most part, this approach remains based on a heuristic interpretation of the topological derivative, a first attempt toward its mathematical justification was done in Bellis et al. (Inverse Problems 29:075012, 2013) for the case of isotropic media with far field data and inhomogeneous refraction index. Our paper extends the analysis there to the case of anisotropic scatterers and background with near field data. Topological derivative-based imaging functional is analyzed using a suitable factorization of the near fields, which became achievable thanks to a new volume integral formulation recently obtained in Bonnet (J. Integral Equ. Appl. 29:271-295, 2017). Our results include justification of sign heuristics for the topological derivative in the isotropic case with jump in the main operator and for some cases of anisotropic media, as well as verifying its decaying property in the isotropic case with near field spherical measurements configuration situated far enough from the probing region.

7.6.6. Elasticity imaging by error in constitutive equation functionals

Participant: Marc Bonnet.

This work is done in collaboration with Wilkins Aquino (Duke University, USA).

We formulate the identification of heterogeneous linear elastic moduli in the context of time-harmonic elastodynamics as the minimization of the modified error in constitutive equation (MECE) functional. Our main goal is to develop theoretical foundations, in a continuous setting, allowing to explain and justify some known beneficial properties of this treatment. A specific feature of MECE formulations is that forward and adjoint solutions are governed by a fully coupled system, whose mathematical properties play a fundamental role in the qualitative and computational aspects of MECE minimization. We prove that this system has a unique and stable solution at any frequency, provided data is abundant enough (in a sense made precise), even though the relevant forward problem is not *a priori* clearly defined. This result has practical implications such as applicability of MECE to partial interior data (with important practical applications including ultrasound elastography), convergence of finite element discretizations and differentiability of the reduced MECE functional. In addition, we establish that usual least squares and pure ECE formulations are limiting cases of MECE formulations for small and large values of the weight of the data misfit component of the functional, respectively. For the latter case, we furthermore show that the reduced MECE Hessian is asymptotically positive for any parameter perturbation supported on the measurement region, thereby corroborating existing computational evidence on convexity improvement brought by MECE functionals. Finally, numerical studies including parameter reconstruction examples using interior data support our findings.

7.6.7. A continuation method for building large invisible obstacles in waveguides

Participants: Antoine Bera, Anne-Sophie Bonnet-Ben Dhia.

This work is done on collaboration with Lucas Chesnel (EPI DEFI).

We are interested in building invisible obstacles in waveguides, at a given frequency. The invisibility is characterized by the nullity of the scattering coefficients associated to propagating modes. In previous papers, a method has been proposed to prove the existence of invisible obstacles and to build them. But its main drawback was its limitation to small obstacles. In order to get larger invisible obstacles, we have developed a new approach which combines the previous idea with a continuation method: we are building a sequence of invisible obstacles, each of them being a small perturbation of the previous one. This algorithm is based, at each step, on the ontoness of an application and on the fixed-point theorem. We have implemented the method in the finite element library XLiFE++, in the case of penetrable obstacles of a two-dimensional acoustic waveguide, in multi-modal regime. A remarkable result is that the ontoness condition can be ensured in many cases, so that the algorithm can be iterated as long as required. Another interesting feature of our approach is that it allows to prescribe some properties of the obstacle (shape of the obstacle, piecewise constant index, ...), but a drawback is that the algorithm can produce non-realistic negative indices. This is a question that we are currently working on. Finally, let us emphasize that the formalism of the method is very general and flexible. In particular, it can be directly extended to 3D waveguides, or to the scattering in free space.

7.7. Acoustics and aeroacoustics

7.7.1. High-order absorbing boundary conditions with corner treatment for high-frequency acoustic scattering

Participant: Axel Modave.

This work is done in collaboration with C. Geuzaine (University of Liège) and X. Antoine (IECL & EPI SPHINX)

We address the design and validation of accurate local absorbing boundary conditions set on convex polygonal computational domains for the finite element solution of high-frequency acoustic scattering problems. While high-order absorbing boundary conditions (HABCs) are accurate for smooth fictitious boundaries, the precision of the solution drops in the presence of corners if no specific treatment is applied. We analyze two strategies to preserve the accuracy of Padé-type HABCs at corners: first by using compatibility relations (derived for right angle corners) and second by regularizing the boundary at the corner. We show that the former strategy is well-adapted to right corners and efficient for nearly-right corners, while the later is better for very obtuse corners. Numerical results are proposed to analyze and compare the approaches for two-and three-dimensional problems.

7.7.2. Time-harmonic acoustic scattering in a vortical flow

Participants: Antoine Bensalah, Patrick Joly, Jean-François Mercier.

We study the time-harmonic acoustic radiation in a fluid in flow. To go beyond the convected Helmholtz equation, only adapted to potential flows, we use Goldstein's equations, coupling exactly the acoustic waves to the hydrodynamic field. We have studied the hydrodynamic part of Goldstein equations, corresponding to a generalized time-harmonic transport equation and we have investigated its well-posedness. The result has been established under the assumption of a domain-filling flow, which in 2D is simply equivalent to a flow that does not vanish. The approach relies on the method of characteristics, which leads to the resolution of the transport equation along the streamlines and on general results of functional analysis. The theoretical results have been illustrated with numerical results obtained with a SUPG Finite Element scheme.

In complement we have developed a new model for Goldstein's equations in which the description of the hydrodynamic phenomena is simplified. The model, initially developed for a carrier flow of low Mach number M, is proved theoretically to remain accurate for moderate Mach numbers, associated to a low error bounded by M^2. Numerical experiments confirm the M^2 law and the good quality of the model for flows of non-small Mach numbers.

7.8. Numerical analysis for PDEs

7.8.1. A family of Crouzeix-Raviart Finite Elements in 3D

Participant: Patrick Ciarlet.

This work is done in collaboration with C. Dunkl (University of Virginia) and S. Sauter (Universität Zürich).

We develop a family of non-conforming "Crouzeix–Raviart" type finite elements in three dimensions. They consist of local polynomials of maximal degree p on simplicial finite element meshes while certain jump conditions are imposed across adjacent simplices. We will prove optimal a priori estimates for these finite elements. The characterization of this space via jump conditions is implicit and the derivation of a local basis requires some deeper theoretical tools from orthogonal polynomials on triangles and their representation. We will derive these tools for this purpose. These results allow us to give explicit representations of the local basis functions. Finally, we will analyze the linear independence of these sets of functions and discuss the question whether they span the whole non-conforming space.

7.8.2. Numerical analysis of the mixed finite element method for the neutron diffusion eigenproblem with heterogeneous coefficients

Participants: Patrick Ciarlet, Léandre Giret, Félix Kpadonou.

This work is done in collaboration with E. Jamelot (CEA).

We study first the convergence of the finite element approximation of the mixed diffusion equations with a source term, in the case where the solution is of low regularity. Such a situation commonly arises in the presence of three or more intersecting material components with different characteristics. Then we focus on the approximation of the associated eigenvalue problem. We prove spectral correctness for this problem in the mixed setting. These studies are carried out without, and then with a domain decomposition method. The domain decomposition method can be non-matching in the sense that the traces of the finite element spaces may not fit at the interface between subdomains. Finally, numerical experiments illustrate the accuracy of the method.

7.8.3. Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients

Participant: Patrick Ciarlet.

This work is done in collaboration with M. Vohralik (EPI SERENA).

We present a posteriori error analysis of diffusion problems where the diffusion tensor is not necessarily symmetric and positive definite and can in particular change its sign. We first identify the correct intrinsic error norm for such problems, covering both conforming and nonconforming approximations. It combines a dual (residual) norm together with the distance to the correct functional space. Importantly, we show the equivalence of both these quantities defined globally over the entire computational domain with the Hilbertian sums of their localizations over patches of elements. In this framework, we then design a posteriori estimators which deliver simultaneously guaranteed error upper bound, global and local error lower bounds, and robustness with respect to the (sign-changing) diffusion tensor. Robustness with respect to the approximation polynomial degree is achieved as well. The estimators are given in a unified setting covering at once conforming, nonconforming, mixed, and discontinuous Galerkin finite element discretizations in two or three space dimensions. Numerical results illustrate the theoretical developments.

7.8.4. On the convergence in H^1-norm for the fractional Laplacian

Participant: Patrick Ciarlet.

This work is done in collaboration with J.P. Borthagaray (University of Maryland).

We consider the numerical solution of the fractional Laplacian of index $s \in (1/2, 1)$ in a bounded domain Ω with homogeneous boundary conditions. Its solution a priori belongs to the fractional order Sobolev space $\widetilde{H}^s(\Omega)$. For the Dirichlet problem and under suitable assumptions on the data, it can be shown that its solution is also in $H^1(\Omega)$. In this case, if one uses the standard Lagrange finite element to discretize the problem, then both the exact and the computed solution belong to $H^1(\Omega)$. A natural question is then whether one can obtain error estimates in $H^1(\Omega)$-norm, in addition to the classical ones that can be derived in the $\widetilde{H}^s(\Omega)$ energy norm. We address this issue, and in particular we derive error estimates for the Lagrange finite element solutions on both quasi-uniform and graded meshes.

RESIST Team

7. New Results

7.1. Monitoring

7.1.1. HTTPS traffic monitoring

Participants: Jérôme François [contact], Pierre-Olivier Brissaud, Olivier Bettan [Thales], Isabelle Chrisment, Thibault Cholez.

While privacy is empowered by encrypted communications such as through the HTTPS protocol, it is also legitimated to allow network monitoring of HTTPS traffic. To be compliant with privacy, we proposed a transparent and passive technique that only detects if an HTTPS request is related to a previously defined action [7]. Our technique is able to detect forbidden searches over a web service such as Google Images. It differs from related work that either focuses on detecting the type of traffic or the used web service. To achieve a high accuracy, our technique relies on learning stage where keywords to be monitored are crawled before we leverage KDE (Kernel Density Estimation). KDE allows us to construct a signature summarizing the sizes of the loaded objects on a page, which strongly depend on the user action or search.

7.1.2. Monitoring Programmable Networks

Participants: Jérôme François [contact], Olivier Festor, Paul Chaignon [Orange Labs], Kahina Lazri [Orange Labs], Thibault Delmas [Orange Labs].

SDN-based monitoring allows us to gather more valuable indicators by specifying or programming the monitoring with a fine granularity. We proposed to use eBPF (extended Berkeley Packet Filter) to apply fine-grained filtering in comparison to OpenFlow. It brings safety guarantees regarding program execution and allows stateful programs. In order to limit the impact on the throughput, we integrated our solution within the regular packet processing pipeline of Open vSwitch, a major software switch for OpenFlow, by extending the cache mechanisms [8].

7.1.3. 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 the last decade, many IoT architectures have been proposed. However, security cannot be guaranteed without failure or by-design. In that context, we are currently investigating predictive security monitoring strategies for large-scale Internet-of-Things. In particular, we are considering the building of behavioral models characterizing such complex networks. The objective is to support both the detection of malicious activities, as well as the selection of security counter-measures.

7.1.4. Quality of Experience Monitoring

Participants: Isabelle Chrisment [contact], Thibault Cholez, Antoine Chemardin, Vassili Rivron [University of Caen], Lakhdar Meftah [University of Lille].

We have pursued our work on smartphone usage monitoring with the SPIRALS team (Inria/Université de Lille) and more specifically on proposing new methods to help measure the QoE and to protect the user's privacy when collecting such data.

In the context of the BottleNet project, to build an adequate instrumented investigation system (mobile applications combining measurements and questionnaires), we decomposed, with a group of students, the network quality concept and the perception of the services in several different approaches. These students worked on bibliographic research, on the smartphone usage and on the perception of the Internet. Structured debates on social issues associated with mobile connectivity were organized. The following topics were dealt: Quality of Service/Quality of Experience; rhythms of life and routines; privacy: diversity of practices and ethical issues; advertising and free: volume, exposure, perception, third-party and cost; quantified self-*: relation to self-quantification; online cultural consumption; information practices on mobile; communication practices.

In the context of the IPL BetterNet project, we continued to work on federating Inria's monitoring tools (APISENSE®, Fathom, Hostview, ACQUA) in a common measurement platform. A first test campaign has been performed with a small set of volunteer users to evaluate the full data collection system built from all these tools.

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: Florent Didier, Alexandre Merlin, Lucas Nussbaum [contact], Olivier Demengeon [SED], Teddy Valette [SED].

The team was again heavily involved in the evolutions and the governance of the Grid'5000 testbed.
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 (9 FTE).
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. Since 2018, SILECS has been listed as part of the French National Roadmap for Very Large Research Infrastructures (TGIR program).
Grid'5000/FIT school We had a central role in the organization of the Grid'5000/FIT school that took place in Sophia-Antipolis in April 2018, gathering 93 participants. Lucas Nussbaum delivered a keynote talk presenting Grid'5000 and its recent evolutions [28]. A successful evaluation of Grid'5000 by its Scientific Advisory Board also took place during the school.
Storage manager A contribution from the team was the design and development of a new storage access manager that allows secure access to NFS home directories, thus closing a widely-spread security vulnerability.

7.2.2. I/O emulation support in Distem

Participants: Alexandre Merlin, Olivier Dautricourt, Abdulqawi Saif, Lucas Nussbaum [contact].

Distem had a new release (version 1.3) at the beginning of 2018. This release mainly focused on bringing it up-to-date in terms of software quality (newer dependencies, added tests) and added some network emulation features that were previously missing.

The emulator was then featured in a tutorial during the Grid'5000/FIT school.

There is ongoing work on adding I/O emulation support in Distem, in order to experiment how Big Data solution can handle degraded situations. This is still pending completion and publication.

7.2.3. I/O access patterns analysis with eBPF

Participants: Abdulqawi Saif, Lucas Nussbaum [contact], Ye-Qiong Song.

We explored the relevance of an emerging instrumentation technology for the Linux kernel, eBPF, and used it to analyze I/O access patterns such as non-sequential accesses, which are particularly harmful on non-SSD drives. We designed a tool to help with such analysis, and applied it to two popular NoSQL databases, MongoDB and Cassandra, outlining severe performance problems [19] with MongoDB, where a workload that should have resulted in sequential accesses was in fact turned into lots of random accesses.

7.2.4. Experiment Monitoring

Participants: Abdulqawi Saif, Alexandre Merlin, Lucas Nussbaum [contact], Ye-Qiong Song.

Most computer experiments include a phase where metrics are gathered from and about various kinds of resources. This phase is often done via manual, non-reproducible and error-prone steps. We designed an experiment monitoring framework called MonEx, built on top of infrastructure monitoring solutions and supporting various monitoring approaches. MonEx fully integrates into the experiment workflow by encompassing all steps from data acquisition to producing publishable figures [18], [29].

7.2.5. Testbed federation and collaborations in the testbeds community

Participant: Lucas Nussbaum [contact].

The Fed4FIRE+ H2020 project started in January 2017 and will run until the end of September 2021. This project aims at consolidating the federation of testbeds in Europe of which Grid'5000 is a member. In 2018, we focused on various aspects related to experiment reproducibility.

We are also active in the GEFI initiative that aims at building links between the US testbeds community (GENI) and their European (FIRE), Japanese and Brazilian counterparts. We participated in the annual GEFI meeting where we chaired two sessions on *Experiment reproducibility* and *Networking experiments*, respectively, and gave one talk on Experiment data management, outlining the recent work that was done on Grid'5000 on disk reservation [27].

7.2.6. Blockchain experimentation

Participants: Jérôme François [Contact], Wazen Shbair [University of Luxembourg, Luxembourg], Radu State [University of Luxembourg, Luxembourg], Mathis Steichen [University of Luxembourg, Luxembourg].

The experimentation of distributed applications like blockchains needs a highly reconfigurable and controllable environment for fine-tuning blockchain and network parameters in different scenarios. Therefore, there might be significant manual operations which lead to human errors and make it hard to reproduce experiments. We proposed an easy to use orchestration framework over the Grid'5000 platform [23]. Our tool can fine-tune blockchain and network parameters before and between experiments. The proposed framework offers insights for private and consortium blockchain developers to identify performance bottlenecks and to assess the behavior of their applications in different circumstances.

7.2.7. 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 transport the traffic of regular web users over an NDN island. Users just need to configure the ingress gateway as a standard web proxy that will be the entry point to the virtualized NDN island, and their traffic is seamlessly transported over NDN, thus benefiting from the good properties of the protocol to deliver content (request mutualization, caching, etc.). HTTP requests/responses are converted into NDN Interest/Data and the answer can either come from the island, or from the web through the egress gateway. Our first functional experimental results of an initial testbed deployment exhibit the capability of our global infrastructure to retrieve the top-1000 most popular web sites without difficulty [17]. This opens the way to wider and more realistic experiments of NDN with real traffic. In particular, the gateway was used to perform QoE experiments involving real users from Nancy and Troyes. They accessed many websites through the NDN network in a very satisfying way.

7.3. Analytics

7.3.1. CPS Security analytics

Participants: Abdelkader Lahmadi [contact], Mingxiao Ma, Isabelle Chrisment.

During 2018, we designed and 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 analyzed the control-theoretic and detectability properties of this attack to assess its impact on reference voltage synchronization at the different control layers of a microgrid. Results from numerical simulation are presented in [15] and demonstrate this attack, in particular the maximum voltage deviation and inaccurate reference voltage synchronization it causes in the microgrid.

7.3.2. Analysis of Internet-wide attacks

Participants: Abdelkader Lahmadi [contact], Giulia de Santis, Jérôme François, Olivier Festor.

Internet-wide scanners are heavily used for malicious activities. In [13], we developed models based on HMMs (Hidden Markov Models) and finite mixture models to identify network scanners from the packets received by a darknet. We used data collected by the darknet hosted in the High Security Lab of Inria Nancy - Grand Est to build these models by characterizing the spatial and temporal movements of the studied scanners (Zmap and Shodan). Our models are able to recognize the scanner with an accuracy of 95% when using spatial movements, and of 98% when using temporal movements.

Under the umbrella of the ThreatPredict project with the International University of Rabat, we have performed preliminary exploratory analysis of Inria darknet data that consists of examining time series of scan activities and the scanning behavior of different attackers [24]. We performed experiments on the clustering of darknet data to extract threat patterns including scanning and DDoS activities. We are still extending the technique with more features and developing Hololens based visualization of the obtained graphs. Based on our experience, traffic analysis faces a major challenge when using machine learning or data-mining techniques due to data which cannot be represented in a meaningful metric space. One major case is TCP or UDP ports. We thus proposed a new semantic based metric between port numbers that does not follow a regular numeric distance but relies on observed attacks of the past.

7.3.3. Cyber Threat Intelligence

Participants: Jérôme François, Abdelkader Lahmadi [contact], Quang Vinh Dang.

We are exploring and validating techniques for learning correlations between vulnerabilities and attack patterns from Cyber threat intelligence data sources including CVE (Common Vulnerabilities and Exposures), CAPEC (Common Attack Pattern Enumeration and Classification) and CWE (Common Weaknesses Enumeration) documents. While there already exist some relations between them, they have been defined manually and so are quite incomplete. Finding these relations is a cumbersome and tedious task and our objective is to guide or even automatically detect relations or correlations between documents. This will ease a better understanding and mitigation of threats. Our work relies on leveraging NLP (Natural Language Processing Techniques) with several techniques such as graph-based or recommendation-based mining. The first results show the ability of our technique to automatically discover missing relations between attack patterns and vulnerability descriptions in the context of SDN [12]. We also consider word and document embedding to identify correlations between them.

7.4. Orchestration

7.4.1. Programming of network functions

Participants: Thibault Cholez [contact], Diane Adjavon [Orange Labs], Anthony Anthony, Raouf Boutaba, Paul Chaignon, Shihabur Rahman Chowdhury, Olivier Festor, Jérôme François, Kahina Lazri [Orange Labs], Xavier Marchal.

NFV is a key technology for the successful deployment of new network protocol stacks like Named Data Networking (NDN). Instead of trying to oddly couple IP and new Information-Centric Networking protocols, one should rather deploy them in different network slices and ensure their isolation. We proposed a complete NFV architecture composed of several Virtual Network Functions (VNF) designed for NDN and orchestrated so that they can dynamically adapt the topology to react against issues such as an ongoing attack [25].

To push even further the possibilities of NFV, we applied the microservice architecture inherited from the software world to design atomic and flexible functions that must be combined to process NDN traffic. The proposed architecture, described in μNDN [16], includes seven orchestrated microservices. Some of them are components extracted from the monolithic and heavy-burden NDN router while others are new on-path functions that can perform specific processing on the traffic like a signature-verification module or a name-filtering module. The evaluation through two realistic scenarios proved the ability of our manager to dynamically scale-up bottleneck functions and mitigate ongoing attacks on the NDN network. We also refined our countermeasure against information leakage attacks in NDN [4].

In [8], we proposed to offload part of the processing of VNF to the programmable switches. The problem resides in guaranteeing a fair scheduling at the switch level assuming the required run-to-completion execution. We thus defined a token-based scheduling approach. In [6], we defined a new scheduler for VNFs that integrates a CPU cycle estimator and a heuristic to avoid wasting idle CPU cycles.

7.4.2. Software-defined security for clouds

Participants: Rémi Badonnel [contact], Olivier Festor, Maxime Compastié, He Ruan [Orange Labs].

We have pursued our work on a software-defined security framework for enabling the enforcement of security policies in distributed clouds. This framework aims at dynamically integrating and configuring security mechanisms for protecting cloud services that are distributed over multi-cloud and multi-tenant environments. In that context, we have described in [11] generation mechanisms for building protected cloud resources based on unikernels in an on-the-fly manner. These unikernels integrate security mechanisms at an early stage, and are characterized by highly-constrained configurations, in order to reduce the attack surface. A demonstration of this work has been showcased during the IFIP/IEEE NOMS 2018 international conference [10]. We have also investigated the exploitation of the TOSCA orchestration language to drive the generation of these unikernels. This language supports the specification of cloud services in the form of topologies and their orchestrations. The objective was to extend this language to both describe the generation of unikernel ressources, and specify different levels of security to be orchestrated. We have designed a framework to interpret this extended language, and to generate and configure protected resources according to these levels. We have evaluated the performance of generation mechanisms through extensive experiments. This generation can be performed in a proactive manner with respect to security levels, in accordance with elasticity and on-demand cloud properties.

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 elaboraboration 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. In particular, we have formalized and extended our approach for generating SDN policies to protect Android applications [21], [22]. We have introduced a system based on inference rules for automating the generation of such chains [20], taking into account both their networking behavior and the OS-level permissions that they request. By using first-order predicates for classifying network traffic observed in flow traces, the composition and factorization of security chains to be applied for several applications becomes straightforward. Our system infers a high-level representation of the security functions, which can be translated into an concrete implementation in the Pyretic language for programming software-defined networks. We showed that the generated chains satisfy several desirable properties such as the absence

of black holes or loops, shadowing freedom, and that they are consistent with the underlying security policy. We are currently working on optimizing and improving the parameterization of the security chains that are generated by our inference system.

SOCRATE Project-Team

6. New Results

6.1. Multi-User Communications

Activities in axis 2 primarily focus on communicating multi-user systems. They represent the core of the research activity that will be pursued in Maracas team.

The first pillar of our research concerns the evaluation of fundamental limits of wireless systems (e.g. capacity) often express as a fundamental tradeoff : energy efficiency - spectral efficiency tradeoff, rate versus reliability, information versus energy transfert,... Our work relies mostly on information theory, signal processing, estimation theory and game theory.

The second pillar concerns the evaluation of real systems and their performance is confronted to the above mentioned fundamental limits. These activities rely on strong collaborations with industry (Nokia, Orange, SigFox, Sequans, SPIE-ICS,...) We also manage the FIT/CorteXlab testbed offering a remote access to a worldwide unique platform.

Beyond these two pillars, we also explore new research areas where our background is relevant. These prospective activities are performed with external collaborations and prepare the future activity of Maracas team. This year we explored molecular communications (supported by an Inria exploratory project), smart grids in collaboration with Sheffield, VLC in association with Agora team or Privacy preservation in collaboration with Privatics team.

6.1.1. Fundamental limits in communications

6.1.1.1. Variations on point to point capacity and related tools

In [31] discrete approximations of the capacity are introduced where the input distribution is constrained to be discrete in addition to any other constraints on the input. For point-to-point memoryless additive noise channels, rates of convergence to the capacity of the original channel are established for a wide range of channels for which the capacity is finite. These results are obtained by viewing discrete approximations as a capacity sensitivity problem, where capacity losses are studied when there are perturbations in any of the parameters describing the channel. In particular, it is shown that the discrete approximation converges arbitrarily close to the channel capacity at rate $O(\Delta)$, where Δ is the discretization level of the approximation. Examples of channels where this rate of convergence holds are also given, including additive Cauchy and inverse Gaussian noise channels.

In [30] the properties of finite frames are explored. Finite frames are sequences of vectors in finite dimensional Hilbert spaces that play a key role in signal processing and coding theory. We studied the class of tight unit-norm frames for $\mathbb{C}^d$ that also form regular schemes, called tight regular schemes (TRS). Many common frames that arise in applications 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 quadrature rules for integrals over spheres in $\mathbb{C}^d$ with weights dependent on the Voronoi regions of each frame element.

6.1.1.2. Interference channel with feedback

The interference channel is a well-known model used to represent simultaneous transmissions in a wireless environment. In the framework of Victor Quintero's PhD, we explored the performance of this model with noisy feedbacks.

In [35], an achievable η-Nash equilibrium (η-NE) region for the two-user Gaussian interference channel with noisy channel-output feedback is presented for all $\eta \geq 1$. This result is obtained in the scenario in which each transmitter-receiver pair chooses its own transmit-receive configuration in order to maximize its own individual information transmission rate. At an η-NE, any unilateral deviation by either of the pairs does not increase the corresponding individual rate by more than η bits per channel use.

In [6], the capacity region of the linear deterministic interference channel with noisy channel-output feedback (LD-IC-NF) is fully characterized. The proof of achievability is based on random coding arguments and rate splitting; block-Markov superposition coding; and backward decoding. The proof of converse reuses some of the existing outer bounds and includes new ones obtained using genie-aided models. Following the insight gained from the analysis of the LD-IC-NF, an achievability region and a converse region for the two-user Gaussian interference channel with noisy channel-output feedback (GIC-NF) are presented. Finally, the achievability region and the converse region are proven to approximate the capacity region of the G-IC-NF to within 4.4 bits.

6.1.1.3. Wiretap channel

The Wiretap channel allows to address the secrecy constraint in an information theory framework. In [13], an analysis of an input distribution that achieves the secrecy capacity of a general degraded additive noise wiretap channel is presented. In particular, using convex optimization methods, an input distribution that achieves the secrecy capacity is characterized by conditions expressed in terms of integral equations. The new conditions are used to study the structure of the optimal input distribution for three different additive noise cases: vector Gaussian; scalar Cauchy; and scalar exponential.

6.1.1.4. Simultaneous Information and Energy Transmission

Simultaneous information and energy transmission (SIET) is an active research problem and aims at providing energy and information simultaneously from transmitters to receivers. We explore the optimal trade-offs in different settings.

In [34], a non-asymptotic analysis of the fundamental limits of simultaneous energy and information transmission (SEIT) is presented. The notion of information-capacity region, i.e., the largest set of simultaneously achievable information and energy rates, is revisited in a context in which transmissions occur within a finite number of channel uses and strictly positive error decoding probability and energy shortage probability are tolerated. The focus is on the case of one transmitter, one information receiver and one energy harvester communicating through binary symmetric memoryless channels. In this case, the information-capacity region is approximated and the trade-off between information rate and energy rate is thoroughly studied.

In [5], the fundamental limits of simultaneous information and energy transmission (SIET) in the two-user Gaussian interference channel (G-IC) with and without perfect channel-output feedback are approximated by two regions in each case, i.e., an achievable region and a converse region. When the energy transmission rate is normalized by the maximum energy rate the approximation is within a constat gap. In the proof of achievability, the key idea is the use of power-splitting between two signal components: an information-carrying component and a no-information component. The construction of the former is based on random coding arguments, whereas the latter consists in a deterministic sequence known by all transmitters and receivers. The proof of the converse is obtained via cut-set bounds, genie-aided channel models, Fano's inequality and some concentration inequalities considering that channel inputs might have a positive mean. Finally, the energy transmission enhancement due to feedback is quantified and it is shown that feedback can at most double the energy transmission rate at high signal to noise ratios.

6.1.1.5. Modeling Interference in Large-Scale Uplink SCMA

Massive connectivity is a fundamental challenge for IoT, as discussed in the next section from a practical perspective. From a theoretical perspective, we propose to relax the assumption of Gaussian interference.

Fast varying active transmitter sets with very short length transmissions arise in communications for the Internet of Things. As a consequence, the interference is dynamic, leading to non-Gaussian statistics. At the same time, the very high density of devices is motivating non-orthogonal multiple access (NOMA) techniques, such as sparse code multiple access (SCMA). In [2], we study the statistics of the dynamic interference from devices using SCMA. In particular, we show that the interference is $\alpha-$stable with non-trivial dependence structure for large scale networks modeled via Poisson point processes. Moreover, the interference on each frequency band is shown to be sub-Gaussian $\alpha-$stable in the special case of disjoint SCMA codebooks. We investigate the impact of the $\alpha-$stable interference on achievable rates and on the optimal density of devices. Our analysis suggests that ultra dense networks are desirable even with $\alpha-$stable interference.

This contribution is a good introduction of the next section where the performance of IoT access techniques are evaluated.

6.1.1.6. General Massive Machine Type Communications Uplink

Non Orthogonal Multiple Access (NOMA) is expected to play an important role for IoT networks, allowing to reduce signaling overheads and to maximize the capacity of dense networks with multiple packets simultaneous transmission. In the uplink, NOMA can improve significantly the performance of an ALOHA random access if the receiver implements a multiuser detection algorithm. In [11], we compared the performance of a code domain NOMA with a classical ALOHA protocol, through simulations. The code domain NOMA uses random Gaussian codes at the transmitters and exploits compressive sensing at the receiver to maximize users detection and to minimize symbol error rates.

As the number of machine type communications increases at an exponential rate, new solutions have to be found in order to deal with the uplink traffic. At the same time, new types of Base Stations (BS) that use a high number of antennas are being designed, and their beamforming capabilities can help to separate signals that have different angles of arrivals. In [15], we consider a network where a BS serves a high number of nodes that lacks a receive chain, and we analyze the evolution of the outage probability as a function of the number of antennas at the BS. We then study the effect of an angle offset between the main beam and the desired node's direction in order to provide realistic results in a beam-switching scenario.

6.1.1.7. Multiple Base Stations Diversity for UNB Systems

In the framework of the long-term collaboration with Sigfox, the PhD of Yuqi Mo defended mast December, explored the performance of Ultra Narrow Band (UNB) with a focus on sophisticated signal processing techniques such as multi-BS processing or successive interference cancellation (SIC). UNB (Ultra Narrow Band) is one of the technologies dedicated to low-power wide-area communication for IoT, currently exploited by SigFox

In [33], [18], the specificity of UNB is the Aloha-type channel access scheme, asynchronous in both time and frequency domain. This randomness can cause partial spectral interference. In this paper, we take advantage of the spatial diversity of multiple base stations to improve the UNB performance, by using selection combining. In the presence of pathloss and spectral randomness of UNB, the channels are considered correlated. A theoretical analysis of outage probability is demonstrated by considering this correlation, for the case of 2 base stations. This methodology of probability computing can be extended to K BSs. The diversity of multiple receivers is proved to be beneficial in enhancing the performance of UNB networks. This gain is shown to be related to the density of the base stations, as well as the distance between each of them. In [8], we propose to apply signal combining and interference cancellation technologies across multiple base stations in UNB networks, in order to take advantage of their spatial diversity. We evaluate and compare the performance enhancement of each technology, compared to single BS case. These technologies exploiting multi-BS diversity are proved to be significantly beneficial in improving UNB networks' scalability. We can gain until 28 times better performance with one iteration global SIC. We highlight that these results provide us a choice among the technologies according to the improvement needs and the implementation complexity.

6.1.2. Contributions in other application fields

6.1.2.1. Molecular communications

Molecular communications is emerging as a technique to support coordination in nanonetworking, particularly in biochemical systems. In complex biochemical systems such as in the human body, it is not always possible to view the molecular communication link in isolation as chemicals in the system may react with chemicals used for the purpose of communication. There are two consequences: either the performance of the molecular communication link is reduced; or the molecular link disrupts the function of the biochemical system. As such, it is important to establish conditions when the molecular communication link can coexist with a biochemical system. In [4], we develop a framework to establish coexistence conditions based on the theory of chemical reaction networks. We then specialize our framework in two settings: an enzyme-aided molecular communication system; and a low-rate molecular communication system near a general biochemical

system. In each case, we prove sufficient conditions to ensure coexistence. In [29], we develop a general framework for the coexistence problem by drawing an analogy to the cognitive radio problem in wireless communication systems. For the particularly promising underlay strategy, we propose a formalization and outline key consequences.

Another key challenge in nanonetworking is to develop a means of coordinating a large number of nanoscale devices. Devices in molecular communication systems—once information molecules are released— are typically viewed as passive, not reacting chemically with the information molecules. While this is an accurate model in diffusion-limited links, it is not the only scenario. In particular, the dynamics of molecular communication systems are more generally governed by reaction-diffusion, where the reaction dynamics can also dominate. This leads to the notion of reaction-limited molecular communication systems, where the concentration profiles of information molecules and other chemical species depends largely on reaction kinetics. In this regime, the system can be approximated by a chemical reaction network. In [14], we exploit this observation to design new protocols for both point-to-point links with feedback and networks for event detection. In particular, using connections between consensus and advection theory and reaction networks lead to simple characterizations of equilibrium concentrations, which yield simple—but accurate— design rules even for networks with a large number of devices.

6.1.2.2. Smart Grids

Smart grids is another application field where information theory and signal processing can be useful. During 2018, we addressed security issues. In [41], random attacks that jointly minimize the amount of information acquired by the operator about the state of the grid and the probability of attack detection are presented. The attacks minimize the information acquired by the operator by minimizing the mutual information between the observations and the state variables describing the grid. Simultaneously, the attacker aims to minimize the probability of attack detection by minimizing the Kullback-Leibler (KL) divergence between the distribution when the attack is present and the distribution under normal operation. The resulting cost function is the weighted sum of the mutual information and the KL divergence mentioned above. The trade-off between the probability of attack detection and the reduction of mutual information is governed by the weighting parameter on the KL divergence term in the cost function. The probability of attack detection is evaluated as a function of the weighting parameter. A sufficient condition on the weighting parameter is given for achieving an arbitrarily small probability of attack detection. The attack performance is numerically assessed on the IEEE 30-Bus and 118-Bus test systems.

6.1.2.3. Privacy and tracking

In a joint work with Privatics team, we presented in [40] the analysis of an Ultrasound-based tracking application. By analyzing several mobile applications along with the network communication and sample of the original audio signal, we were able to reverse engineer the ultrasonic communications and some other elements of the system. Based on those finding we show how arbitrary ultrasonic signal can be generated and how to perform jamming. Finally we analyze a real world deployment and discuss privacy implications.

6.1.2.4. VLC

In a joint work with Agora, we present in [12] our efforts to design a communication system between an ordinary RGB light emitting diode and a smart-phone. This work in progress presents our preliminary findings obtained investigating this poorly known and unusual channel. We give engineering insights on driving an RGB light emitting diode for camera communication and discuss remaining challenges. Finally, we propose possible solutions to cope with these issues that are blockers for a user ready implementation.

6.1.2.5. Intelligent Transport

On-demand transport has been disrupted by Uber and other providers, which are challenging the traditional approach adopted by taxi services. Instead of using fixed passenger pricing and driver payments, there is now the possibility of adaptation to changes in demand and supply. Properly designed, this new approach can lead to desirable tradeoffs between passenger prices, individual driver profits and provider revenue. However, pricing and allocations - known as mechanisms - are challenging problems falling in the intersection of economics and computer science. In [3], we develop a general framework to classify mechanisms in on-demand transport.

Moreover, we show that data is key to optimizing each mechanism and analyze a dataset provided by a real-world on-demand transport provider. This analysis provides valuable new insights into efficient pricing and allocation in on-demand transport.

6.2. Flexible Radio Front-End

Activities in this axis could globally be divided in two main topics: low-power wireless sensors (with applications in wearable devices, guided propagation for ventilation systems, and tag-to-tag RFID), and optimization of waveforms (for wake-up radio receivers and wireless power transfer).

6.2.1. Low-Power WSN

Wearable sensors for health monitoring can enable the early detection of various symptoms, and hence rapid remedial actions may be undertaken. In particular, the monitoring of cardiac events by using such wearable sensors can provide real-time and more relevant diagnosis of cardiac arrhythmia than classical solutions. However, such devices usually use batteries, which require regular recharging to ensure long-term measurements. In the framework of a local collaborative project, we therefore designed and evaluated a connected sensor for the ambulatory monitoring of cardiac events, which can be used as an autonomous device without the need of a battery. Even when using off-the-shelf, low-cost integrated circuits, by optimizing both the hardware and software embedded in the device, we were able to reduce the energy consumption of the entire system to below 0.4 mW while measuring and storing the ECG on a non-volatile memory. Moreover, in this project, a power-management circuit able to store energy collected from the radio communication interface is proposed, able to make the connected sensor fully autonomous. Initial results show that this sensor could be suitable for a truly continuous and long-term monitoring of cardiac events [32].

In collaboration with Atlantic, we have done here a preliminary study [37], [23] of wireless transmissions using the ventilation metallic ducts as waveguides. Starting from the waveguide theory, we deeply studied in simulation the actual attenuation encountered by radiowaves in such a specific medium. This kind of wireless link appears to be really efficient, and therefore highly promising to implement Internet of Things (IoT) in old buildings to make them smarter. This study also expresses a very simple empirical model in order to ease dimensioning a wireless network in such conditions and a specific antenna design enabling both good performance and high robustness to the influence of the environment.

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.

6.2.2. Optimization of waveforms for wake-up radio and energy harvesting

First Filter Bank Multi Carrier (FBMC) signals are employed in order to improve the performance of a quasi-passive wake-up radio receiver (WuRx) for which the addressing is performed by the means of a frequency fingerprint. The feasibility of such kind of WuRx was already demonstrated by using orthogonal frequency-division multiplexing (OFDM) signals to form the identifiers. Together with the main advantage of this approach (i.e. no base band processing needed and consequently a reduced energy consumption), one of the drawbacks is their low sensitivity. Through a set of circuit-system co-simulations, it is shown that by their characteristics, especially high Peak to Average Power Ratio (PAPR) and high out of band attenuation, FBMC signals manage to boost the sensitivity and moreover to enhance the robustness of this kind of WuRx. Moreover, we introduced robust wake-up IDs for quasi-passive wake-up receivers in an Internet of Things context[16]. These IDs can address single devices and are based on the Hadamard codes. Further a novel wake-up threshold is implemented to make the device more sensitive and robust against false wake-ups (FWUs). The wake-up procedure is simulated with a tap delay line (TDL) model for a line of sight (LOS) channel and a non line of sight (NLOS) channel. In both scenarios sufficient wake-up distances are reached with low false wake-up probabilitys (FWUPs). Additionally, the system is tested against the influence of an external bandwidth use. Finally, a recommendation for the global system is given.

In [21], we are proposing a way to maximize the DC power collected in the case of a wireless power transfer (WPT) scenario. Three main aspects are taken into account: the RF (radio frequency) source, the propagation channel and the rectifier as the main part of the energy collecting circuit. This problem is formulated as a convex optimization one. Then, as a first step towards solving this problem, a rectifier circuit was simulated by using Keysight's ADS software and, by using a classical model identification strategy i.e. Vector Fitting algorithm, the state-space model of the passive parts of this rectifier were extracted. In order to verify the extracted model, S11 input reflection coefficients and DC output voltages of the original circuit and the state-space model are compared.

6.2.3. UWB for localization

Ultra Wide Band (UWB) is a wireless communication technology that is characterized, in its *impulse radio* scheme [55], by very short duration waveforms called *pulses* (in the order of few nanoseconds), using a wide band and low power spectral density. Among the many advantages offered by this technology is the fact that the arrival time of a pulse can be determined quite precisely, giving the opportunity to measure the distance between two communicating devices by estimating the flight time of the signal.

Although this technology has been known for a long time, it is only recently that cheap UWB chips have been commercialized for civilian applications. As the UWB technology is sensitive to many parameters, the effective performance of localization systems based on UWB may vary a lot compared to what is announced in datasheets. Some accuracy studies have been performed [47], [48] but few of them focus on rapid movement of the transceivers.

Indeed, indoor ranging is in itself dependent on many parameter and very difficult to evaluate objectively, but when the transceivers are moving fast (say as if they were attached to dancer's wrists), more parameters are to be taken into account: transceiver calibration, random errors, presence of obstacle, antenna orientation etc.

In [20], we study experimentally the precision of UWB ranging for rapid movements in an indoor environment, based on the technology proposed by Decawave (DW1000 [45]) whose chips have already been integrated in many commercial devices. We show in particular how to improve the precision of the distance measured by averaging the ranging over successive samples.

6.3. Software Radio Programming Model

6.3.1. Non Uniform Memory Access Analyzer

Non Uniform Memory Access (NUMA) architectures are nowadays common for running High-Performance Computing (HPC) applications. In such architectures, several distinct physical memories are assembled to create a single shared memory. Nevertheless, because there are several physical memories, access times to these memories are not uniform depending on the location of the core performing the memory request and on the location of the target memory. Hence, threads and data placement are crucial to efficiently exploit such architectures. To help in taking decision about this placement, profiling tools are needed. In [36], we propose NUMA MeMory Ana-lyzer (NumaMMA), a new profiling tool for understanding the memory access patterns of HPC applications. NumaMMA combines efficient collection of memory traces using hardware mechanisms with original visualization means allowing to see how memory access patterns evolve over time. The information reported by NumaMMA allows to understand the nature of these access patterns inside each object allocated by the application. We show how NumaMMA can help understanding the memory patterns of several HPC applications in order to optimize them and get speedups up to 28% over the standard non optimized version.

6.3.2. Environments for transiently powered devices

An important research initiative is being followed in Socrate today: the study of the new NVRAM technology and its use in ultra-low power context. NVRAM stands for Non-Volatile Radom Access Memory. Non-Volatile memory has been extising for a while (Nand Flash for instance) but was not sufficiently fast to be used as main memory. Many emerging technologies are forseen for Non-Volatile RAM to replace current RAM [50].

Socrate has started a work on the applicability of NVRAM for *transiantly powered systems*, i.e. systems which may undergo power outage at any time. This study resulted in the Sytare software presented at the NVMW conference [25] and also to the starting of an Inria Project Lab [39]: ZEP.

The Sytare software introduces a checkpointing system that takes into account peripherals (ADC, leds, timer, radio communication, etc.) present on all embedded system. Checkpointing is the natural solution to power outage: regularly save the state of the system in NVRAM so as to restore it when power is on again. However, no work on checkpointing took into account the restoration of the states of peripherals, Sytare provides this possibility. A complete description of Sytare has been accepted to IEEE Transaction on Computers [1], special issue on NVRAM.

6.3.3. Dynamic memory allocation for heterogeneous memory systems

In a low power system-on-chip the memory hierarchy is traditionally composed of Static RAM (SRAM) and NOR flash. The main feature of SRAM is a fast access time, while Flash memory is dense, and also non-volatile i.e. it does not require power to retain data. Because of its low writing speed, Flash memory is mostly used in a read-only fashion (e.g. for code) and the amount of SRAM is kept to a minimum in order to lower leakage power.

Emerging memory technologies exhibit different trade-offs and more heterogeneity. Non-Volatile RAM technologies like MRAM (Magnetic RAM) or RRAM (Resistive RAM) open new perspectives on power-management since they can be switched on or off at very little cost. Their characteristics are very dependent on the technology used, but it is now widely known that they will provide a high integration density and fast read access time to persistent data. NVRAM is usually not as fast as SRAM and some technologies have a limited endurance hence are not suited to store frequently modified data. In addition, most NVRAM technologies have asymmetric access times, writes being slower than reads.

In the context of embedded systems, the hardware architecture is evolving towards a model where different memory banks, with different hardware characteristics, are directly exposed to software, as it has been the case for scratchpad memories (SPM). This raises questions including:

- What is the expected performance impact of adding fast memory to a system based on NVRAM? In particular: will the addition of a small amount of fast memory result in significant performance improvement?
- How should one adapt and optimize their software memory management to leverage these new technologies?

In [10], [28], we study these questions in the perpective of dynamic memory allocation. In this first study we show, with extensive profiling how much can be gained with a clever dynamic memory allocation in the context of heterogeneous memory. We limit the study to two different memories, RAM and NVRAM for instance. This gain can go up to 15% of performance, depending of course of the performances of the different memories used. These results will be helpfull to design a clever dynamic allocator for these new architectures and also will help in the design process of new architecture for low power systems that will include NVRAM for normaly-off systems for instance.

6.3.4. Arithmetic for signal processing

Linear Time Invariant (LTI) filters are often specified and simulated using high-precision software, before being implemented in low-precision fixed-point hardware. A problem is that the hardware does not behave exactly as the simulation due to quantization and rounding issues. The article [7] advocates the construction of LTI architectures that behave as if the computation was performed with infinite accuracy, then converted to the low-precision output format with an error smaller than its least significant bit. This simple specification guarantees the numerical quality of the hardware, even for critical LTI systems. Besides, it is possible to derive the optimal values of all the internal data formats that ensure that the specification is met. This requires a detailed error analysis that captures not only the quantization and rounding errors, but also their infinite accumulation in recursive filters. This generic methodology is detailed for the case of low-precision LTI filters in the Direct Form I implemented in FPGA logic. It is demonstrated by a fully automated and open-source architecture generator tool integrated in FloPoCo, and validated on a range of Infinite Impulse Response filters.

6.3.5. Karatsuba multipliers on modern FPGAs

The Karatsuba method is a well-known technique to reduce the complexity of large multiplications. However it is poorly suited to the rectangular 17x25-bit multipliers embedded in recent Xilinx FPGAs: The traditional Karatsuba approach must under-use them as square 18x18 ones. In [17], the Karatsuba method is extended to efficiently use such rectangular multipliers to build larger multipliers. Rectangular multipliers can be efficiently exploited if their input word sizes have a large greatest common divider. In the Xilinx FPGA case, this can be obtained by using the 17x25 embedded multipliers as 16x24. The obtained architectures are implemented with due detail to architectural features such as the pre-adders and post-adders available in Xilinx DSP blocks. They are synthesized and compared with traditional Karatsuba, but also with (non-Karatsuba) state-of-the-art tiling techniques that make use of the full rectangular multipliers. The proposed technique improves resource consumption and performance for multipliers of numbers larger than 64 bits.

6.3.6. PyGA: a Python to FPGA compiler prototype

In a collaboration with Intel, Yohann Uguen has worked on a compiler of Python to FPGA [22]. Based on the Numba Just-In-Time (JIT) compiler for Python and the Intel FPGA SDK for OpenCL, it allows any Python user to use a FPGA card as an accelerator for Python seamlessly, albeit with limited performance so far.

6.3.7. General computer arithmetic

A second edition of the Handbook for Floating-Point Arithmetic has been published [38].

With colleagues from Aric, we have worked on a critical review [42] of the Posit system, a proposed alternative to the prevalent floating-point format.