RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2015

Section New Results

Edition: 2016-03-21

DISTRIBUTED SYSTEMS AND MIDDLEWARE
1. ASAP Project-Team 5
2. ATLANMODELS Team 14
3. CIDRE Project-Team 17
4. COAST Project-Team 23
5. CTRL-A Team 26
6. MIMOVE Team 30
7. MYRIADS Project-Team 34
8. REGAL Project-Team 42
9. SCALE Team 46
10. SPIRALS Project-Team 50
11. WHISPER Project-Team 51
DISTRIBUTED AND HIGH PERFORMANCE COMPUTING
12. ALPINES Project-Team 54
13. AVALON Project-Team 58
14. HIEPACS Project-Team 66
15. KERDATA Project-Team 76
16. MESCAL Project-Team 82
17. MOAIS Project-Team 88
18. ROMA Project-Team 90
19. STORM Team 101
20. TADAAM Team 105
DISTRIBUTED PROGRAMMING AND SOFTWARE ENGINEERING
21. ASCOLA Project-Team 110
22. DIVERSE Project-Team 117
23. FOCUS Project-Team 124
24. INDES Project-Team 129
25. PHOENIX Project-Team 133
26. RMOD Project-Team 135
27. TACOMA Team 141
NETWORKS AND TELECOMMUNICATIONS
28. COATI Project-Team 147
29. DANTE Project-Team 157
30. DIANA Project-Team 163
31. DIONYSOS Project-Team 167
32. DYOGENE Project-Team 176
33. EVA Team 186
34. FUN Project-Team 194
35. GANG Project-Team 199
36. INFINE Team 208
37. MADYNES Project-Team 214

38. MAESTRO Project-Team 225
39. MUSE Team 236
40. RAP Project-Team 239
41. SOCRATE Project-Team 243
42. URBANET Team 251

ASAP Project-Team

6. New Results

6.1. Models and Theory of Distributed Systems

6.1.1. Asynchronous Byzantine Systems: From Multivalued to Binary Consensus with t < n/3, $O(n^2)$ Messages, O(1) Time, and no Signature

Participant: Michel Raynal.

This work [39] presents a new algorithm that reduces multivalued consensus to binary consensus in an asynchronous message-passing system made up of n processes where up to t may commit Byzantine failures. This algorithm has the following noteworthy properties: it assumes t < n/3 (and is consequently optimal from a resilience point of view), uses $O(n^2)$ messages, has a constant time complexity, and does not use signatures. The design of this reduction algorithm relies on two new all-to-all communication abstractions. The first one allows the non-faulty processes to reduce the number of proposed values to c, where c is a small constant. The second communication abstraction allows each non-faulty process to compute a set of (proposed) values such that, if the set of a non-faulty process contains a single value, then this value belongs to the set of any non-faulty process. Both communication abstractions have an $O(n^2)$ message complexity and a constant time complexity. The reduction of multivalued Byzantine consensus to binary Byzantine consensus is then a simple sequential use of these communication abstractions. To the best of our knowledge, this is the first asynchronous message-passing algorithm that reduces multivalued consensus to binary consensus with $O(n^2)$ messages and constant time complexity (measured with the longest causal chain of messages) in the presence of up to t < n/3 Byzantine processes, and without using cryptography techniques. Moreover, this reduction algorithm tolerates message reordering by Byzantine processes.

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

6.1.2. Atomic Read/Write Memory in Signature-free Byzantine Asynchronous Message-passing Systems

Participant: Michel Raynal.

In this work [54] we designed a signature-free distributed algorithm which builds an atomic read/write shared memory on top of an n-process asynchronous message-passing system in which up to t < n/3 processes may commit Byzantine failures. From a conceptual point of view, this algorithm is designed to be as close as possible to the algorithm proposed by Attiya, Bar-Noy and Dolev (JACM 1995), which builds an atomic register in an n-process asynchronous message-passing system where up to t < n/2 processes may crash. The proposed algorithm is particularly simple. It does not use cryptography to cope with Byzantine processes, and is optimal from a t-resilience point of view (t < n/3). A read operation requires O(n) messages, and a write operation requires O(n 2) messages.

This work was done in collaboration with Achour Mostefaoui, Matoula Petrolia, and Claude Jard from the LINA laboratory in Nantes.

6.1.3. Intrusion-Tolerant Broadcast and Agreement Abstractions in the Presence of Byzantine Processes

Participant: Michel Raynal.

A process commits a Byzantine failure when its behavior does not comply with the algorithm it is assumed to execute. Considering asynchronous message-passing systems, this work [18] presents distributed abstractions, and associated algorithms, that allow non-faulty processes to correctly cooperate, despite the uncertainty created by the net effect of asynchrony and Byzantine failures. These abstractions are broadcast abstractions (namely, no-duplicity broadcast, reliable broadcast, and validated broadcast), and agreement abstraction (namely, consensus). While no-duplicity broadcast and reliable broadcast are well-known one-to-all communication abstractions, validated broadcast is a new all-to-all communication abstraction designed to address agreement problems. After having introduced these abstractions, we also presented an algorithm implementing validated broadcast on top of reliable broadcast. Then we presented two consensus algorithms, which are reductions of multivalued consensus to binary consensus. The first one is a generic algorithm, that can be instantiated with unreliable broadcast or no-duplicity broadcast, while the second is a consensus algorithm based on validated broadcast. Finally, a third algorithm is presented that solves the binary consensus problem. This algorithm is a randomized algorithm based on validated broadcast and a common coin. The presentation of all the abstractions and their algorithms is done incrementally. This work was done in collaboration with Achour Mostefaoui from the LINA laboratory in Nantes.

6.1.4. Minimal Synchrony for Asynchronous Byzantine Consensus

Participants: Michel Raynal, Zohir Bouzid.

Solving the consensus problem requires in one way or another that the underlying system satisfies some synchrony assumption. Considering an asynchronous message-passing system of n processes where (a) up to t < n/3 may commit Byzantine failures, and (b) each pair of processes is connected by two uni-directional channels (with possibly different timing properties), this work [24] investigates the synchrony assumption required to solve consensus, and presents a signature-free consensus algorithm whose synchrony requirement is the existence of a process that is an eventual t+1bisource. Such a process p is a correct process that eventually has (a) timely input channels from t correct processes and (b) timely output channels to t correct processes (these input and output channels can connect p to different subsets of processes). As this synchrony condition was shown to be necessary and sufficient in the stronger asynchronous system model (a) enriched with message authentication, and (b) where the channels are bidirectional and have the same timing properties in both directions, it follows that it is also necessary and sufficient in the weaker system model considered in this work. In addition to the fact that it closes a long-lasting problem related to Byzantine agreement, a noteworthy feature of the proposed algorithm lies in its design simplicity, which is a first-class property.

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

6.1.5. Signature-Free Asynchronous Binary Byzantine Consensus with t<n/3, O(n^2) Messages, and O(1) Expected Time

Participant: Michel Raynal.

This work [17] is on broadcast and agreement in asynchronous message-passing systems made up of n processes, and where up to t processes may have a Byzantine Behavior. Its first contribution is a powerful, yet simple, all-to-all broadcast communication abstraction suited to binary values. This abstraction, which copes with up to t < n/3 Byzantine processes, allows each process to broadcast a binary value, and obtain a set of values such that (1) no value broadcast only by Byzantine processes can belong to the set of a correct process, and (2) if the set obtained by a correct process contains a single value v, then the set obtained by any correct process contains v. The second contribution of this work is a new round-based asynchronous consensus algorithm that copes with up to t < n/3 Byzantine processes. This algorithm is based on the previous binary broadcast abstraction and a weak common coin. In addition of being signature-free and optimal with respect to the value of t, this consensus algorithm has several noteworthy properties: the expected number of rounds to decide is constant; each round is composed of a constant number of communication steps and involves O(n^2) messages; each message is composed of a round number plus a constant number of bits. Moreover, the algorithm tolerates message reordering by the adversary (i.e., the Byzantine processes). This work was done in collaboration with Achour Mostefaoui from the LINA laboratory in Nantes, and Hamouma Moumen from Université de Béjaïa.

6.1.6. Specifying Concurrent Problems: Beyond Linearizability and up to Tasks

Participants: Michel Raynal, Zohir Bouzid.

Tasks and objects are two predominant ways of specifying distributed problems. A task specifies for each set of processes (which may run concurrently) the valid outputs of the processes. An object specifies the outputs the object may produce when it is accessed sequentially. Each one requires its own implementation notion, to tell when an execution satisfies the specification. For objects linearizability is commonly used, while for tasks implementation notions are less explored. Sequential specifications are very convenient, especially important is the locality property of linearizability, which states that linearizable objects compose for free into a linearizable object. However, most well-known tasks have no sequential specification. Also, tasks have no clear locality property. This work [26] introduces the notion of interval-sequential object. The corresponding implementation notion of interval-linearizability generalizes linearizability. Interval-linearizability allows to specify any task. However, there are sequential one-shot objects that cannot be expressed as tasks, under the simplest interpretation of a task. We also showed that a natural extension of the notion of a task is expressive enough to specify any interval-sequential object.

This work was done in collaboration with Armando Castaneda and Sergio Rajsbaumm from UNAM, Mexico.

6.1.7. Test-and-Set in Optimal Space

Participant: George Giakkoupis.

The test-and-set object is a fundamental synchronization primitive for shared memory systems. In [35] we address the number of registers (supporting atomic reads and writes) required to implement a one-shot test-and-set object in the standard asynchronous shared memory model with n processes. The best lower bound is $\log n - 1$ for obstruction-free and deadlock-free implementations, and recently a deterministic obstruction-free implementation using $O(\sqrt{n})$ registers was presented.

In [35] we close the gap between these existing upper and lower bounds by presenting a deterministic obstruction-free implementation of a one-shot test-and-set object from $\Theta(\log n)$ registers of size $\Theta(\log n)$ bits. Combining our obstruction-free algorithm with techniques from previous research, we also obtain a randomized wait-free test-and-set algorithm from $\Theta(\log n)$ registers, with expected step-complexity $\Theta(\log^* n)$ against the oblivious adversary. The core tool in our algorithm is the implementation of a deterministic obstruction-free *sifter* object, using only 6 registers. If k processes access a sifter, then when they have terminated, at least one and at most $\mathsf{floor}(2k+1)/3$ processes return "win" and all others return "lose".

This is a joint work with Maryam Helmi (U. of Calgary), Lisa Higham (U. of Calgary), and Philipp Woelfel (U. of Calgary), supported by the RADCON Inria Associate Team.

6.2. Graph and Probabilistic Algorithms

6.2.1. On the Quadratic Shortest Path Problem

Participant: Davide Frey.

Finding the shortest path in a directed graph is one of the most important combinatorial optimization problems, having applications in a wide range of fields. In its basic version, however, the problem fails to represent situations in which the value of the objective function is determined not only by the choice of each single arc, but also by the combined presence of pairs of arcs in the solution. In this work [40] we model these situations as a Quadratic Shortest Path Problem, which calls for the minimization of a quadratic objective function subject to shortest-path constraints. We prove strong NP-hardness of the problem and analyze polynomially solvable special cases, obtained by restricting the distance of arc pairs in the graph that appear jointly in a quadratic monomial of the objective function. Based on this special case and problem structure, we devise fast lower bounding procedures for the general problem and show computationally that they clearly outperform other approaches proposed in the literature in terms of its strength.

6.2.2. Tight Bounds on Vertex Connectivity Under Vertex Sampling

Participant: George Giakkoupis.

A fundamental result by Karger (SODA 1994) states that for any λ-edge-connected graph with n nodes, independently sampling each edge with probability $p = \Omega(\log n/\lambda)$ results in a graph that has edge connectivity $\Omega(\lambda p)$, with high probability. In [27] we prove the analogous result for vertex connectivity, when sampling vertices. We show that for any k-vertex-connected graph G with n nodes, if each node is independently sampled with probability $p = \Omega(\sqrt{\log n/k})$, then the subgraph induced by the sampled nodes has vertex connectivity $\Omega(kp^2)$, with high probability. This bound improves upon the recent results of Censor-Hillel et al. (SODA 2014) and is existentially optimal.

This is a joint work with Keren Censor-Hillel (Technion), Mohsen Ghaffari (MIT), Bernhard Haeupler (Carnegie Mellon U.), and Fabian Kuhn (U. of Freiburg).

6.3. Scalable Systems

6.3.1. Being prepared in a sparse world: the case of KNN graph construction

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

This work presents KIFF [41], a generic, fast and scalable KNN graph construction algorithm. KIFF directly exploits the bipartite nature of most datasets to which KNN algorithms are applied. This simple but powerful strategy drastically limits the computational cost required to rapidly converge to an accurate KNN solution, especially for sparse datasets. Our evaluation on a representative range of datasets show that KIFF provides, on average, a speed-up factor of 14 against recent state-of-the art solutions while improving the quality of the KNN approximation by 18

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

6.3.2. Cheap and Cheerful: Trading Speed and Quality for Scalable Social Recommenders

Participants: Anne-Marie Kermarrec, François Taïani, Juan M. Tirado Martin.

Recommending appropriate content and users is a critical feature of on-line social networks. Computing accurate recommendations on very large datasets can however be particularly costly in terms of resources, even on modern parallel and distributed infrastructures. As a result, modern recommenders must generally trade-off quality and computational cost to reach a practical solution. This trade-off has however so far been largely left unexplored by the research community, making it difficult for practitioners to reach informed design decisions. In this work [37], we investigate to which extent the additional computing costs of advanced recommendation techniques based on supervised classifiers can be balanced by the gains they bring in terms of quality. In particular, we compare these recommenders against their unsupervised counterparts, which offer lightweight and highly scalable alternatives. We propose a thorough evaluation comparing 11 classifiers against 7 lightweight recommenders on a real Twitter dataset. Additionally, we explore data grouping as a method to reduce computational costs in a distributed setting while improving recommendation quality. We demonstrate how classifiers trained using data grouping can reduce their computing time by 6 while improving recommendations up to 22% when compared with lightweight solutions.

6.3.3. Fast Nearest Neighbor Search

Participants: Fabien André, Anne-Marie Kermarrec.

Nearest Neighbor (NN) search in high dimension is an important feature in many applications, such as multimedia databases, information retrieval or machine learning. Product Quantization (PQ) is a widely used solution which offers high performance, i.e., low response time while preserving a high accuracy. PQ represents high-dimensional vectors by compact codes. Large databases can therefore be stored in memory, allowing NN queries without resorting to slow I/O operations. PQ computes distances to neighbors using cache-resident lookup tables, thus its performance remains limited by (i) the many cache accesses that the algorithm requires, and (ii) its inability to leverage SIMD instructions available on modern CPUs.

To address these limitations, we designed a novel algorithm, PQ Fast Scan [19], that transforms the cache-resident lookup tables into small tables, sized to fit SIMD registers. This transformation allows (i) in-register lookups in place of cache accesses and (ii) an efficient SIMD implementation. PQ Fast Scan has the exact same accuracy as PQ, while having 4 to 6 times lower response time (e.g., for 25 million vectors, scan time is reduced from 74ms to 13ms).

This work was done in collaboration with Nicolas Le Scouarnec.

6.3.4. Holons: towards a systematic approach to composing systems of systems

Participants: Yérom-David Bromberg, François Taïani.

The world's computing infrastructure is increasingly differentiating into self-contained distributed systems with various purposes and capabilities (e.g. IoT installations, clouds, VANETs, WSNs, CDNs,. . .). Furthermore, such systems are increasingly being composed to generate systems of systems that offer value-added functionality. Today, however, system of systems composition is typically ad-hoc and fragile. It requires developers to possess an intimate knowledge of system internals and low-level interactions between their components. In this work [21], we outline a vision and set up a research agenda towards the generalised programmatic construction of distributed systems as compositions of other distributed systems. Our vision, in which we refer uniformly to systems and to compositions of systems as holons, employs code generation techniques and uses common abstractions, operations and mechanisms at all system levels to support uniform system of systems composition. We believe our holon approach could facilitate a step change in the convenience and correctness with which systems of systems can be built, and open unprecedented opportunities for the emergence of new and previously-unenvisaged distributed system deployments, analogous perhaps to the impact the mashup culture has had on the way we now build web applications.

This work was done in collaboration with Gordon Blair Geoff Coulson, and Yehia Elkhatib from Lancaster University (UK), Laurent Réveillère from University of Bordeaux / Labri, and Heverson Borba Ribeiro and Etienne Rivière from University of Neuchâtel (Switzerland).

6.3.5. Hybrid datacenter scheduling

Participant: Anne-Marie Kermarrec.

We address the problem of efficient scheduling of large clusters under high load and heterogeneous workloads. A heterogeneous workload typically consists of many short jobs and a small number of large jobs that consume the bulk of the cluster's resources.

Recent work advocates distributed scheduling to overcome the limitations of centralized schedulers for large clusters with many competing jobs. Such distributed schedulers are inherently scalable, but may make poor scheduling decisions because of limited visibility into the overall resource usage in the cluster. In particular, we demonstrate that under high load, short jobs can fare poorly with such a distributed scheduler.

We propose instead a new hybrid centralized/ distributed scheduler, called Hawk. In Hawk, long jobs are scheduled using a centralized scheduler, while short ones are scheduled in a fully distributed way. Moreover, a small portion of the cluster is reserved for the use of short jobs. In order to compensate for the occasional poor decisions made by the distributed scheduler, we propose a novel and efficient randomized work-stealing algorithm.

We evaluate Hawk using a trace-driven simulation and a prototype implementation in Spark. In particular, using a Google trace, we show that under high load, compared to the purely distributed Sparrow scheduler, Hawk improves the 50th and 90th percentile runtimes by 80% and 90% for short jobs and by 35% and 10% for long jobs, respectively. Measurements of a prototype implementation using Spark on a 100-node cluster confirm the results of the simulation. This work has been done in the context of the Inria/epfl research center and in collaboration with Pamela delgado, Florin Dinu and Willy Zwaenepoel from EPFL and published in Usenix ATC in 2015 [30].

6.3.6. Out-of-core KNN Computation

Participants: Nitin Chiluka, Anne-Marie Kermarrec, Javier Olivares.

This work proposes a novel multi threading approach to compute KNN on large datasets by leveraging both disk and main memory efficiently. The main rationale of our approach is to minimize random accesses to disk, maximize sequential access to data and efficient usage of only a fraction of the available memory. This approach is evaluated by comparing its performance with a fully in-memory implementation of KNN, in terms of execution time and memory consumption. This multithreading approach outperforms the in-memory baseline in all cases when the large dataset does not fit in memory.

6.3.7. Scaling Out Link Prediction with SNAPLE

Participants: Anne-Marie Kermarrec, François Taïani, Juan M. Tirado Martin.

A growing number of organizations are seeking to analyze extra large graphs in a timely and resource-efficient manner. With some graphs containing well over a billion elements, these organizations are turning to distributed graph-computing platforms that can scale out easily in existing data-centers and clouds. Unfortunately such platforms usually impose programming models that can be ill suited to typical graph computations, fundamentally undermining their potential benefits. In this work [38], we consider how the emblematic problem of link-prediction can be implemented efficiently in gather-apply-scatter (GAS) platforms, a popular distributed graph-computation model. Our proposal, called Snaple, exploits a novel highly-localized vertex scoring technique, and minimizes the cost of data flow while maintaining prediction quality. When used within GraphLab, Snaple can scale to very large graphs that a standard implementation of link prediction on GraphLab cannot handle. More precisely, we show that Snaple can process a graph containing 1.4 billions edges on a 256 cores cluster in less than three minutes, with no penalty in the quality of predictions. This result corresponds to an over-linear speedup of 30 against a 20-core standalone machine running a non-distributed state-of-the-art solution.

6.3.8. Similitude: Decentralised Adaptation in Large-Scale P2P Recommenders

Participants: Davide Frey, Anne-Marie Kermarrec, Pierre-Louis Roman, François Taïani.

Decentralised recommenders have been proposed to deliver privacy-preserving, personalised and highly scalable on-line recommendations. Current implementations tend, however, to rely on a hard-wired similarity metric that cannot adapt. This constitutes a strong limitation in the face of evolving needs. In this work [33], we propose a framework to develop dynamically adaptive decentralized recommendation systems. Our proposal supports a decentralised form of adaptation, in which individual nodes can independently select, and update their own recommendation algorithm, while still collectively contributing to the overall system's mission.

This work was done in collaboration with Christopher Maddock and Andreas Mauthe (Univ. of Lancaster, UK).

6.3.9. Transactional Memory Recommenders

Participant: Anne-Marie Kermarrec.

The Transactional Memory (TM) paradigm promises to greatly simplify the development of concurrent applications. This led, over the years, to the creation of a plethora of TM implementations delivering wide ranges of performance across workloads. Yet, no universal TM implementation fits each and every workload. In fact, the best TM in a given workload can reveal to be disastrous for another one. This forces developers to face the complex task of tuning TM implementations, which significantly hampers the wide adoption of TMs. In this work, we address the challenge of automatically identifying the best TM implementation for a given workload. Our proposed system, ProteusTM, hides behind the TM interface a large library of implementations. Under the hood, it leverages an innovative, multi-dimensional online optimization scheme, combining two popular machine learning techniques: Collaborative Filtering and Bayesian Optimization. We integrated ProteusTM in GCC and demonstrated its ability to switch TM implementations and adapt several configuration parameters (e.g., number of threads). We extensively evaluated ProteusTM, obtaining average performance 3% less than the optimal, and gains up to 100 over static alternatives.

This work has been done in collaboration with Rachid Guerraoui from EPFL, Diego Didona Nuno Diegues, Ricardo Neves and Paolo Romano from INESC, Lisboa) and will be published in ASPLOS 2016 [31].

6.3.10. Want to scale in centralized systems? Think P2P

Participants: Anne-Marie Kermarrec, François Taïani.

Peer-to-peer (P2P) systems have been widely researched over the past decade, leading to highly scalable implementations for a wide range of distributed services and applications. A P2P system assigns symmetric roles to machines, which can act both as client and server. This distribution of responsibility alleviates the need for any central component to maintain a global knowledge of the system. Instead, each peer takes individual decisions based on a local and limited knowledge of the rest of the system, providing scalability by design. While P2P systems have been successfully applied to a wide range of distributed applications (multicast, routing, caches, storage, pub-sub, video streaming), with some highly visible successes (Skype, Bitcoin), they tend to have fallen out of fashion in favor of a much more cloud-centric vision of the current Internet. We think this is paradoxical, as cloud-based systems are themselves large-scale, highly distributed infrastructures. They reside within massive, densely interconnected datacenters, and must execute efficiently on an increasing number of machines, while dealing with growing volumes of data. Today even more than a decade ago, large-scale systems require scalable designs to deliver efficient services. In this work [16] we argue that the local nature of P2P systems is key for scalability regardless whether a system is eventually deployed on a single multi-core machine, distributed within a data center, or fully decentralized across multiple autonomous hosts. Our claim is backed by the observation that some of the most scalable services in use today have been heavily influenced by abstractions and rationales introduced in the context of P2P systems. Looking to the future, we argue that future large-scale systems could greatly benefit from fully decentralized strategies inspired from P2P systems. We illustrate the P2P legacy through several examples related to Cloud Computing and Big Data, and provide general guidelines to design large-scale systems according to a P2P philosophy.

6.3.11. WebGC: Browser-based gossiping

Participants: Raziel Carvajal Gomez, Davide Frey, Anne-Marie Kermarrec.

The advent of browser-to-browser communication technologies like WebRTC has renewed interest in the peer-to-peer communication model. However, the available WebRTC code base still lacks important components at the basis of several peer-to-peer solutions. Through a collaboration with Mathieu Simonin from the Inria SED in the context of the Brow2Brow ADT project, we started to tackle this problem by proposing WebGC, a library for gossip-based communication between web browsers. Due to their inherent scalability, gossip-based, or epidemic protocols constitute a key component of a large number of decentralized applications. WebGC thus represents an important step towards their wider spread. We demonstrated the final version of the library at WISE 2015 [53].

6.4. Privacy in User Centric Applications

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

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

Whether we are shopping for an interesting book or selecting a movie to watch, the chances are that a recommendation system will help us decide what we want. Recommendation systems collect information about our own preferences, compare them to those of other users, and provide us with suggestions on a variety of topics. But is the information gathered by a recommendation system safe from potential attackers, be them other users, or companies that access the recommendation system? And above all, can service providers protect this information while still providing effective recommendations? In this work, we analyze the effect of Sybil attacks on collaborative-filtering recommendation systems, and discuss the impact of different similarity metrics in the trade-off between recommendation quality and privacy. Our results, on a state-of-the-art recommendation framework and on real datasets show that existing similarity metrics exhibit a wide range of behaviors in the presence of Sybil attacks. Yet, they are all subject to the same trade off: Sybil resilience for recommendation quality. We therefore propose and evaluate a novel similarity metric that combines the best of both worlds: a low RMSE score with a prediction accuracy for Sybil users of only a few percent. A preliminary version of this work was published at EuroSec 2015 [32].

This work was done in collaboration with Antoine Boutet, and Rachid Guerraoui.

6.4.2. Decentralized view prediction for global content placement

Participants: Stéphane Delbruel, Davide Frey, François Taïani.

A large portion of today's Internet traffic originates from streaming and video services. Storing, indexing, and serving these videos is a daily engineering challenge that requires increasing amounts of efforts and infrastructures. One promising direction to improve video services consists in predicting at upload time where and when a new video might be viewed, thereby optimizing placement and caching decisions. Implementing such a prediction service in a scalable manner poses significant technical challenges. In this work [28], we address these challenges in the context of a decentralized storage system consisting of set-top boxes or end nodes. Specifically, we propose a novel data placement algorithm that exploits information about the tags associated with existing content, such as videos, and uses it to infer the number of views that newly uploaded content will have in each country.

6.4.3. Distance-Based Differential Privacy in Recommenders

Participant: Anne-Marie Kermarrec.

The upsurge in the number of web users over the last two decades has resulted in a significant growth of online information. This information growth calls for recommenders that personalize the information proposed to each individual user. Nevertheless, personalization also opens major privacy concerns. We designed D2P, a novel protocol that ensures a strong form of differential privacy, which we call distance-based differential privacy, and which is particularly well suited to recommenders. D2P avoids revealing exact user profiles by creating altered profiles where each item is replaced with another one at some distance. We evaluate D2P analytically and experimentally on MovieLens and Jester datasets and compare it with other private and non-private recommenders. This work has been done in the context of the Web-Alter-ego Google focused award and in collaboration with Rachid guerraoui, Rhicheek Patra and Masha Taziki from EPFL and published in PVLVB in 2015 [15].

6.4.4. Privacy-Conscious Information Diffusion in Social Networks

Participants: George Giakkoupis, Arnaud Jégou, Anne-Marie Kermarrec, Nupur Mittal.

This work presents a distributed algorithm – Riposte [47], for information dissemination in social networks which guarantees to preserve the privacy of its users. RIPOSTE ensures that information spreads widely if and only if a large fraction of users find it interesting, and this is done in a "privacy-conscious" manner, namely without revealing the opinion of any individual user. Whenever an information item is received by a user, RIPOSTE decides to either forward the item to all the user's neighbors, or not to forward it to anyone. The decision is randomized and is based on the user's (private) opinion on the item, as well as on an upper bound s on the number of user's neighbors that have not received the item yet.

This work was done in collaboration with Rachid Guerraoui from EPFL, Switzerland.

6.4.5. Hide & Share: Landmark-based Similarity for Private knn Computation

Participants: Davide Frey, Anne-Marie Kermarrec, Antoine Rault, François Taïani.

Computing k-nearest-neighbor graphs constitutes a fundamental operation in a variety of data-mining applications. As a prominent example, user-based collaborative-filtering provides recommendations by identifying the items appreciated by the closest neighbors of a target user. As this kind of applications evolve, they will require KNN algorithms to operate on more and more sensitive data. This has prompted researchers to propose decentralized peer-to-peer KNN solutions that avoid concentrating all information in the hands of one central organization. Unfortunately, such decentralized solutions remain vulnerable to malicious peers that attempt to collect and exploit information on participating users.

In this work [22], we seek to overcome this limitation by proposing H&S (Hide & Share), a novel landmark-based similarity mechanism for decentralized KNN computation. Landmarks allow users (and the associated peers) to estimate how close they lay to one another without disclosing their individual profiles.

We evaluate H&S in the context of a user-based collaborative-filtering recommender with publicly available traces from existing recommendation systems. We show that although landmark-based similarity does disturb similarity values (to ensure privacy), the quality of the recommendations is not as significantly hampered. We also show that the mere fact of disturbing similarity values turns out to be an asset because it prevents a malicious user from performing a profile reconstruction attack against other users, thus reinforcing users' privacy. Finally, we provide a formal privacy guarantee by computing the expected amount of information revealed by H&S about a user's profile.

This work was done in collaboration with Antoine Boutet from the University of St. Etienne, and with Jingjing Wang and Rachid Guerraoui from EPFL, Switzerland.

ATLANMODELS Team

7. New Results

7.1. Reverse Engineering & Evolution

Model Driven Reverse Engineering (MDRE), with its applications on software modernization or tool evolution for example, is a discipline in which model-based principles and techniques are used to treat various kinds of (sometimes very large) existing systems. In the continuity the work started several years ago, AtlanMod has been working actively on this research area this year again. The main contributions are the following:

- In the context of the ARTIST FP7 project, the work has been continued on reusing (and extending accordingly) MoDisco and several of its components to provide the Reverse Engineering support required within the project (and more particularly in the context of the use cases provided by our industrial partners). This has been an important year for the team in this project since it successfully ended in November 2015 after final review at the European Commission. At conceptual-level, the proposed overall approach (as a main result of the ARTIST project) and the main lessons we have learned from its application to concrete industrial scenarios have been published and promoted to a large and high-level audience [11]. The ARTIST project in itself, the various research aspects it addresses and the offered technical solutions have also been presented to the Modeling community [22]. At tooling-level, several (MoDisco-based) model discovery components from Java and SQL have been enhanced while made available as part of a second version of the official ARTIST OS Release [0]. A promising work has also been started on studying deeper the automated discovery of behavioral aspects of software applications, notably by working on a pragmatic mapping between a programming language (Java) and a modeling language (the OMG fUML standard) that focuses on executable aspects.
- To facilitate the understanding of existing software applications via the different models describing them, a significant work has been performed related to providing a generic support for dealing with viewpoints and views expressed on set of possibly heterogeneous and large models. To this intent, and directly capitalizing on the work performed in the TEAP FUI project that ended by the end of 2014, the EMF Views prototype has been significantly refined and enhanced with a ViewPoint Definition Language (the VPDL domain-specific language having a SQL-like syntax) notably [18]. Based on this same model viewpoints/views approach, and more particularly on its underlying (meta)model virtualization support, the general problem of lightweight (meta)model extension has been studied more deeply in the context of our work within the MoNoGe FUI project (national). This has already resulted in a corresponding prototype and a DSL for expressing metamodel extensions [17]. Within the coming year, the plan is to continue further this global work on model viewpoints/views in a software understanding and evolution context.
- Software development projects are notoriously complex and difficult to deal with. Several support tools have been introduced in the past decades to ease the development activities such as issue tracking, code review and Source Control Management (SCM) systems. While such tools efficiently track the evolution of a given aspect of the project (e.g., issues or code), they provide just a partial view of the software project and they often lack of querying mechanisms beyond basic support (e.g., command line, simple gui). This is particularly true for projects that rely on Git, one of the most popular SCM systems. Nowadays many tools are built on top of it, however, they do not complement Git with query functionalities and currently none of them proposes a mechanism that unifies the project information scattered in such different tools. In [28], we propose a conceptual schema for Git and an approach that, given a Git repository, exports its data to a relational database in order to (1) promote data integration with other existing Git-based tools relying on databases

[0] http://www.artist-project.eu/tools-of-toolbox/193

and (2) provide query functionalities expressed through standard SQL syntax. To ensure efficiency, our approach comes with an incremental propagation mechanism that refreshes the database content with the latest modifications.

7.2. MDE Scalability

The increasing number of companies embracing MDE methods and tools have exceeded the limits of the current model-based technologies, presenting scalability issues while facing the growing complexity of their data. Since further research and development is imperative in order to maintain MDE techniques as relevant as they are in less complex contexts, we have focused our research in three axes, (i) scalable persistence solutions, (ii) scalable model transformation engines, and (iii) testing of large scale distributed systems.

In [33], we introduce and evaluate a map-based persistence model for MDE tools. We use this model to build a transparent persistence layer for modeling tools, on top of a map-based database engine. The layer can be plugged into the Eclipse Modeling Framework, lowering execution times and memory consumption levels of other existing approaches. Empirical tests are performed based on a typical industrial scenario, model-driven reverse engineering, where very large software models originate from the analysis of massive code bases. The layer is freely distributed and can be immediately used for enhancing the scalability of any existing Eclipse Modeling tool. We learned that—in terms of performance—typical model-access APIs, with fine-grained methods that only allow for one-step-navigation queries, do not benefit from complex relational or graph-based data structures. Much better results are potentially obtained by optimized low-level data structures, like hash-tables, which guarantee low and constant access times. Additional features that may be of interest in scenarios where performance is not an issue (such as versioning and transactional support provided by CDO) have not been considered. In [32] we extend our persistent mechanism to distributed environments by presenting NeoEMF/HBase, a model-persistence backend for the Eclipse Modeling Framework (EMF) built on top of the Apache HBase data store. Model distribution is hidden from client applications, that are transparently provided with the model elements they navigate. Access to remote model elements is decentralized, avoiding the bottleneck of a single access point. The persistence model is based on key-value stores that allow for efficient on-demand model persistence.

Once we develop a high-performance and distributed persistence mechanism for very-large models, we can exploit it to run high-performance computing over such models. One of the central operations in MDE is rule-based model transformation (MT). It is used to specify manipulation operations over structured data coming in the form of model graphs. However, being based on computationally expensive operations like subgraph isomorphism, MT tools are facing issues on both memory occupancy and execution time while dealing with the increasing model size and complexity. One way to overcome these issues is to exploit the wide availability of distributed clusters in the Cloud for the distributed execution of MT. In [24] and [23], we propose an approach to automatically distribute the execution of model transformations written in a popular MT language, ATL, on top of a well-known distributed programming model, MapReduce. We show how the execution semantics of ATL can be aligned with the MapReduce computation model. We describe the extensions to the ATL transformation engine to enable distribution, and we experimentally demonstrate the scalability of this solution in a reverse-engineering scenario.

Another fundamental operation in MDE is model querying. The Object Constraint Language (OCL) is the standard query language proposed by OMG and is a central component in other modeling and transformation languages such as the Unified Modeling Language (UML), the Meta Object Facility (MOF), and Query View Transformation (QVT). OCL is standardized as a strict functional language. In [34], we propose a lazy evaluation strategy for OCL. We argue that a lazy evaluation semantics is beneficial in some model-driven engineering scenarios for: i) lowering evaluation times on very large models; ii) simplifying expressions on models by using infinite data structures (e.g., infinite models); iii) increasing the reusability of OCL libraries. We implement the approach on the ATL virtual machine EMFTVM.

Finally an important class of operations in MDE is bidirectional (i.e. reversible) computation. Especially bidirectional model transformation is a key technology when two models that can change over time have to be kept constantly consistent with each other. In Hidaka et al. we clarify and visualize the space of design

choices for bidirectional transformations from an MDE point of view, in the form of a feature model. The selected list of existing approaches are characterized by mapping them to the feature model. Then the feature model is used to highlight some unexplored research lines in bidirectional transformations, especially in the scalability of such systems.

7.3. Software Quality

We initiated a new line of research in order to investigate Novelty Search (NS) for the automatic generation of test data, in collaboration with the DiverSE team. Our goal is to explore the huge space of test data within the input domain. In this approach, we select test data based on a novelty score showing how different they are compared to all other solutions evaluated so far [25], [26].

CIDRE Project-Team

7. New Results

7.1. Intrusion detection

7.1.1. Alert Correlation in Distributed Systems

In large systems, multiple (host and network) Intrusion Detection Systems (IDS) and many sensors are usually deployed. They continuously and independently generate notifications (event's observations, warnings and alerts). To cope with this amount of collected data, alert correlation systems have to be designed. An alert correlation system aims at exploiting the known relationships between some elements that appear in the flow of low level notifications to generate high semantic meta-alerts. The main goal is to reduce the number of alerts returned to the security administrator and to allow a higher level analysis of the situation. However, producing correlation rules is a highly difficult operation, as it requires both the knowledge of an attacker, and the knowledge of the functionalities of all IDSes involved in the detection process. In [59], [38], [19], we focus on the transformation process that allows to translate the description of a complex attack scenario into correlation rules and its assessment. We show that, once a human expert has provided an action tree derived from an attack tree, a fully automated transformation process can generate exhaustive correlation rules that would be tedious and error prone to enumerate by hand. The transformation relies on a detailed description of various aspects of the real execution environment (topology of the system, deployed services, etc.). Consequently, the generated correlation rules are tightly linked to the characteristics of the monitored information system. The proposed transformation process has been implemented in a prototype that generates correlation rules expressed in an attack description language called Adele. Additionally, a work has been performed to assess the approach on real environment, and to evaluate the accuracy of the rules built.

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

7.1.2. Android Malware Analysis

We explore how information flows induced by a tainted application are helpful to understand how this tainted application interacts within other components inside the operating system. For that purpose, we have defined a new data structure called System Flow Graph representing in a graph how a marked data is disseminated (inside the operating system). We have shown that this data structure is helpful to understand and represent malicious behaviors [31]. Our main challenge is thus to be able to produce relevant graphs which means being able to really observe malicious executions.

For that purpose we developed GroddDroid [25] a tool dedicated to the automatic triggering of Android malware. GroddDroid makes a first static analysis of the application bytecode. During this analysis, GroddDroid identifies the suspicious parts of the bytecode and modifies the bytecode in order to exhibit an execution path that leads to these suspicious parts. The application is later reconstructed/recompiled in order to be executed. This way, GroddDroid offers a way to force the suspicious code to be executed and then observed.

7.1.3. Comparative Study of Alert Formats

In the context of the SECEF project, we conduced a comparative study of different existing alert formats [39]. We analyzed two proprietary formats, CEF (HP ArcSight) and LEEF (IBM QRADAR), as well as 4 standard formats, IDMEF (IETF), CEE (MITRE), CIM and CADF (DMTF). We proposed several metrics to compare them based on an accurate revue of every fields proposed by each format. The results show that IDMEF is the most expressive and structured format. However, some fields proposed by other formats are not covered in IDMEF. We proposed some modification of the alert format to take those limitations into account.

7.1.4. Visualization

This year, research on viusalization for security was oriented towards two objectives. First, as we did during the previous years, we tried to provide solution for security analysts to better analyze *a posteriori* events related to security that are happening on a system. Christopher Humphries, who was the first CIDre Ph.D. student on this topic defended his Ph.D. Thesis *User-Centered Security Events Visualization* this December. We should also mention that we presented a prototype of our tool ELVIS during the FIC 2015 in Lille on the Pôle Cyber-Défense area.

This year, we also started research on a new topic in visualization for security. By contrast with our previous work that was dedicated to forensics, i.e. *a posteriori* analysis of security events, we started this year to study real time analysis of alerts generated by an IDS. The idea here is to allow better monitoring of what is currently happening on a system. We proposed VEGAS, a tool that allows front-line security operators to perform a first triage of the alerts to provide consistent groups of alerts to security analysts. A new Ph.D. student, Damien Crémilleux, was hired on a DGA-MI funding, to work on this topic. VEGAS was presented during the poster session of VizSec 2015 [58] that took place in Chicago, Illinois, USA on the 26th of October 2015.

7.2. Privacy

7.2.1. The Right to be Forgotten

The right to be forgotten, or to oblivion, is an aspect of privacy rights. It relates to the need for individuals to be able to leave a part of their past behind them, to change their mind about something or to take a new start in a given domain. The final report of the DAO project [53] presents an analysis of the concept from a multidisciplinary point of view, including a sociological study, a legal state of the art assorted with insights of possible evolutions, and a technical state of the art along with the proposal of a new architecture [22]. A joint technical and legal analysis of the conceptual and technical issues specific to social networks is also proposed. From the point of view of a computer scientist, the most obvious issue with the right to be forgotten is the ability to control the deletion of a piece of information once it has been disclosed and disseminated. In the general case, no guarantees can be provided, but under certain conditions it is possible to enforce remote deletion with reasonable guarantees. In general, it implies that architectural and applicative choices are made beforehand, either to allow for future decisions regarding data made available in a controlled framework, like late modifications of its access policy or the triggering of its destruction, or to plan deletion from the beginning and set a time-to-leave when disclosing the data within a particular environment, or . The approach designed in CIDRE, relying on both ephemeral publication and data degradation techniques, falls in the latter category, improving the utility for third parties (when compared to existing ephemeral publication techniques) and building a new trade-off with the users' privacy needs, by making different versions of the original data, more or less precise, available for different durations, the more detailed information being lost the quickliest.

CIDRE also contributes, through the B<>com IRT, to the supervision (by Annie Blandin, professor at Télécom Bretagne, and Guillaume Piolle) of Gustav Malis's doctoral work in law in the domain of a restrictive case of the right to be forgotten. In this context, very original contributions have been made at the intersection between the two fields. In particular, a joint analysis has been proposed on the roles of legal and computing tools for the implementation of the right to be forgotten [50]. In particular, it seems that the two domains consider the issue with very different perspectives: the computer scientist almost takes for granted that he cannot rely on regulations and on "security through legality", hence the tools he designs are intended to directly empower the

user, putting him in control of his data by using preventive protection techniques. The tools may fail though, or more likely their applicability conditions may not suit all scenarios. When issues arise they may be captured by the regulatory framework, which intends to provide means for reparation and restoration. Both approaches fail to encompass all possible situations and to solve all potential issues, but they provide users and citizens with complementary tools.

The work combining computer science and law conducted in the DAO projet as well as the main conclusions of the project have also been presented in interdisciplinary colloquium by Sébastien Gambs and Maryline Boizard [48], [47].

7.2.2. Private and Secure Location-based Services

Mobility has always been an important aspect of human activities. Nowadays problems of congestion in urban areas due to the massive usage of cars, last-minutes travel needs and progress in information and communication technologies encourage the rise of new transport modes. Among those are carpooling services, which let car owners share the empty seats of their cars with other travellers having the same travel destination. However, the way carpooling services are implemented today raises several privacy issues. In a recent paper, together with researchers from LAAS-CNRS we have proposed to use privacy enhancing technologies to improve the quality of carpooling services by specially taking in consideration privacy aspects [46].

In addition, publishing directly human mobility data raises serious privacy issues due to its inference potential, such as the (re-)identification of individuals. To address these issues and to foster the development of such applications in a privacy-preserving manner, we propose in a recent paper [26] a novel approach in which Call Detail Records (CDRs) are summarized under the form of a differentially-private Bloom filter for the purpose of privately estimating the number of mobile service users moving from one area (region) to another in a given time frame. Our sanitization method is both time and space efficient, and ensures differential privacy while solving the shortcomings of a solution recently proposed. We also report on experiments conducted using a real life CDRs dataset, which show that our method maintains a high utility while providing strong privacy.

Finally, in authentication protocols, a relay attack allows an adversary to impersonate a legitimate prover, possibly located far away from a verifier, by simply forwarding messages between these two entities. The effectiveness of such attacks has been demonstrated in practice in many environments, such as ISO 14443-compliant smartcards and car-locking mechanisms. Distance-bounding (DB) protocols, which enable the verifier to check his proximity to the prover, are a promising countermeasure against relay attacks. In such protocols, the verifier measures the time elapsed between sending a challenge and receiving the associated response of the prover to estimate their proximity. So far, distance bounding has remained mainly a theoretical concept. Indeed in practice, up to our knowledge only three ISO 14443-compliant implementations of DB protocols exist. The first two are implemented on proprietary smartcards while the last one is available on a highly-customized and dedicated hardware. In a recent paper [35], we demonstrated a proof-of-concept implementation of the Swiss-Knife DB protocol on smartphones running in RFID-emulation mode. To our best knowledge, this is the first time that such an implementation has been performed. Our experimental results are encouraging as they show that relay attacks introducing more than 1.5 ms are directly detectable (in general off-the-shelf relay attacks introduce at least 10 ms of delay). We also leverage on the full power of the ISO-DEP specification to implement the same protocol with 8-bit challenges and responses, thus reaching a better security level per execution without increasing the possibility of relay attacks. The analysis of our results leads to new promising research directions in the area of distance bounding.

7.3. Trust

Reputation mechanisms allow users to mutually evaluate their trust. This is achieved through the computation of a reputation score summarizing their past behaviors. Depending on these scores, users are free to accept or refuse to interact with each other. Existing solutions often rely on costly cryptographic tools that may lead to impractical solutions. We have proposed in [41], [40], [28] usable privacy preserving reputation mechanisms. These mechanisms are distributed and handles non-monotonic ratings. Evaluation made on our mechanism reveals it to be fully usable even with cheap on-board computers. This is a very encouraging result as it shows

that privacy does not impede utility and accuracy. This has been achieved by combining distributed algorithms and cryptographic schemes. Our mechanism is independent of the reputation model, that is, our system can integrate any reputation model, preferably one using both positive and negative ratings.

In a mobile ad hoc network we have also considered the problem of designing a reputation system that allows to update and to propagate the computed reputation scores while tolerating Byzantine failures [42]. Each time a correct node uses directly a service, it can determine by itself the quality of service currently provided. This fresh and valid rating information is broadcast immediately to all its current neighbors. Then, while the mobile node moves, it can receive from other nodes other recommendations also related to the same service. Thus it updates continuously its own opinion. Meanwhile it continues to broadcast this updated information. The freshness and the validity of the received/sent information become questionable. We propose a protocol that allows a node to ignore a second hand information when this information is not fresh or not valid. In particular, fake values provided by Byzantine nodes are eliminated when they are not consistent with those gathered from correct nodes. When the quality of service stabilizes, the correct nodes are supposed to provide quite similar recommendations. In this case, we demonstrate that the proposed protocol ensures convergence to a range of possible reputation scores if a necessary condition is satisfied by the mobile nodes. Simulations are conducted in random mobility scenarios. The results show that our algorithm has a better performance than typical methods proposed in previous works.

7.4. Other Topics Related to Security or Distributed Computing

7.4.1. Detection of distributed denial of service attacks

A Denial of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Denial of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term “iceberg” has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs [29]. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modeled as items) prior to sending them to the server, and throw away all the items that appear with a small probability p_i, and such that the sum of these small probabilities is modeled by probability p_0. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the time needed to collect c distinct items among n frequent ones. A thorough analysis of the time needed to collect c distinct items appears in [16], [15].

7.4.2. Metrics Estimation on Very Large Data Streams

Huge data flows have become very common in the last decade. This has motivated the design of online algorithms that allow the accurate estimation of statistics on very large data flows. A rich body of algorithms and techniques have been proposed for the past several years to efficiently compute statistics on massive data streams. In particular, estimating the number of times data items recur in data streams in real time enables, for example, the detection of worms and denial of service attacks in intrusion detection services, or the traffic monitoring in cloud computing applications. Two main approaches exist to monitor in real time massive data streams. The first one consists in regularly sampling the input streams so that only a limited amount of data items is locally kept. This allows to exactly compute functions on these samples. However, accuracy of this computation with respect to the stream in its entirety fully depends on the volume of data items that has been

sampled and their order in the stream. In contrast, the streaming approach consists in scanning each piece of data of the input stream on the fly, and in locally keeping only compact synopses or *sketches* that contain the most important information about these data. This approach enables us to derive some data streams statistics with guaranteed error bounds without making any assumptions on the order in which data items are received at nodes. Sketches highly rely on the properties of hashing functions to extract statistics from them. Sketches vary according to the number of hash functions they use, and the type of operations they use to extract statistics. The *Count-Min sketch* algorithm proposed by Cormode and Muthukrishnan in 2005 so far predominates all the other ones in terms of space and time needed to guarantee an additive ε-accuracy on the estimation of item frequencies. Briefly, this technique performs t random projections of the set of items of the input stream into a much smaller co-domain of size k, with $k = \lceil e/\epsilon \rceil$ and $t = \lceil \log(1/\delta) \rceil$ in which $0 < \epsilon, \delta < 1$. The user defined parameters ϵ and δ represent respectively the accuracy of the approximation, and the probability with which the accuracy holds. However, because k is typically much smaller than the total number of distinct items in the input stream, hash collisions do occur. This affects the estimation of item frequency when the size of the stream is large. In this work, we have proposed an alternative approach to reduce the impact of collisions on the estimation of item frequency. The intuition of our idea is that by keeping track of the most frequent items of the stream, and by removing their weight from the one of the items with which these frequent items collide, the over-estimation of non frequent items is drastically decreased [21].

We have also proposed a metric, called codeviation, that allows to evaluate the correlation between distributed streams [27]. This metric is inspired from classical metric in statistics and probability theory, and as such allows us to understand how observed quantities change together, and in which proportion. We then propose to estimate the codeviation in the data stream model. In this model, functions are estimated on a huge sequence of data items, in an online fashion, and with a very small amount of memory with respect to both the size of the input stream and the values domain from which data items are drawn. We give upper and lower bounds on the quality of the codeviation, and provide both local and distributed algorithms that additively approximates the codeviation among n data streams by using a sublinear number of bits of space in the size of the domain value from which data items are drawn, and the maximal stream length. To the best of our knowledge, such a metric has never been proposed so far.

7.4.3. Stream Processing Systems

Stream processing systems are today gaining momentum as a tool 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. 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 operator's ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource utilization and thus larger throughputs and reduced tuple processing latencies. We have proposed a new key grouping technique targeted toward applications working on input streams characterized by a skewed value distribution [44]. Our solution is based on the observation that when the values used to perform the grouping have skewed frequencies, e.g. they can be approximated with a Zipfian distribution, the few most frequent values (the *heavy hitters*) drive the load distribution, while the remaining largest fraction of the values (the *sparse items*) appear so rarely in the stream that the relative impact of each of them on the global load balance is negligible. We have shown, through a theoretical analysis, that our solution provides on average near-optimal mappings using sub-linear space in the number of tuples read from the input stream in the learning phase and the support (value domain) of the tuples. In particular this analysis presents new results regarding the expected error made on the estimation of the frequency of heavy hitters.

7.4.4. Randomized Message-Passing Test-and-Set

In [30], we have presented a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes,

returns yes to a unique process and returns no to all the others. Recently many advances in implementing Test&Set objects have been achieved, however all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number p $\leq$ n of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two groups of processes. We propose a message-passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

7.4.5. Population Protocol Model

The population protocol model, introduced by Angluin et his colleagues in 2006, provides theoretical foundations for analyzing global properties emerging from pairwise interactions among a large number of anonymous agents. In the population protocol model, agents are modeled as identical and deterministic finite state machines, *i.e.* each agent can be in a finite number of states while waiting to execute a transition. When two agents interact, they communicate their local state, and can move from one state to another according to a joint transition function. The patterns of interaction are unpredictable, however they must be fair, in the sense that any interaction that should possibly appear cannot be avoided forever. The ultimate goal of population protocols is for all the agents to converge to a correct value independently of the interaction pattern. Examples of systems whose behavior can be modeled by population protocols range from molecule interactions of a chemical process to sensor networks in which agents, which are small devices embedded on animals, interact each time two animals are in the same radio range.

In this work, we focus on an quite important related question. Namely, is there a population protocol that exactly counts the difference κ between the number of agents that initially set their state to A and the one that initially set it to B, and can it be solved in an efficient way, that is with the guarantee that each agent should converge to the exact value of κ after having triggered a sub-linear number of interactions in the size of the system [43].

We answer this question by the affirmative by presenting a $O(n^{3/2})$-state population protocol that allows each agent to converge to the exact solution by interacting no more than $O(\log n)$ times. The proposed protocol is very simple (as is true for most known population protocols), but is general enough to be used to solve different types of tasks.

COAST Project-Team

5. New Results

5.1. Probabilistic Partial Orderings

Participants: Jordi Martori Adrian, Pascal Urso.

Ordering events in a distributed system fundamentally consists in delaying event delivery. Partial ordering, such as FIFO and causal order, has many usage in practical distributed and collaborative systems and can be obtained in arbitrarily large and dynamic networks. However, partial orderings imply that messages cannot be sent and delivered as soon as produced.

In [14], we study the latency induced by such partial orderings. We obtain a probabilistic measure of the moment a message can be delivered according the different characteristics of the distributed system. Having such a measure helps to understand the systems behaviour and to design new protocols. For instance, our measure allows us to parametrize a naive, albeit efficient, fault-tolerant causal delivery mechanism. We experimentally validate our approach using Internet-scale production distribution latency including faults.

5.2. Effect of Delay on Group Performance

Participants: François Charoy, Claudia-Lavinia Ignat [contact], Gérald Oster.

We continued our work on studying the effect of delay in real-time collaborative editing. Delays exist between the execution of one user's modification and the visibility of this modification to the other users. Such delays are in part fundamental to the network, as well as arising from the consistency maintenance algorithms and underlying architecture of collaborative editors. Existing quantitative research on collaborative document editing does not examine either concern for delay or the efficacy of compensatory strategies.

In [12] we studied a collaborative note taking task where we introduced simulated delay. The study was done with 20 groups of 4 users which were asked to listen to a short interview and take notes. We found out a general effect of delay on performance related to the ability to manage redundancy and errors across the document. We interpret this finding as a compromised ability to maintain awareness of team member activity, and a reversion to independent work. Measures of common ground in accompanying chat indicate that groups with less experienced team members attempt to compensate for the effect of delay. In contrast, more experienced groups do not adjust their communication in response to delay, and their performance remains sensitive to the delay manipulation. Results of this study support our team assertion that delay associated with conventional consistency maintenance algorithms will impede group performance. Therefore, these results promote the use of novel algorithms such as CRDTs and motivate the pursuance of research and development on these approaches.

5.3. A CRDT Supporting Selective Undo for Collaborative Text Editing

Participants: Luc André, Claudia-Lavinia Ignat [contact].

Selective undo is an important feature in collaborative editors. With selective undo, a user can undo an earlier operation, regardless of when and where the operation was generated. Current systems that support selective undo are subject to two main limitations. Firstly, they only support undo of operations on atomic objects (e.g. characters or un- breakable lines). In the case of string-wise operations such as copy-paste, find-replace or select-delete, users can typically only undo earlier operations character by charac- ter. Secondly, selective undo may lead to undesirable effects. For example, a user first inserts a misspelled word and then makes a correction. The correction depends on the first insertion of the word. It is undesirable to undo the insertion alone and leave the correction behind as a groundless modification. In [15] we proposed a novel consistency maintenance approach relying on a layered commutative replicated data type (CRDT) that supports selective undo of string-wise operations in collaborative editing. This is the first work that manages undesirable effects of undo. Our performance study shows that it provides sufficient responsiveness to the end users.

5.4. A Trust-Based Formal Delegation Framework for Enterprise Social Networks

Participants: Ahmed Bouchami, Olivier Perrin [contact].

Collaborative environments raise major challenges to secure them. These challenges increase when it comes to the domain of Enterprise Social Networks (ESNs) as ESNs aim to incorporate the social technologies in an organization setup while asserting greater control of information security. In this context, the security challenges have taken a new shape as an ESN may not be limited to the boundaries of a single organization and users from different organizations can collaborate in a common federated environment.

We address the problem of the authorization's delegation in federated collaborative environments like ESNs with an approach based on event-calculus, a temporal logic programming formalism. While the traditional approaches are either user-centric or organization-centric, the approach bridges the gap between these two views and the proposed framework enhances the delegation scheme. We have proposed a behavior monitoring mechanism, that permits to assess principals' trust level within the federated collaborative environment [10].

5.5. Risk Management in the Cloud. Application to Business Process Deployment

Participants: Claude Godart [contact], Elio Goettelmann.

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

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

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

Elio Goettelmann has defended his PhD thesis entitled "Risk-aware Business Process Modeling and Trusted Deployment in the Cloud" on October 2015 [1] based on this result. This framework has been combined to an access control model for strengthening access controls in the context of a collaborative federation of components [9].

5.6. Secure Business Process Deployment in SaaS Contexts

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

Business process (BP) stakeholders want the benefits of the cloud, but they are also reluctant to expose their BP models which express the know-how of their companies. To prevent such a know-how exposure, we are developing a design-time approach for obfuscating a BP model by splitting its model into a collaboration of BP fragments semantically equivalent to the initial BP. This breaking down renders the discovery of the deep content of a critical fragment or of the whole process semantics, by cloud providers much harder when these fragments are deployed in a multi-cloud context. While existing contributions on this topic remain at the level of principles, we propose an algorithm supporting such a BP model transformation [11]. To validate this approach, we are developing a new metric of obfuscation. Complementary to obfuscation, we are developing techniques to reuse, at design time, business process fragments from the cloud, but with limited security risks [8].

5.7. Web Services Selection with QoS

Participants: Amina Ahmed Nacer, Kahina Bessai, Claude Godart [contact], Samir Youcef.

The development of the web technologies and the increase of available services raise the issue of the selection of the most appropriate service among a set of candidate web services. First of all, the services offering a given functionality are discovered. Then, the service section process assists users in choosing the services that better meets their preferences. These preferences are generally expressed as potentially objective functions often conflicting.

Most of existing works trying to select the best web services are based either on a single evaluation criterion or, at best, on the use of an aggregation function like weighted sum of several quantitative evaluation criteria, or the use of the Pareto optimality notion.

In this work, we address some shortcomings of existing approaches by introducing a new optimality notion based on two tests: (i) concordance and (ii) discordance tests. It presents an efficient algorithm to select only the best services using the introduced optimality notion. Moreover, the proposed algorithm exhibits encouraging results as supported by a series of experiments [7].

CTRL-A Team

7. New Results

7.1. Discrete control and reactive language support

Participants: Gwenaël Delaval, Eric Rutten, Stéphane Mocanu, Alia Hajjar, Abdoul-Razak Hassimi Harouna.

Concerning language support, we have designed and implemented BZR, a mixed imperative/declarative programming language: declarative contracts are enforced upon imperatively described behaviors (see 6.1). The semantics of the language uses the notion of Discrete Controller Synthesis (DCS) [5]. This work is done in close cooperation with the Inria team Sumo at Inria Rennes (H. Marchand).

New results concern the master internship of Alia Hajjar, co-directed by Gwenaël Delaval an Stéphane Mocanu, on the subject of Application of control of reactive environments and probabilistic models on Transactional Memory. Multiprocessor environments which use concurrent programs and data structures showed the need of techniques to organize the usage of the shared structures, to reduce the unpredicted delay and reduce the contention between concurrent processors. Transactional Memory (TM) is a programming model that eases development of concurrent applications. Concurrent programming causes conflicts and TM is a way to resolve these conflicts with the transaction paradigm. To control conflict, techniques are provided to optimize (identify the best) degree of parallelism. In this framework, the aim is to control the TM system by adapting the degree of parallelism in order to maximize the throughput, i.e., number of committed transactions per time. The main objective is to minimize the execution time of a parallel application, thus maximize the throughput. During this master's thesis, the behavior of a multithreaded TM environment has been modeled as a stochastic discrete event system. The Heptagon/BZR language has then been used to implement this model for simulation, and evaluation of control strategies.

Ongoing work concerns aspects of compilation and debugging and exploring the notion of adaptive discrete control, which is yet an open question in discrete control in contrast to the well-known adaptive continuous control.

Another activity related to discrete control is or work with Leiden University and CWI (N. Khakpour, now at Linnaeus U., and F. Arbab) on enforcing correctness of the behavior of an adaptive software system during dynamic adaptation is an important challenge along the way to realize correct adaptive systems [11].

7.2. Design and programming

7.2.1. Component-based approaches

Participants: Frederico Alvares de Oliveira Junior, Eric Rutten.

Architecting in the context of variability has become a real need in today's software development. Modern software systems and their architecture must adapt dynamically to events coming from the environment (e.g., workload requested by users, changes in functionality) and the execution platform (e.g., resource availability). Component-based architectures have shown to be very suited for self-adaptation especially with their dynamical reconfiguration capabilities. However, existing solutions for reconfiguration often rely on low level, imperative, and non formal languages. We have defined Ctrl-F, a domain-specific language whose objective is to provide high-level support for describing adaptation behaviors and policies in component-based architectures. It relies on reactive programming for formal verification and control of reconfigurations. We integrate Ctrl-F with the FraSCAti Service Component Architecture middleware platform, and apply it to the Znn.com self-adaptive case study [20], [15], [14], [18].

We work on the topic in cooperation with the Spirals Inria team at Inria Lille (L. Seinturier). It constitutes a follow-up on previous work in the ANR Minalogic project MIND, industrializing the Fractal component-based framework, with a continuation of contacts with ST Microelectronics (V. Bertin). Our integration of BZR and Fractal [4], [2] is at the basis of our current work.

7.2.2. Rule-based systems

Participants: Adja Sylla, Eric Rutten.

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

The transactional nature of the LINC platform insures the correct execution of each of the rules constituting the program, but there still is a need to insure the safety of their coordination, and of the behavior resulting from their sequential execution. For example, in the smart environments application domain, we must insure safety of control decisions, so that all the configurations that can be reached are safe, as well as the sequences of actions in switching between them. For this we will rely on automata-based models and control, using the BZR language, and integrating it in a domains specific language. Our work builds upon preliminary results involving colored Petri nets models [17].

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

7.3. Infrastructure-level support

We apply the results of the previous axes of the team's activity to a range of infrastructures of different natures, but sharing a transversal problem of reconfiguration control design. From this very diversity of validations and experiences, we draw a synthesis of the whole approach [13], towards a general view of Feedback Control as MAPE-K loop in Autonomic Computing [21].

7.3.1. Autonomic Cloud and Big-Data systems

7.3.1.1. Coordination in multiple-loop autonomic Cloud systems

Participants: Soguy Gueye, Gwenaël Delaval, Eric Rutten.

Complex computing systems are increasingly self-adaptive, with an autonomic computing approach for their administration. Real systems require the co-existence of multiple autonomic management loops, each complex to design. However their uncoordinated co-existence leads to performance degradation and possibly to inconsistency. There is a need for methodological supports facilitating the coordination of multiple autonomic managers. To tackle this problem, we take a global view and underscore that Autonomic Management Systems (AMS) are intrinsically reactive, as they react to flows of monitoring data by emitting flows of reconfiguration actions. Therefore we propose a new approach for the design of AMSs, based on synchronous programming and discrete controller synthesis techniques. They provide us with high-level languages for modeling the system to manage, as well as means for statically guaranteeing the absence of logical coordination problems. Hence, they suit our main contribution, which is to obtain guarantees at design time about the absence of logical inconsistencies in the taken decisions. We detail our approach, illustrate it by designing an AMS for a realistic multi-tier application, and evaluate its practicality with an implementation [10].

In order to coordinate managers without breaking their natural modularity. we address the problem with a method stressing modularity, and focusing on the discrete control of the interactions of managers. We make proposals for the distributed execution of modular controllers, first in synchronized way, and then relaxing this synchronization. We apply and validate our method on a multi-loop multi-tier system in a data-center [16].

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

7.3.1.2. Control for Big data

Participants: Bogdan Robu [Gipsa-lab], Mihaly Berekmeri [Gipsa-lab], Nicolas Marchand [Gipsa-lab].

Companies have a fast growing amounts of data to process and store, a data explosion is happening next to us. Currently one of the most common approaches to treat these vast data quantities is the MapReduce parallel programming paradigm. While it?s use is widespread in the industry, ensuring performance constraints, while also minimizing costs, provides considerable challenges. To deal with these issues we propose a control theoretical approach, based on techniques that have already proved their usefulness in the control community. We developed an algorithm to create the first linear dynamic model for a Big Data MapReduce Cloud system, running a concurrent workload. Furthermore we identify two important control use cases: relaxed performance - minimal resource and strict performance. We developed the first feedback control mechanism for such systems. Then to minimize the number of control actuations, an event-based feedback controller was also introduced. Furthermore to address the strict performance challenges a feedforward controller that efficiently suppresses the effects of large workload size variations is developed. On top of this issues an optimal predictive control which deals with concurrent objectives (dependability and performance) is implemented. The approach is validated online in a benchmark running in a real 60 node MapReduce cluster, using a data intensive Business Intelligence [22], [23].

This work is performed in cooperation with LIG (S. Bouchenak) in the framework of the PhD of M. Berekmeri.

7.3.2. Reconfiguration control in DPR FPGA

Participant: Eric Rutten.

Dynamically reconfigurable hardware has been identified as a promising solution for the design of energy efficient embedded systems. However, its adoption is limited by the costly design effort including verification and validation, which is even more complex than for non dynamically reconfigurable systems. We worked on this topic in the context of a ensign environment, developed in the framework of the ANR project Famous, in cooperation with LabSticc in Lorient and Inria Lille (DaRT team) [12]. We proposed a tool-supported formal method to automatically design a correct-by-construction control of the reconfiguration. By representing system behaviors with automata, we exploit automated algorithms to synthesize controllers that safely enforce reconfiguration strategies formulated as properties to be satisfied by control. We design generic modeling patterns for a class of reconfigurable architectures, taking into account both hardware architecture and applications, as well as relevant control objectives. We validate our approach on two case studies implemented on FPGAs [1].

We are currently valorizing results in more publications [12], [9], and extending the use of control techniques by evaluating the new tool ReaX developed at Inria Rennes (Sumo).

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

7.3.3. Autonomic memory management in HPC

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

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

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

7.3.4. Control of smart environments

Participants: Adja Sylla, Mengxuan Zhao, Eric Rutten, Hassane Alla [Gipsa-lab].

7.3.4.1. Generic supervision architecture

New application domains of control, such as in the Internet of Things (IoT) and Smart Environments, require generic control rules enabling the systematization and the automation of the controller synthesis. We worked on an approach for the generation of Discrete Supervisory Controllers for these applications. A general modeling framework is proposed for the application domain of smart home. We formalize the design of the environment manager as a Discrete Controller Synthesis (DCS) problem, w.r.t. multiple constraints and objectives, for example logical issues of mutual exclusion, bounding of power peaks. We validate our models and manager computations with the BZR language and an experimental simulator This work was performed in cooperation with Orange labs (G. Privat) in the framework of the Cifre PhD of M. Zhao [8].

7.3.4.2. Rule-based specification

In the context of IoT applications like mart home environments, the rules for programming in the LINC framework are used as a flexible tool to govern the relations between sensors and actuators. Runtime coordination and formal analysis becomes a necessity to avoid side effects mainly when applications are critical. In cooperation with CEA LETI/DACLE, we are working on a case study for safe applications development in IoT and smart home environments [17].

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

MIMOVE Team

7. New Results

7.1. Introduction

MiMove's research activities in 2015 have focused on a set of areas directly related to the team's research topics. Hence, we have worked on QoS for Emergent Mobile Systems (§ 7.2) in relation to our research topic regarding Emergent Mobile Distributed Systems (§ 3.2). Furthermore, our effort on SoundCity (§ 7.3) is linked to our research on Mobile Social Crowd-sensing (§ 3.4). Still in the context of Mobile Social Crowd-sensing (§ 3.4), we have developed AppCivist-PB (§ 7.4) related to our interest in social applications aiming to actively involve citizens (see § 4.1); this is further linked to our research on composition of Emergent Mobile Distributed Systems (§ 3.2).

7.2. QoS for Emergent Mobile Systems

Participants: Georgios Bouloukakis, Nikolaos Georgantas, Rachit Agarwal, Valérie Issarny, Raphael de Aquino Gomes.

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

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

- While VSB ensures functional interoperability of heterogeneous choreography interactions, differences in timing requirements and constraints of such interactions can severely affect their latencies and success rates. To model timeliness, we introduce the *lease* and *timeout* parameters. The former captures data availability and validity in time, while the latter represents intermittent availability of data recipients due to mobility and disconnection. By precisely studying the related timing thresholds using timed automata models, we verify conditions for successful interactions with VSB connectors. Furthermore, we statistically analyze through simulations, the effect of varying lease and timeout periods to ensure higher probabilities of successful interactions. Simulation experiments are compared with experiments run on the VSB implementation testbed to evaluate the accuracy of results. This work can provide application developers with precise design time information when setting these timing thresholds in order to ensure accurate runtime behavior [23].
- Choreography peers deployed in mobile environments are typically characterized by intermittent connectivity and asynchronous reception of data. In such environments, it is essential to guarantee acceptable levels of timeliness between the data sources and mobile users. In order to provide QoS guarantees in different application scenarios and contexts, it is necessary to model the system performance by incorporating the intermittent connectivity. Queueing Network Models (QNMs) offer a simple modeling environment, which can be used to represent various application scenarios, and provide accurate analytical solutions for performance metrics, such as system response time. We provide an analytical solution regarding the end-to-end response time between the users and the

data sources by modeling the intermittent connectivity of mobile users with product-form QNMs. We utilize the publish/subscribe middleware as the underlying communication infrastructure for the mobile users. To represent the subscriber's connections/disconnections, we model and solve analytically an ON/OFF queueing system by applying a mean value approach. Finally, we validate our model using both simulations with real-world workload traces and comparison with an actual implementation of a Java Messaging Service middleware. The deviations between the performance results foreseen by the analytical model and the ones provided by the simulator and the prototype implementation of a real system are shown to be less than 5% for a variety of scenarios.

- Large-scale mobile environments are characterized by, among others, a large number of mobile users, intermittent connectivity and non-homogeneous arrival rate of data to the users, depending on the region's context. Multiple application scenarios in major cities need to address the above situation for the creation of robust mobile systems. Towards this, it is fundamental to enable system designers to tune a communication infrastructure using various parameters depending on the specific context. We take a first step towards enabling an application platform for large-scale information management relying on mobile social crowd-sourcing [26]. To inform the stakeholders of expected loads and costs, we model a large-scale mobile pub/sub system as a queueing network. We introduce additional timing constraints such as (i) mobile user's intermittent connectivity period; and (ii) data validity lifetime period (e.g. that of sensor data). Using our MobileJINQS simulator (http://xsb.inria.fr/d4d#mobilejinqs), we parameterize our model with realistic input loads derived from the D4D CDR (Call Detail Record) dataset (http://www.d4d.orange.com/en/home) and varied lifetime periods in order to analyze the effect on response time. This work provides system designers with coarse grain design time information when setting realistic loads and time constraints [18].

7.3. Urban Civics: An IoT Middleware for Democratizing Crowdsensed Data in Smart Societies

Participants: Valérie Issarny, Fadwa Rebhi, Animesh Pathak, Sara Hachem.

The growth of our cities comes along with the aggravation of urban nuisances (e.g., air pollution), which significantly alters the citizens' quality of life and especially their health. It then becomes essential to ensure the growth of cities is both environmentally and socially sustainable. As computer scientists, it is our vision that ICT shall play a key role in achieving the above sustainability requirements, as already put forward by the smart city/society concept. However, smart cities have mostly emphasized the big data dimension and related knowledge engineering to ease the management of the city's infrastructure and resources. While this is an important part of smart cities, we believe that ICT should be leveraged to promote participatory democracy so that citizens and government can communicate openly about the issues facing their societies as much as about their solutions. Toward that goal, we have introduced the Urban Civics middleware, which addresses three complementary research questions underlying participatory democracy from an ICT perspective [20], [21]:

(RQ1) How to leverage the richness of urban sensors of the new digital era that features the Internet of Things, open data, social networking, and mobile computing to serve both citizens and government with better insights? Our answer lies in connecting those various data sources where probabilistic protocols combined with semantic technology allow for an urban-scale middleware solution.

(RQ2) How to assimilate urban data so as to generate explanatory city models to inform urban problem solving? Our solution leverages data assimilation (developed by the Inria CLIME team) that has proven successful in geosciences and paves the way to the comprehensive integration of heterogeneous data sources whose accuracy may vary significantly.

(RQ3) How to integrate the solutions to the above into a scalable urban middleware and further ensure citizen participation? Building on our past experience in developing middleware solutions for the mobile environment and especially the – mobile – Internet of Things, we have conceived and introduced the architecture of Urban Civics, a novel IoT middleware solution for democratizing

crowd-sensed data in smart societie. We are in particular confident that, in addition to leveraging existing incentive mechanisms, the citizen participation will also be prompted by the very nature of participatory democracy. However, such an assumption needs to be validated through actual experiments at an urban scale for which we deploy use cases in the Paris and San Francisco Bay areas.

7.4. AppCivist: Engineering Software Assemblies for Participatory Democracy

Participants: Valérie Issarny, Cristhian Parra Trepowski, Animesh Pathak.

Information and communication technologies (ICT) are profoundly changing the nature of human social and environmental interactions. One such change concerns innovations in the way that citizens both interact with government institutions and engage in greater self-government through democratic assembly and collective action. Our research focuses on this transformation of politics, asking how new social media can contribute to new forms of democracy. The pervasive use of ICT suggests that they present an unprecedented opportunity to rethink the constraints of time and space that are generally thought to make the exercise of a more direct and engaging democracy at a large scale practically impossible. In effect, ICT challenge the assumption that citizens of large political units must be content with systems of representative democracy that typically produce a more passive and legalistic citizenship than an active and participatory one.

To consider this challenge, we undertake a pragmatic and modest investigation of how ICT and more precisely software systems can contribute to enabling direct democracy at a large scale. Our research has two immediate objectives. One is to engineer software that leverages the reach of the Internet and the powers of computation to enhance the experience and efficacy of civic participation. The second is to use the ICT software platform to induce the associational forms of a new digitally-inspired citizenship among residents.

Our research is multi-disciplinary in nature, bringing together anthropologists and computer scientists to coinvestigate how to build software systems that promote the development of such digital democratic assemblies and citizens. Our initiative is further rooted in the principles of social activism in that we want to provide citizens with new software systems that help them articulate projects, deliberate directly among themselves, and mobilize activities. A number of digital tools and in particular social networks and web-based content management systems already support aspects of social activism. However, these tools need to be customized as much as composed to become really useful for activists. To that end, we have set the principles of the AppCivist service-oriented software platform in [24]. AppCivist is built around the vision of letting activist users compose their own applications, called Assemblies, using relevant Internet-based components that enable various aspects of democratic assembly and collective action. Starting from a social science perspective, we identified the following high-level categories of functions for AppCivist Assemblies: Mobilizing people, Co-creating proposals, Acting collectively, and Communicating.

Following, we have concentrated on developing the first instance of AppCivist for Participatory Budgeting (PB), as a representative use case of participatory democracy. As a result, we are able to account for various initiatives in citizen participation, including lessons learned from existing PB campaigns worldwide since their emergence in Brazil in the late 1980s. Research contributions more specifically relate to [22]:

- *State of the art survey and analysis of software systems that contribute to enabling participatory democracy*, which lacks an adequate bottom-up approach to digital proposal making. Such an approach would allow groups of citizens to self-assemble on the basis of common interests and enable the resulting citizen assemblies to initiate ideas and elaborate on them using convenient assemblies of software services.
- *State of the art survey and analysis of digital tools oriented towards Participatory Budgeting*, where leveraging ICT to enable truly urban-scale participation in PB campaigns remains unrealized. AppCivist-PB utilizes the concepts of *citizen assembly* and *software assembly* to address this challenge.
- *AppCivist-PB software architecture* enabling citizen and software assemblies, which following the design of AppCivist introduced in [24] strictly adheres to the principles of service orientation. In

that framework, citizen assemblies allow registered users and groups of users to self assemble into higher-level groups to coordinate idea generation and to elaborate proposals through versioning. In a complementary way, software assemblies adhere to the well-known principle of service composition, configuring software services and components oriented towards the implementation of functions supporting participatory democracy.

- *AppCivist-PB prototype* permits an early assessment of the effectiveness of AppCivist-PB in supporting actual urban-scale PB campaigns, such as the one of Paris in 2015. In addition, the prototype provides an opportunity to experiment with developing service wrappers to integrate third-party services (e.g., Etherpad.org) into its software assemblies. In the near future, we intend to automate this integration as much as possible, building on our background in the synthesis of mediators [13], [12].

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

MYRIADS Project-Team

7. New Results

7.1. Cloud Resource Management

Participants: Ancuta Iordache, Christine Morin, Ghada Moualla, Guillaume Pierre, Matthieu Simonin, Lodewijck Vogelzang.

7.1.1. Application Performance Modeling in Heterogeneous Cloud Environments

Participants: Ancuta Iordache, Lodewijck Vogelzang, Guillaume Pierre.

Heterogeneous cloud platforms offer many possibilities for applications to make fine-grained choices over the types of resources they execute on. This opens for example opportunities for fine-grained control of the tradeoff between expensive resources likely to deliver high levels of performance, and slower resources likely to cost less. We designed a methodology for automatically exploring this performance vs. cost tradeoff when an arbitrary application is submitted to the platform. Thereafter, the system can automatically select the set of resources which is likely to implement the tradeoff specified by the user. We significantly improved the speed at which the system can characterize the performance of an arbitrary application. A first publication on this topic has been published [26], and a second one is in preparation.

7.1.2. Heterogeneous Resource Management

Participants: Ancuta Iordache, Guillaume Pierre.

During her internship at Maxeler Technologies, Ancuta Iordache developed an original technique for virtualizing FPGAs such that they can be used as high-performance computing devices in cloud infrastructures. Virtual FPGAs can be accessed remotely by any VM in the system. They can span multiple physical FPGA, they are elastic, and they can also be shared between multiple tenants. A publication on this topic is currently under evaluation.

7.1.3. Self-adaptatable Hadoop Virtual Clusters

Participants: Christine Morin, Ghada Moualla, Matthieu Simonin.

In the context of Ghana Moualla's Master internship, we designed the Elastic MapReduce Adaptation (EMRA) system to execute Hadoop MapReduce applications with user-defined deadlines in cloud virtual clusters. EMRA integrates an algorithm to automatically adapt the Hadoop cluster size at runtime in order to meet user-defined deadlines. We proposed an automatic scaling algorithm, which monitors the progress of the Map phase of the application during its execution and estimate if the user-defined deadline can be met. If the current allocated resources are not sufficient to meet the deadline, more resources are provisioned. The adaptation service comprises of three main components: (i) a monitor to check the progress of the running application, (ii) an estimator to predict the time needed to complete the application based on its current progress ; (iii) a controller to adapt the size of the virtual cluster by adding virtual machines as needed. The controller takes into account the start-up overhead of the new virtual machines and the time needed for the VM to fetch their input data from the original nodes over the network in order to start their map tasks. We implemented a prototype of the EMRA system in the context of Sahara, an environment for managing Hadoop virtual clusters on top of OpenStack IaaS clouds. We experimented the EMRA system on Grid'5000 with traditional MapReduce benchmarks. We evaluated the relative error of the estimator, the cost for scaling up or down a virtual cluster and showed that the proposed adaptation algorithm allows user-defined deadlines to be met.

7.2. Distributed Cloud Computing

Participants: Teodor Crivat, Yvon Jégou, Vlad Mirel, Christine Morin, Anne-Cécile Orgerie, Edouard Outin, Nikolaos Parlavantzas, Jean-Louis Pazat, Guillaume Pierre, Aboozar Rajabi, Carlos Ruiz Diaz, Arnab Sinha, Genc Tato, Cédric Tedeschi.

7.2.1. A multi-objective adaptation system for the management of a Distributed Cloud

Participants: Yvon Jégou, Edouard Outin, Jean-Louis Pazat.

In this project, we consider a "Distributed Cloud" made of multiple data/computing centers interconnected by a high speed network and belonging to the same administration domain. Moreover, in the Cloud organization targeted here, the network capabilities can be dynamically configured in order to guarantee QoS for streaming or to negotiate bandwidth for example.

As a first step, we are focusing on a single centralised Cloud.

Due to the dynamic capabilities of the Clouds, often referred to as elasticity, there is a strong need to dynamically adapt both platforms and applications to users needs and environmental constraints such as electrical power consumption.

We address the management of a Cloud in order to consider both optimization for energy consumption and for users' QoS needs. The objectives of this optimization will be negotiated as contracts on Service Level Agreement (SLA). A special emphasis will be put on the distributed aspect of the platform and include both servers and network adaptation capabilities.

The design of the system relies on self-* techniques and on adaptation mechanisms at any level (from IaaS to SaaS). The MAPE-k framework (Monitor-Analysis-Planning-Execution based on knowledge) is used for the implementation of the system. The technical developments are based on the Openstack framework.

We have implemented a system that uses a genetic algorithm to optimize Cloud energy consumption and machine learning techniques to improve the fitness function regarding a real distributed cluster of servers. We have carried out experiments on the OpenStack platform to validate our solution. This experimentation shows that the machine learning produces an accurate energy model, predicting precise values for the simulation.

We are currently refining this model and comparing it to real measurements on the platform.

This work is done in cooperation with the DIVERSE team and in cooperation with Orange under the umbrella of the B-COM Technology Research Center.

7.2.2. Dynamic reconfiguration for multi-cloud applications

Participants: Nikolaos Parlavantzas, Aboozar Rajabi, Carlos Ruiz Diaz, Arnab Sinha.

In the context of the PaaSage European project, we are working on model-based, continuous self-optimization of multi-cloud applications. In particular, we are developing a dynamic adaptation system, capable of transforming the currently running application configuration into a target configuration in a cost-effective and safe manner. In 2015, we have improved and extended the Adapter prototype [45]. The system now fully supports dynamic configuration, including detecting changes, generating reconfiguration plans, validating plans based on a cost-benefit calculation, and executing plans in parallel, improving adaptation performance. Moreover, we have performed initial investigations on the use of PaaSage for supporting Internet of Things (IoT) applications [27]. Finally, in the context of Carlos Ruiz's stay, we are defining a model for managing the configuration of cloud applications and environments. This model is based on feature modeling and the derived configurations are mapped to PaaSage models.

7.2.3. Towards a distributed cloud inside the backbone

Participants: Christine Morin, Anne-Cécile Orgerie, Genc Tato, Cédric Tedeschi.

The DISCOVERY proposal officially started at the end of 2015. It is an Inria Project Lab (IPL) led by Adrien Lebre from the ASCOLA team, and currently on leave at Inria. It aims at designing a distributed cloud, leveraging the resources we can find in the network backbone. [0] In practice, this work is intended to get integrated within the OpenStack softwarehttps://www.openstack.org/ so as to decentralize its whole architecture.

[0]The DISCOVERY website: http://beyondtheclouds.github.io

In this context, and in collaboration with ASCOLA and ASAP teams, we started the design of an overlay network whose purpose is to be able, with a limited cost, to locate geographically-close nodes from any point of the network. In this framework, the PhD thesis of Genc Tato started in December 2015. It aims at developing locality mechanisms at the data management layer.

We have also started an energy/cost-benefit analysis of a decentralized Cloud infrastructure like the one proposed within Discovery. This work is conducted by Anthony Simonet, a post-doctoral researcher on an Inria contract for the Discovery IPL and co-supervised by Adrien Lebre from the ASCOLA team and Anne-Cécile Orgerie from Myriads team.

7.2.4. Mobile edge cloud computing with ConPaaS

Participants: Teodor Crivat, Vlad Mirel, Guillaume Pierre.

Interactive multi-user applications usually rely on intermediate cloud servers to mediate the inter-user interaction. However, current mobile networks exhibit network latencies in the order of 50-150 ms between the device and any cloud. Such latencies make it impossible to create smooth interactions with the end user. To enable an "instantaneous" feeling, augmented reality applications require that end-to-end latencies should remain below 20 ms.

To address these issues, we extended ConPaaS to support the deployment of cloud applications in a distributed set of Raspberry Pi machines. The motivation is to reduce the latency compared to a traditional deployment where the backend is located in an external cloud: instead of reaching the cloud through a wide-area network, in this setup each cloud node is also equipped with a wifi hotspot which allows local users to access it directly.

7.2.5. Fog Computing

Participant: Jean-Louis Pazat.

The concept of "Fog Computing" is currently developed on the idea of hosting instances of services not on centralized datacenters (i.e. the "Cloud"), but on a highly distributed infrastructure: the Internet Edge (i.e. the "Fog"). This infrastructure consists in geographically distributed computing resources with relatively small capabilities. Compared with datacenters, a "Fog" infrastructure is able to offer to Service Providers a shorter distance from the service to the user but with the same flexibility of software deployment and management.

This work focus on the problem of resource allocation in such infrastructure when considering services in the area of Internet of Things, Social Networks or Online Gaming. For such use-cases, service-to-user latency is a critical parameter for the quality of experience. Optimizing such a parameter is an objective for the platform built on top of the Fog Infrastructure that will be dedicated to the deployment of the considered service. In order to achieve such a goal, the platform needs to select some strategies for the allocation of network and computing resources, based on the initial requirements for service distribution.

We are designing a prototype based on micro services and we are considering low overhead virtualization systems using containers. This prototype is intended to run inside an Internet Box or inside a LAN disk server at user's home. The whole system will be intended to be used very small or medium size user communities willing to share devices and data. The main characteristics of the system will be reliable distributed storage and distributed execution of services.

This work is part of Bruno Stevant's PhD thesis, which began in December 2014. It is done in cooperation with the REOP team, Institut Mines telecom/IRISA.

7.3. Cloud Security

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

7.3.1. Security Monitoring of Clouds

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

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 evaluating SLA terms,
- making the SLA terms enforcement mechanisms self-adaptable to cope with the dynamic nature of clouds.

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

In 2015 we started to study the state of the art about SLA for security monitoring in clouds, as well as about evaluating security monitoring setups in clouds.

In 2015 we also studied the self-adaptation issues of security monitoring with two kinds of security monitoring components: a network intrusion detection system (NIDS), and a secured application-level firewall. Moreover a new approach to secure an application-level firewall has been proposed.

To experiment with both kinds of components, a prototype called SAIDS has been implemented in the OpenStack-based IaaS cloud testbed that was setup in 2014. The NIDS software used is Snort. The application-level firewall is based on Linux nftables and Open vSwitch. In order to study more complex security monitoring setups, SAIDS will be extended in 2016.

A preliminary evaluation of SAIDS has been published in the doctoral symposium of CCGrid 2015. A more complete evaluation of SAIDS as well as the evaluation of the application-level firewall will be done in 2016.

7.4. Greening Clouds

Participants: Maria Del Mar Callau Zori, Ismael Cuadrado Cordero, David Guyon, Sabbir Hasan Rochi, Yunbo Li, Christine Morin, Anne-Cécile Orgerie, Jean-Louis Pazat, Guillaume Pierre.

7.4.1. Energy-aware IaaS-PaaS co-design

Participants: Maria Del Mar Callau Zori, Anne-Cécile Orgerie, Guillaume Pierre.

The wide adoption of the cloud computing paradigm plays a crucial role in the ever-increasing demand for energy-efficient data centers. Driven by this requirement, cloud providers resort to a variety of techniques to improve energy usage at each level of the cloud computing stack. However, prior studies mostly consider resource-level energy optimizations in IaaS clouds, overlooking the workload-related information locked at higher levels, such as PaaS clouds. 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. A publication on this topic is in preparation.

7.4.2. Energy-efficient cloud elasticity for data-driven applications

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

Distributed and parallel systems offer to users tremendous computing capacities. They rely on distributed computing resources linked by networks. They require algorithms and protocols to manage these resources in a transparent way for users. Recently, the maturity of virtualization techniques has allowed for the emergence of virtualized infrastructures (Clouds). These infrastructures provide resources to users dynamically, and adapted to their needs. By benefiting from economies of scale, Clouds can efficiently manage and offer virtually unlimited numbers of resources, reducing the costs for users.

However, the rapid growth for Cloud demands leads to a preoccupying and uncontrolled increase of their electric consumption. In this context, we will focus on data driven applications which require to process large amounts of data. These applications have elastic needs in terms of computing resources as their workload varies over time. While reducing energy consumption and improving performance are orthogonal goals, this internship aims at studying possible trade-offs for energy-efficient data processing without performance impact. As elasticity comes at a cost of reconfigurations, these trade-offs will consider the time and energy required by the infrastructure to dynamically adapt the resources to the application needs.

The master internship work of David Guyon on this topic has been presented at IEEE GreenCom 2015 [39]. This work will be continued during David's PhD thesis.

7.4.3. Energy-efficient and network-aware resource allocation in Cloud infrastructures

Participants: Ismael Cuadrado Cordero, Christine Morin, Anne-Cécile Orgerie.

Energy consumption is cloud computing has become a key environmental and economic concern. Our work aims at designing energy-efficient resource allocation for Cloud infrastructures. The ever-growing appetite of new applications for network resources leads to an unprecedented electricity bill, and for these bandwidth-hungry applications, networks can become a significant bottleneck. New algorithms have to be designed integrating the data locality dimension to optimize computing resource allocation while taking into account the fluctuating limits of network resources. Towards this end, we proposed GRaNADA, a semi-decentralized Platform-as-a-service (PaaS) architecture for real-time multiple-users applications. Our architecture geographically distributes the computation among the clients of the cloud, moving the computation away from the datacenter to save energy - by shutting down or downgrading non utilized resources such as routers and switches, servers, etc. - and provides lower latencies for users. GRaNADA implements the concept of micro-cloud, a fully autonomous energy-efficient subnetwork of clients of the same service, designed to keep the greenest path between its nodes. Along with GRaNADA, we proposed DEEPACC, a cloud-aware routing protocol which distributes the connection between the nodes. Our system GRaNADA targets services where the geographical distribution of clients working on the same data is limited - for example, a shared on-line document - or those services where, even if the geographical distribution of clients is high, the upload data communication to the cloud is small - for instance a light social network like Twitter. We compared our approach with two main existing solutions - replication of data in the edge and traditional centralized cloud computing. Our approach based on micro-clouds exhibits interesting properties in terms of QoS and especially latency. Simulations show that, using the proposed PaaS, one can save up to 75% of the spent network energy compared to traditional centralized cloud computing approaches. Our approach is also more energy-efficient than the most popular semi-decentralized solutions, like nano data centers. This work has been presented at IEEE GreenCom 2015 [18].

We also evaluated the suitability of using micro-clouds in the context of smart cities. We investigated the idea to build a local cloud on top of networking resources spread across a defined area and including the mobile devices of the users. This local cloud is managed by lightweight mechanisms in order to handle users who can appear/disappear and move. We used a scenario considering a platform for neighborhood services and showed that micro-clouds make better use of the network, reducing the amount of unnecessary data traveling through external networks. This work is currently under review for a conference.

7.4.4. Resource allocation in a Cloud partially powered by renewable energy sources

Participants: Yunbo Li, Anne-Cécile Orgerie.

We propose here to design a disruptive approach to Cloud resource management which takes advantage of renewable energy availability to perform opportunistic tasks. To begin with, the considered Cloud is mono-site (i.e. all resources are in the same physical location) and performs tasks (like web hosting or MapReduce tasks) running in virtual machines. This Cloud receives a fixed amount of power from the regular electric Grid. This power allows it to run usual tasks. In addition, this Cloud is also connected to renewable energy sources (such as windmills or solar cells) and when these sources produce electricity, the Cloud can use it to run more tasks.

The proposed resource management system needs to integrate a prediction model to be able to forecast these extra-power periods of time in order to schedule more work during these periods. Batteries will be used to guarantee that enough energy is available when switching on a new server working exclusively on renewable energy. Given a reliable prediction model, it is possible to design a scheduling algorithm that aims at optimizing resource utilization and energy usage, problem known to be NP-hard. The proposed heuristics will thus schedule tasks spatially (on the appropriate servers) and temporally (over time, with tasks that can be planed in the future).

This work is done in collaboration with Ascola team from LINA in Nantes. Two publications have been accepted this year on this topic for: SmartGreens 2015 [15] and IEEE GreenCom 2015 [21].

7.4.5. SLA driven Cloud Auto-scaling for optimizing energy footprint

Participants: Sabbir Hasan Rochi, Jean-Louis Pazat.

As a direct consequence of the increasing popularity of Internet and Cloud Computing services, data centers are amazingly growing and hence have to urgently face energy consumption issues. At the Infrastructure-as-a-Service (IaaS) layer, Cloud Computing allows to dynamically adjust the provision of physical resources according to Platform-as-a-Service (PaaS) needs while optimizing energy efficiency of the data center.

The management of elastic resources in Clouds according to fluctuating workloads in the Software-as-a-Service (SaaS) applications and different Quality-of-Service (QoS) end-user's expectations is a complex issue and cannot be done dynamically by a human intervention. We advocate the adoption of Autonomic Computing (AC) at each XaaS layer for responsiveness and autonomy in front of environment changes. At the SaaS layer, AC enables applications to react to a highly variable workload by dynamically adjusting the amount of resources in order to keep the QoS for the end users. Similarly, at the IaaS layer, AC enables the infrastructure to react to context changes by optimizing the allocation of resources and thereby reduce the costs related to energy consumption. However, problems may occur since those self-managed systems are related in some way (e.g. applications depend on services provided by a cloud infrastructure): decisions taken in isolation at given layer may interfere with other layers, leading whole system to undesired states.

We have defined a scheme for green energy management in the presence of explicit and implicit integration of renewable energy in datacenter [13]. More specifically we propose three contributions: i) we introduce the concept of virtualization of green energy to address the uncertainty of green energy availability, ii) we extend the Cloud Service Level Agreement (CSLA) language to support Green SLA introducing two new threshold parameters and iii) we introduce greenSLA algorithm which leverages the concept of virtualization of green energy to provide per interval specific Green SLA. Experiments were conducted with real workload profile from PlanetLab and server power model from SPECpower to demonstrate that, Green SLA can be successfully established and satisfied without incurring higher cost.

This work is done in collaboration with Ascola team from LINA in Nantes.

7.5. Energy-efficient Computing Infrastructures

Participants: Christine Morin, Anne-Cécile Orgerie, Martin Quinson.

7.5.1. Simulating the impact of DVFS within SimGrid

Participants: Christine Morin, Anne-Cécile Orgerie, Martin Quinson.

Simulation is a a popular approach for studying the performance of HPC applications in a variety of scenarios. However, simulators do not typically provide insights on the energy consumption of the simulated platforms. Furthermore, studying the impact of application configuration choices on energy is a difficult task, as not many platforms are equipped with the proper power measurement tools. The goal of this work is to enable energy-aware experimentation within the SimGrid simulation toolkit, by introducing a model of application energy consumption and enabling the use of Dynamic Voltage and Frequency Scaling (DVFS) techniques for the simulated platforms. We provide the methodology used to obtain accurate energy estimations, highlighting the simulator calibration phase. The proposed energy model is validated by means of a large set of experiments featuring several benchmarks and scientific applications. This work is available in the latest SimGrid release. This work is done in collaboration with the Mescal team from LIG in Grenoble. A paper is currently under preparation on this work.

7.5.2. Simulating Energy Consumption of Wired Networks

Participants: Timothée Haudebourg, Anne-Cécile Orgerie.

Predicting the performance of applications, in terms of completion time and resource usage for instance, is critical to appropriately dimensioning resources that will be allocated to these applications. Current applications, such as web servers and Cloud services, require lots of computing and networking resources. Yet, these resource demands are highly fluctuating over time. Thus, adequately and dynamically dimensioning these resources is challenging and crucial to guarantee performance and cost-effectiveness. In the same manner, estimating the energy consumption of applications deployed over heterogeneous cloud resources is important in order to provision power resources and make use of renewable energies. Concerning the consumption of entire infrastructures, some studies show that computing resources represent the biggest part in the Cloud's consumption, while others show that, depending on the studied scenario, the energy cost of the network infrastructure that links the user to the computing resources can be bigger than the energy cost of the servers.

In this work, we aim at simulating the energy consumption of wired networks which receive little attention in the Cloud computing community even though they represent key elements of these distributed architectures. To this end, we are contributing to the well-known open-source simulator ns3 by developing an energy consumption module named ECOFEN.

In 2015, this simulator has been extended to integrate two more green levers: low power idle (IEEE 802.3az) and adaptive link rate. This work has been done during the internship of Timothée Haudebourg (L3 ENS Rennes) and a publication is currently under preparation.

7.5.3. Multicriteria scheduling for large-scale HPC environments

Participant: Anne-Cécile Orgerie.

Energy consumption is one of the main limiting factor for the design and deployment of large scale numerical infrastructures. The road towards "Sustainable Exascale" is a challenge with a target of 50 Gflops per watt. Energy efficiency must be taken into account and must be combined with other criteria like performance, resilience, Quality of Service.

As platforms become more and more heterogeneous (co-processors, GPUs, low power processors...), an efficient scheduling of applications and services at large scale remains a challenge. In this context, we will explore and propose a multicriteria scheduling model and framework for large scale HPC systems. Based on real energy measurements and calibrations, we will propose some performance and energy models and will build a multi criteria scheduler. Simulation on selected scenario will be explored and a prototype will be designed for ensuring experimental validation.

This work is done in collaboration with ROMA and Avalon teams from LIP in Lyon.

7.6. Decentralized and Adaptive workflows

Participants: Jean-Louis Pazat, Javier Rojas Balderrama, Matthieu Simonin, Cédric Tedeschi, Palakiyem Wallah.

7.6.1. Adaptive Workflows with Chemical Computing

Participants: Javier Rojas Balderrama, Matthieu Simonin, Cédric Tedeschi.

We have designed a high-level programming model based on the HOCL rule-based language to express workflow adaptation. It was specifically designed to support changes in the workflow logic at run time. This mechanism was implemented within the GinFlow software and experimented over the Grid'5000 platform. An article was just accepted for publication at the IPDPS 2016 conference.

7.6.2. Best-effort decentralized workflow execution

Participants: Jean-Louis Pazat, Cédric Tedeschi, Palakiyem Wallah.

We are currently proposing a simple workflow model for workflow execution in platforms with limited computing resources and services. The key idea is to devise a best-effort workflow engine that does not require a strong centralized orchestrator. Such a workflow engine relies on point-to-point cooperation between nodes supporting the execution. A minimalistic demonstrator of these concepts has been devised and implemented. Early experiments have been conducted on a single machine.

7.7. Experimental Platforms

Participants: Julien Lefeuvre, David Margery.

7.7.1. Contribution to Fed4FIRE testbed

Participants: Julien Lefeuvre, David Margery.

In Fed4FIRE, two key technologies have been adopted as common protocols to enable experimenters to interact with testbeds: Slice Federation Architecture (SFA), to provision resources, and Control and Management Framework for Networking Testbeds (OMF) to control them. In 2015, the main area of work has been the implementation of an SFA API to BonFIRE, still on-going. In the process, we wrote the reference documentation to write a new delegate for geni-tools, the reference implementation of SFA maintained by the GENI project office. This codebase has now been made public on github, in part because of our interactions with the code and suggested changes to ease writing new delegates. We have also contributed to the design of a service layer proxy mechanisms so that testbeds with http based APIs can be queried by any Fed4FIRE user using a standard authentications mechanism. The BonFIRE API has been made available through that mechanism, based on XML documents signed using the XML Signature specification.

REGAL Project-Team

6. New Results

6.1. Distributed algorithms for dynamic networks

Participants: Luciana Bezerra Arantes [correspondent], Marjorie Bournat, Swan Dubois, Denis Jeanneau, Mohamed Hamza Kaaouachi, Sébastien Monnet, Franck Petit [correspondent], Pierre Sens, Julien Sopena.

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

We have obtained results both on fundamental aspects of distributed algorithms and on specific emerging large-scale applications.

We study various key topics of distributed algorithms: agreement, failure detection, data dissemination and data finding in large scale systems, self-stabilization and self-* services.

6.1.1. Agreement and failure detection in dynamic Distributed Systems

Distributed systems should provide reliable and continuous services despite the failures of some of their components. A classical way for a distributed system to tolerate failures is to detect them and then to recover. It is now well recognized that the dominant factor in system unavailability lies in the failure detection phase. In 2015, we obtain the following results on failure detection:

Assuming a message-passing environment with a majority of correct processes, the necessary and sufficient information about failures for implementing a general state machine replication scheme ensuring consistency is captured by the Ω failure detector. We show in [46] that in such a message-passing environment, Ω is also the weakest failure detector to implement an eventually consistent replicated service, where replicas are expected to agree on the evolution of the service state only after some (a priori unknown) time.

We also study the k-set agreement problem is a generalization of the consensus problem where processes can decide up to k different values. Very few papers have tackled this problem in dynamic networks. Exploiting the formalism of the Time Varying Graph model, we propose in [70]a new quorum-based failure detector for solving k-set agreement in dynamic networks with asynchronous communications. We present two algorithms that implement this new failure detector using graph connectivity and message pattern assumptions. We also provide an algorithm for solving k-set agreement using our new failure detector.

We propose several algorithms to implement efficient failure detection services. We introduce in [60] the Two Windows Failure Detector (2WFD), an algorithm that provides QoS and is able to react to sudden changes in network conditions, a property that currently existing algorithms do not satisfy. We ran tests on real traces and compared the 2W-FD to state-of-the-art algorithms. Our results show that our algorithm presents the best performance in terms of speed and accuracy in unstable scenarios. In [62], we propose a new approach towards the implementation of failure detectors for large and dynamic networks: we study reputation systems as a means to detect failures. The reputation mechanism allows efficient node cooperation via the sharing of views about other nodes. Our experimental results show that a simple prototype of a reputation-based detection service performs better than other known adaptive failure detectors, with improved flexibility. It can thus be used in a dynamic environment with a large and variable number of nodes.

6.1.2. Probabilistic Byzantine Tolerance allocation strategies in Hybrid Cloud Environments

We explore the node allocation challenges in providing probabilistic Byzantine fault tolerance in a hybrid cloud environment, consisting of nodes with varying reliability levels, compute power, and monetary cost. We consider hybrid computing architectures that combine edge nodes with cloud hosted computing. In such a system, a large fraction of the computation is performed by donated machines at the edge of the network, which significantly reduces the cost to the owner of the computation.

Considering "bag of tasks" (BoT) applications where a large computational problem is broken into a large number of independent tasks, the probabilistic Byzantine fault tolerance guarantee refers to the confidence level that the result of a given computation is correct despite potential Byzantine failures. In [36] we explore probabilistic Byzantine tolerance, in which computation tasks are replicated on dynamic replication sets whose size is determined based on ensuring probabilistic thresholds of correctness.

6.1.3. Covering problems in dynamic systems

We study covering problems (such as minimal dominating set or maximal matching) in the context of highly dynamic distributed systems. We first obtain some general results. In [48], we first propose a new definition of this family of problems since classical ones are meaningless in such systems. We generalize the classical definition of time complexity (for static systems) to our setting. We also provided in [40] a generic tool to help the writing of impossibility proofs in dynamic distributed systems. Then, we focus on the particular case of the minimal dominating set problem. We characterize the necessary and sufficient condition to construct deterministically a minimal dominating set in a dynamic system according to our definition.

6.1.4. Self-Stabilization

Self-stabilization is a generic paradigm to tolerate transient faults (*i.e.*, faults of finite duration) in distributed systems. Results obtained in this area by Regal members in 2015 follow.

Spanning tree construction is a well-studied problem in distributed computing for its numerous applications like routing, broadcast...Properties of the obtained trees, efficiency of the construction, and fault-tolerance guarantees are naturally at the heart of many researches. In this context, we propose in [39] a new self-stabilizing algorithm for the minimum diameter spanning tree that achieves better time and space complexity than existing solutions. Moreover, our solution tolerates a fully asynchronous adversary.

A classical way to endowed self-stabilization with (permanent) fault tolerance is *confinement*. That is, we ensure that the self-stabilizing system moreover ensures that the effect of permanent faults is limited to some topological areas of the system. In [27], we propose a characterization of optimal confinement areas for a large set of spanning tree metrics in presence of Byzantine faults. In [24], we propose a stabilizing implementation of an atomic register in presence of crash faults. By avoiding the propagation of fault effects further than a given radius, confinement is clearly a *spatial* approach. Another approach, called *temporal*, consists in recovering as quick as possible to a configuration from which some forms of safety are satisfied.

In [68], we introduce the notion of *gradual stabilization* and provide a gradually self-stabilizing algorithm that solves the *unison* problem, *i.e.*, the problem that consists in synchronizing logical clocks locally maintained by the processes.

6.1.5. Team of Mobile Robots

Swarm of autonomous mobile sensor devices (or, robots) recently emerged as an attractive issue in the study of dynamic distributed systems permits to assess the intrinsic difficulties of many fundamentals tasks, such as exploring or gathering in a discrete space. We consider autonomous robots that are endowed with visibility sensors (but that are otherwise unable to communicate) and motion actuators. The robots we consider are weak, *i.e.*, they are anonymous, uniform, unable to explicitly communicate, and oblivious (they do not remember any of their past actions). Despite their weakness, those robots must collaborate to solve a collective tasks such as exploration, gathering, flocking, to quote only a few.

In [45], we first show that it is impossible to explore any simple torus of arbitrary size with (strictly) less than four robots, even if the algorithm is probabilistic. Next, we propose an optimal (*w.r.t.* the number of robots) solution for the terminating exploration of torus-shaped networks by a team of k such robots in the SSYNC model. The proposed algorithm is probabilistic and works for any simple torus of size $\ell \times L$, where $7 \leq \ell \leq L$. Since the optimal number of robots is also four in rings, our result shows that increasing the number of possible symmetries in the network (due to increasing dimensions) does not necessarily come at an extra cost *w.r.t.* the number of robots that are necessary to solve the problem.

6.2. Management of distributed data

Participants: Rudyar Cortes, Mesaac Makpangou, Olivier Marin, Sébastien Monnet [correspondent], Pierre Sens.

6.2.1. Long term durability and storage load distribution

In 2014, we had proposed SPLAD (for Scattering and PLAcing Data replicas to enhance long-term durability), a model that allows us to vary the data scattering degree by tuning a selection range width. We have enhanced our model [57] and we have focused on the study of the policy used while choosing a storing node within the selection range. Some policies may lead to heavily unbalanced storage load distribution which can be harmful for the system. Simple policies to balance the load (e.g. storing new blocks on least loaded nodes) may induce network congestion and thus data losses. We have shown that the "power of two choices" policy (choosing the least loaded node among two random ones) brings good results both in terms of storage load distribution and fault tolerance.

6.2.2. Management of dynamic big data

Managing and processing Dynamic Big Data, where multiple sources produce new data continuously, is very complex. Static cluster- or grid-based solutions are prone to induce bottleneck problems, and are therefore ill-suited in this context. Our objective in this domain is to design and implement a Reliable Large Scale Distributed Framework for the Management and Processing of Dynamic Big Data. In 2015, we focused on Spatio-temporal range queries over Big Location Data aim to extract and analyze relevant data items generated around a given location and time. They require concurrent processing of massive and dynamic data flows. We proposed a scalable architecture for continuous spatio-temporal range queries built by coalescing multiple computing nodes on top of a Distributed Hash Table. The key component of our architecture is a distributed spatio-temporal indexing structure which exhibits low insertion and low index maintenance costs. We assessed our solution with a public data set released by Yahoo! which comprises millions of geotagged multimedia files [43].

6.3. CISE Logic and tool for proving invariants in distributed databases

Participants: Marc Shapiro [correspondent], Mahsa Najafzadeh, Alexey Gotsman, Carla Ferreira.

We have developed a new sound logic for proving the correctness of a distributed database under concurrent updates, showing whether the application maintains the database's *integrity invariants*. An operation of the application is specified as a *preparator*, which checks the operation's precondition at an origin replica and generates an *effector*. The effector abstracts the update to be applied to every replica. The application also specifies which operations are allowed to take place concurrently. In summary, the logic shows that the application maintains the invariant if the three following rules are satisfied:

- Each operation individually maintains the invariant. It follows that operations' preconditions are sufficiently strong to ensure correctness in a sequential execution.
- The effectors of any two operations that can execute concurrently commute. This implies that the database replicas all converge to the same state.
- For any pair of operations u and v that can execute concurrently, the precondition of u is stable under the effector of v, and vice-versa.

This result is published at POPL 2016 [50].

We have implemented a tool (based on the Z3 SMT solver) that implements these rules. A demo of the tool is available online [78]. If the application passes the tool, it is correct. If not, the tool returns a counter-example, which the application developer can inspect to find the source of the error. Generally speaking, the developer can either weaken the invariants or the effects of operations, or strengthen consistency by disallowing concurrency. By choosing one or the other, the developer performs a co-design of the application with its consistency protocol, in order to have the highest possible concurrency that still ensures correctness.

For instance, consider a database of bank accounts, with the invariant that an account's balance must be positive. The banking application has operations $credit(acct, amt)$, $debit(acct, amt)$, and $accrue - interest(acct)$. The first rule dictates that $debit$ has the precondition $amt = balance$. The second rule dictates that $accrue - interest$ computes the amount of interest according to the state at the origin, not at every replica. The third rule is violated if concurrent $debit$s are allowed; if the bank wishes to uphold the invariant, the only correct solution is to disallow concurrent $debit$s.

6.4. Memory management for big data

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

6.4.1. Automated file cache pooling

Some applications, like online sales servers, intensively use disk I/Os. Their performance is tightly coupled with I/Os efficiency. To speed up I/Os, operating systems use free memory to offer caching mechanisms. Several I/O intensive applications may require a large cache to perform well. However, nowadays resources are virtualized. In clouds, for instance, virtual machines (VMs) offer both isolation and flexibility. This is the foundation of cloud elasticity, but it induces fragmentation of the physical resources, including memory. This fragmentation reduces the amount of available memory a VM can use for caching I/Os. Previously, we proposed Puma (for Pooling Unused Memory in Virtual Machines) which allows I/O intensive applications running on top of VMs to benefit of large caches. This was realized by providing a remote caching mechanism that provides the ability for any VM to extend its cache using the memory of other VMs located either in the same or in a different host.

We have performed an extensive evaluation of Puma [53] and we have enhanced our solution: Puma adapts automatically the amount a memory that a VM offers to another VM. Furthermore, if the network becomes overloaded, Puma detects a performance degradation and stops using a remote cache.

SCALE Team

7. New Results

7.1. Programming Languages for Distributed Systems

7.1.1. Multi-active Objects

Participants: Ludovic Henrio, Justine Rochas, Vincenzo Mastandrea.

The active object programming model is particularly adapted to easily program distributed objects: it separates objects into several *activities*, each manipulated by a single thread, preventing data races. However, this programming model has its limitations in terms of expressiveness – risk of deadlocks – and of efficiency on multicore machines. We proposed to extend active objects with *local multi-threading*. We rely on declarative *annotations* for expressing potential concurrency between requests, allowing easy and high-level expression of concurrency. This year we realized the following:

- We proved the correctness of our compiler from ABS language into ProActive multi-active objects. This translation can be generalised to many other active object languages. This work has been published as a research report, and is under submission to a conference. The proof brought us very deep and interesting understanding on the differences between the languages.
- We started to work on static detection of deadlocks for multi-active object. This is the work of Vincenzo Mastandrea who is starting a Labex PhD in collaboration with the FOCUS EPI (Univ of Bologna). An article is currently submitted to a conference on this subject.
- We are formalising in Isabelle/HOL a first version of the semantics of multiactive objects. This work was done in collaboration with Florian Kammuller
- We organised a workshop on active object languages with the main teams in Europe involved in the development of active-object languages. A journal survey paper on the subject is currently being written.
- We implemented a debugger for multi active object programs.

We plan to continue to improve the model, especially about compile-time checking of annotations and about fault tolerance of multiactive objects.

7.1.2. Behavioural Semantics

Participants: Ludovic Henrio, Eric Madelaine, Min Zhang, Siqi Li.

We are conducting a large study on Parameterised Networks of Automata (pNets) from a theoretical perspective. We started last year with some 'pragmatic' expressiveness of the pNets formalism, showing how to express a wide range of classical constructs of (value-passing) process calculi, but also complex interaction patterns used in modern distributed systems. After publishing those results [13], we focused on open systems and our formalism is able to represent operators of composition of processes, they are represented as hierarchically composed automata with holes and parameters. We defined a semantics for open pNets and a bisimulation theory for them. This study was driven by several usecase examples including a hierarchical broadcast algorithm and several operators of concurrent processes. A short presentation is accepted for publication in the journal "Science China: Information Sciences -". A full paper on the subject of the semantics and bisimulation for open pNets is under submission to a conference.

In parallel we have started the study of a denotational semantics for open pNets, based on the Universal Theory of Processes (UTP). The idea in the long term would be to draw links between the operational, denotational, and algebraic models of the pNet formalism. A short presentation of our preliminary results will be presented at the conference PDP'16 (work in progress session).

7.1.3. GPU-based High Performance Computing for finance

Participants: Michael Benguigui, Françoise Baude.

We have pursued our work on pricing American multi-dimensional (so very computation intensive) options in finance and we hav been able to extend this to the computation of Value At Risk (consists in repeating the American option pricing, but we have found a financial grounded optimization that avoids us to replicate the most time consuming phase).

Moreover, the balancing of work is taking in consideration the heterogeneous nature of the involved GPUs, and is capable to harness the computing power of multi-core CPUs that also support running OpenCL codes. As our scheduling solution is capable to get a reasonable prediction of the workload of each slave computation, we have leveraged this to run the whole pricing and VaR computations onhybrid and heterogeneous clusters. These last results have been incorporated in the PhD thesis of M. Benguigui.

7.1.4. Scalable and robust Middleware for distributed event based computing

Participants: Maeva Antoine, Fabrice Huet, Françoise Baude.

In the context of the FP7 STREP PLAY and French SocEDA ANR research projects terminated late 2013, we initiated and pursued the design and development of the Event Cloud.

As a distributed system handling huge amount of information, this middleware can suffer from data imbalances. In a journal extension of a previous workshop paper [6], we have enlarged our litterature review of structured peer to peer systems regarding the way they handle load imbalance to the case of distributed big data systems. We have generalized those popular approaches by proposing a core API that we have proved to be indeed also applicable to the Event Cloud middleware way of implementing a load balancing policy.

7.1.5. Vercors: Integrated environment for verifying and running distributed components

Participants: Ludovic Henrio, Oleksandra Kulankhina, Eric Madelaine.

It is the general prurpose of the Vercors platform to target the generation of distributed applications with safety guarantees. In Vercors, the approach starts from graphical specification formalisms allowing the architectural and behavioral description of component systems. From this point, the user can automatically verify application properties using model-checking techniques. Finally, the specified and verified component model can be translated into executable Java code. The Vercors tool suite is distributed as an Eclipse plugin. This year

- we implemented a first reliable version of the whole tool chain including generation of verifiable models and executable Java code.
- We applied the approach to several examples including Peterson's leader election algorithm, a workflow executor, and the control and management of service composition [7].
- A paper accepted at FASE'2016 presents an overview of this work; a research report provides the full version of the paper [20]. The theoretical background was published as a research report and an improve version is being submitted as a journal paper.

The practical implementation allowed us to improve the presentation of the theory and better evaluate it.

7.2. Run-time/middle-ware level

7.2.1. Virtual Machines Scheduling

Participants: Fabien Hermenier, Vincent Kherbache.

In [19], we present BtrPlace as an application of the dynamic bin packing problem with a focus on its dynamic and heterogeneous nature. We advocate flexibility to answer these issues and present the theoretical aspects of BtrPlace and its modeling using Constraint Programming.

We also continued our work on scheduling VM migrations. In [14], [17], we propose a model for VM migration that consider their memory workload and the network topology. This model was then implemented in place of the previous migration scheduler in BtrPlace. Experiments on a real testbed show the new scheduler outperforms state-of-the-art approaches that cap the migration parallelism by a constant to reduce the completion time. Besides an optimal capping, it reduces the migration duration by 20.4% on average and the completion time by 28.1%. In a maintenance operation involving 96 VMs to migrate between 72 servers, it saves 21.5% Joules against the native BtrPlace. Finally, its current library of 6 constraints allows administrators to address temporal and energy concerns, for example to adapt the schedule and fit a power budget.

Finally, in [10] we transfer the principles of using Constraint Programming to propose a multi-objective job placement algorithm devoted to High Performance Computing (HPC). One of the key decisions made by both MapReduce and HPC cluster management frameworks is the placement of jobs within a cluster. To make this decision, they consider factors like resource constraints within a node or the proximity of data to a process. However, they fail to account for the degree of collocation on the cluster's nodes. A tight process placement can create contention for the intra-node shared resources, such as shared caches, memory, disk, or network bandwidth. A loose placement would create less contention, but exacerbate network delays and increase cluster-wide power consumption. Finding the best job placement is challenging, because among many possible placements, we need to find one that gives us an acceptable trade-off between performance and power consumption. We then propose to tackle the problem via multi-objective optimization. Our solution is able to balance conflicting objectives specified by the user and efficiently find a suitable job placement.

7.3. Application level

7.3.1. DEVS-based Modeling & Simulation

Participants: Olivier Dalle, Damian Vicino.

DEVS is a formalism for the specification of discrete-event simulation models, proposed by Zeigler in the 70's, that is still the subject of many research in the simulation community. Surprisingly, the problem of representing the time in this formalism has always been somehow neglected, and most DEVS simulators keep using Floating Point numbers for their arithmetics on time values, which leads to a range of systematic errors, including severe ones such as breaking the causal relations in the model.

In [15] we propose simulation algorithms, based on the Discrete Event System Specification (DEVS) formalism, that can be used to simulate and obtain every possible output and state trajecto-ries of simulations that receive input values with uncertainty quantification. Then, we present a subclass of DEVS models, called Finite Forkable DEVS (FF-DEVS), that can be simulated by the proposed algorithms. This subclass ensures that the simulation is forking only a finite number of processes for each simulation step. Finally, we discuss the simulation of a traffic light model and show the trajectories obtained when it is subject to input uncertainty.

We have also worked on improving the simulation of DEVS models in some particular situations[16]. Parallel Discrete Event System Specification (PDEVS), for example, is a well-known formalism used to model and simulate Discrete Event Systems. This formalism uses an abstract simulator that defines a set of abstract algorithms that are parallel by nature. To implement simulators using these abstract algorithms , several architectures were proposed. Most of these architectures follow distributed approaches that may not be appropriate for single core processors or microcontrollers. In order to reuse efficiently PDEVS models in this type of systems, we define a new architecture that provides a single threaded execution by passing messages in a call/return fashion to simplify the execution time analysis.

This work has also been presented and defended in the PhD Thesis of D. Vicino[5].

7.3.2. Simulation of Software-Defined Networks

Participants: Olivier Dalle, Damian Vicino.

Software Defined Networks (SDN) is a new technology that has gained a lot of attention recently. It introduces programmatic ways to reorganize the network logical topology. To achieve this, the network interacts with a set of controllers, that can dynamically update the configuration of the network routing equipments based on the received events. As often with new network technologies, discrete-event simulation proves to be an invaluable tool for understanding and analzing the performance and behavior of the new systems. In [8], we use such smulations for evaluating the impact of Software-Defined Networks' Reactive Routing on BitTorrent performance. Indeed, BitTorrent uses choking algorithms that continuously open and close connections to different peers. Software Defined Networks implementing Reactive Routing may be negatively affecting the performances of the system under specific conditions because of it lack of knowledge of BitTorrent strategies.

SPIRALS Project-Team

7. New Results

7.1. Traceability of Concerns in Large Software Systems

In 2015, we obtained new results in the domain of the analysis of large software systems. The purpose is to be able to deal the complexity of such systems by slicing them depending on different concerns. The slicing enables to gain a view and a better understanding on how the concern evolves over time and through the different refinement layers of the software system. For that, we present a systematic approach based on model driven engineering and basic models of software components, in order to better manage software complexity and traceability of functional and non-functional requirements. We provide in particular three major contributions. First, we provide an integrated set of meta-models for describing the concerns of software requirements, software components, and traceability between the concerns and software components. By providing an abstract model, we are independent of any implementation and thus allow existing approaches relying on that model to expand their support. With the second contribution, we propose a formal support of our model to allow formal verification. We focus on temporal property verification. For this, our design model is translated into timed automata for which we can apply a timed model checker. Instead of using temporal logic, which is difficult to handle by non-experts, we use patterns of temporal properties. For each pattern, we propose timed automata that can be applied directly into a timed model checking tool. These timed automata are seen as observers or watch dogs that check the system under observation. Finally, with the last contribution, we propose a software component-based development and verification approach, called SARA, and included in V-lifecycle widely used in the railway domain. These contributions have been validated with case studies from the domain of railway control systems especially for the new European train control system ERTMS/ETCS. These results contribute to our objective on self-optimizing software systems (see Section 3.3) and are part of the PhD thesis by Marc Sango [13].

7.2. Automatic Analysis and Repair of Exception Bugs for Java Programs

In 2015, we obtained new results in the field of automated software repair, that is a new and emerging domain of software engineering. The goal of automatic repair is to increase the quality of software systems by automatizing tasks related to fixing of defects and bugs. The new results that we bring are related with the management of runtime exceptions. These results contribute to our objective on self-healing software systems (see Section 3.2) and are part of the PhD thesis by Benoit Cornu [11], defended on 26 November 2015. To improve the available information about exceptions, we have presented a characterization of the exceptions (expected or not, anticipated or not), and of their corresponding resilience mechanisms [16]. We have provided definitions about what is a bug when facing exceptions and what are the already-in-place corresponding resilience mechanisms. We have formalized two formal resilience properties: source-independence and pure-resilience as well as an algorithm to verify them. Then, we have presented two dynamic analysis techniques based on code transformation for analyzing exceptions. Casper is an approach to make bug fixing easier by providing information about the origin of null pointer dereferences. NpeFix is a system to tolerate null pointer dereferences. Both systems are empirically validated on real-world null dereference bugs from large-scale open-source projects

WHISPER Project-Team

7. New Results

7.1. Software engineering for infrastructure software

Tracking code fragments of interest is important in monitoring a software project over multiple versions. Various approaches, including our previous work on Herodotos, exploit the notion of Longest Common Subsequence, as computed by readily available tools such as GNU Diff, to map corresponding code fragments. Nevertheless, the efficient code differencing algorithms are typically line-based or word-based, and thus do not report changes at the level of language constructs. Furthermore, they identify only additions and removals, but not the moving of a block of code from one part of a file to another. Code fragments of interest that fall within the added and removed regions of code have to be manually correlated across versions, which is tedious and error-prone. When studying a very large code base over a long time, the number of manual correlations can become an obstacle to the success of a study. In a paper published at the IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER) [14], we investigate the effect of replacing the current line-based algorithm used by Herodotos by tree-matching, as provided by the algorithm of the differencing tool GumTree. In contrast to the line-based approach, the tree-based approach does not generate any manual correlations, but it incurs a high execution time. To address the problem, we propose a hybrid strategy that gives the best of both approaches.

Understanding the severity of reported bugs is important in both research and practice. In particular, a number of recently proposed mining-based software engineering techniques predict bug severity, bug report quality, and bug-fix time, according to this information. Many bug tracking systems provide a field "severity" offering options such as "severe", "normal", and "minor", with "normal" as the default. However, there is a widespread perception that for many bug reports the label "normal" may not reflect the actual severity, because reporters may overlook setting the severity or may not feel confident enough to do so. In many cases, researchers ignore "normal" bug reports, and thus overlook a large percentage of the reports provided. On the other hand, treating them all together risks mixing reports that have very diverse properties. In a study published at the Working Conference on Mining Software Repositories (MSR) 2015 [16], we investigate the extent to which "normal" bug reports actually have the "normal" severity. We find that many "normal" bug reports in practice are not normal. Furthermore, this misclassification can have a significant impact on the accuracy of mining-based tools and studies that rely on bug report severity information.

Software is continually evolving, to fix bugs and add new features. Industry users, however, often value stability, and thus may not be able to update their code base to the latest versions. This raises the need to selectively backport new features to older software versions. Traditionally, backporting has been done by cluttering the backported code with preprocessor directives, to replace behaviors that are unsupported in an earlier version by appropriate workarounds. This approach however involves writing a lot of error-prone backporting code, and results in implementations that are hard to read and maintain. In a paper published at the 2015 European Dependable Computing Conference (EDCC) [15], we consider this issue in the context of the Linux kernel, for which older versions are in wide use. We present a new backporting strategy that relies on the use of a backporting compatability library and on code that is automatically generated using the program transformation tool Coccinelle. This approach reduces the amount of code that must be manually written, and thus can help the Linux kernel backporting effort scale while maintaining the dependability of the backporting process.

Logging is a common and important programming practice, but choosing how to log is challenging, especially in a large, evolving software code base that provides many logging alternatives. Insufficient logging may complicate debugging, while logging incorrectly may result in excessive performance overhead and an overload of trivial logs. The Linux kernel has over 13 million lines of code, over 1100 different logging functions, and the strategies for when and how to log have evolved over time. To help developers log correctly

we propose, in a paper published at BEneVol 2015 [18], a framework that will learn existing logging practices from the software development history, and that will be capable of identifying new logging strategies, even when the new strategies just start to be adopted.

7.2. Java runtime support

Java class loaders are commonly used in application servers to load, unload and update a set of classes as a unit. However, unloading or updating a class loader can introduce stale references to the objects of the outdated class loader. A stale reference leads to a memory leak and, for an update, to an inconsistency between the outdated classes and their replacements. To detect and eliminate stale references, in a paper published at DSN 2015 [12], we propose Incinerator, a Java virtual machine extension that introduces the notion of an outdated class loader. Incinerator detects stale references and sets them to null during a garbage collection cycle. We evaluate Incinerator in the context of the OSGi framework and show that Incinerator correctly detects and eliminates stale references, including a bug in Knopflerfish. We also evaluate the performance of Incinerator with the DaCapo benchmark on VMKit and show that Incinerator has an overhead of at most 3.3%.

7.3. Parallel and Distributed Computing

The scalability of multithreaded applications on current multicore systems is hampered by the performance of lock algorithms, due to the costs of access contention and cache misses. In an article published in ACM Transactions on Computer Systems (TOCS), we present a new locking technique, Remote Core Locking (RCL) [10], that aims to accelerate the execution of critical sections in legacy applications on multicore architectures. The idea of RCL is to replace lock acquisitions by optimized remote procedure calls to a dedicated server hardware thread. RCL limits the performance collapse observed with other lock algorithms when many threads try to acquire a lock concurrently and removes the need to transfer lock-protected shared data to the hardware thread acquiring the lock because such data can typically remain in the server's cache. Other contributions presented in this article include a profiler that identifies the locks that are the bottlenecks in multithreaded applications and that can thus benefit from RCL, and a reengineering tool that transforms POSIX lock acquisitions into RCL locks. Eighteen applications were used to evaluate RCL: the nine applications of the SPLASH-2 benchmark suite, the seven applications of the Phoenix 2 benchmark suite, Memcached, and Berkeley DB with a TPC-C client. Eight of these applications are unable to scale because of locks and benefit from RCL on an x86 machine with four AMD Opteron processors and 48 hardware threads. By using RCL instead of Linux POSIX locks, performance is improved by up to 2.5 times on Memcached, and up to 11.6 times on Berkeley DB with the TPC-C client. On a SPARC machine with two Sun Ultrasparc T2+ processors and 128 hardware threads, three applications benefit from RCL. In particular, performance is improved by up to 1.3 times with respect to Solaris POSIX locks on Memcached, and up to 7.9 times on Berkeley DB with the TPC-C client.

Software Transactional Memory (STM) is an optimistic concurrency control mechanism that simplifies parallel programming. Still, there has been little interest in its applicability for reactive applications in which there is a required response time for certain operations. In an article published in ACM Transactions on Parallel Computing (TOPC) [11], we propose supporting such applications by allowing programmers to associate time with atomic blocks in the forms of deadlines and QoS requirements. Based on statistics of past executions, we adjust the execution mode of transactions by decreasing the level of optimism as the deadline approaches. In the presence of concurrent deadlines, we propose different conflict resolution policies. Execution mode switching mechanisms allow meeting multiple deadlines in a consistent manner, with potential QoS degradations being split fairly among several threads as contention increases, and avoiding starvation. Our implementation consists of extensions to a STM runtime that allow gathering statistics and switching execution modes. We also propose novel contention managers adapted to transactional workloads subject to deadlines. The experimental evaluation shows that our approaches significantly improve the likelihood of a transaction meeting its deadline and QoS requirement, even in cases where progress is hampered by conflicts and other concurrent transactions with deadlines.

A challenge in designing a peer-to-peer (P2P) system is to ensure that the system is able to tolerate selfish nodes that strategically deviate from their specification whenever doing so is convenient. In a paper published at SRDS 2015 [13], we propose *RACOON*, a framework for the design of P2P systems that are resilient to selfish behaviours. While most existing solutions target specific systems or types of selfishness, *RACOON* proposes a generic and semi-automatic approach that achieves robust and reusable results. Also, *RACOON* supports the system designer in the performance-oriented tuning of the system, by proposing a novel approach that combines Game Theory and simulations. We illustrate the benefits of using *RACOON* by designing two P2P systems: a live streaming and an anonymous communication system. In simulations and a real deployment of the two applications on a testbed comprising 100 nodes, the systems designed using *RACOON* achieve both resilience to selfish nodes and high performance.

7.4. From Sets to Bits in Coq

Sets form the building block of mathematics, while finite sets are a fundamental data structure of computer science. In the world of mathematics, finite sets enjoy appealing mathematical properties, such as a proof-irrelevant equality and extensionality of functions. Computer scientists, on the other hand, have devised efficient algorithms for set operations based on the representation of finite sets as bit vectors and on bit twiddling, exploiting the hardware's ability to efficiently process machine words.

With interactive theorem provers, sets are reinstituted as mathematical objects. While there are several finite set libraries in Coq, these implementations are far removed from those used in efficient code. Recent work on modeling low-level architectures, such as x86 [41] processors, however, have brought the world of bit twiddling within reach of our proof assistants. We are now able to specify and reason about low-level programs.

In this work, we have implemented bitsets and their associated operations in the Coq proof assistant, thus allowing us to transparently navigate between the concrete world of bit vectors and the abstract world of finite sets. This work grew from a puzzled look at the first page of Warren's *Hacker's Delight* [77], where lies the cryptic formula $x\&(x-1)$ to turn off the rightmost bit in a word. How do we translate the English specification given in the book into a formal definition? How do we prove that this formula meets its specification? Could Coq generate efficient and trustworthy code from it? And how efficiently could we simulate it within Coq itself?

In our work, we have established a bijection between bitsets and sets over finite types. Following a refinement approach, we have shown that a significant part of SSReflect*finset* library can be refined to operations manipulating bitsets. We have also developed a trustworthy extraction of bitsets down to OCaml's machine integers. While we were bound to axiomatize machine integers, we adopted a methodology based on exhaustive testing to gain greater confidence in our model. Finally, we have demonstrateed the usefulness of our library through two applications, a certified implementation of Bloom filters and a verified implementation of the n-queens algorithm.

ALPINES Project-Team

7. New Results

7.1. Communication avoiding algorithms for dense linear algebra

Our group continues to work on algorithms for dense linear algebra operations that minimize communication. During this year we focused on improving the performance of communication avoiding QR factorization as well as designing algorithms for computing rank revealing and low rank approximations of dense and sparse matrices.

In [2] we discuss the communication avoiding QR factorization of a dense matrix. The standard algorithm for computing the QR decomposition of a tall and skinny matrix (one with many more rows than columns) is often bottlenecked by communication costs. The algorithm which is implemented in LAPACK, ScaLAPACK, and Elemental is known as Householder QR. For tall and skinny matrices, the algorithm works column-by-column, computing a Householder vector and applying the corresponding transformation for each column in the matrix. When the matrix is distributed across a parallel machine, this requires one parallel reduction per column. The TSQR algorithm, on the other hand, performs only one reduction during the entire computation. Therefore, TSQR requires asymptotically less inter-processor synchronization than Householder QR on parallel machines (TSQR also achieves asymptotically higher cache reuse on sequential machines). However, TSQR produces a different representation of the orthogonal factor and therefore requires more software development to support the new representation. Further, implicitly applying the orthogonal factor to the trailing matrix in the context of factoring a square matrix is more complicated and costly than with the Householder representation.

We show how to perform TSQR and then reconstruct the Householder vector representation with the same asymptotic communication efficiency and little extra computational cost. We demonstrate the high performance and numerical stability of this algorithm both theoretically and empirically. The new Householder reconstruction algorithm allows us to design more efficient parallel QR algorithms, with significantly lower latency cost compared to Householder QR and lower bandwidth and latency costs compared with Communication-Avoiding QR (CAQR) algorithm. Experiments on supercomputers demonstrate the benefits of the communication cost improvements: in particular, our experiments show substantial improvements over tuned library implementations for tall-and-skinny matrices. We also provide algorithmic improvements to the Householder QR and CAQR algorithms, and we investigate several alternatives to the Householder reconstruction algorithm that sacrifice guarantees on numerical stability in some cases in order to obtain higher performance.

In [4] we introduce CARRQR, a communication avoiding rank revealing QR factorization with tournament pivoting. Revealing the rank of a matrix is an operation that appears in many important problems as least squares problems, low rank approximations, regularization, nonsymmetric eigenproblems. In practice the QR factorization with column pivoting often works well, and it is widely used even if it is known to fail, for example on the so-called Kahan matrix. However in terms of communication, the QR factorization with column pivoting is sub-optimal with respect to lower bounds on communication. If the algorithm is performed in parallel, then typically the matrix is distributed over P processors by using a two-dimensional block cyclic partitionning. This is indeed the approach used in the `psgeqpf` routine from ScaLAPACK. At each step of the decomposition, the QR factorization with column pivoting finds the column of maximum norm and permutes it to the leading position, and this requires exchanging $O(n)$ messages, where n is the number of columns of the input matrix. For square matrices, when the memory per processor used is on the order of $O(n^2/P)$, the lower bound on the number of messages to be exchanged is $\Omega(\sqrt{P})$. The number of messages exchanged during the QR factorization with column pivoting is larger by at least a factor of $n/\sqrt{P}$ than the lower bound.

In this paper we introduce CARRQR, a communication optimal (modulo polylogarithmic factors) rank revealing QR factorization based on tournament pivoting. The factorization is based on an algorithm that computes the decomposition by blocks of b columns (panels). For each panel, tournament pivoting proceeds in two steps. The first step aims at identifying a set of b candidate pivot columns that are as well-conditioned as possible. These columns are permuted to the leading positions, and they are used as pivots for the next b steps of the QR factorization. To identify the set of b candidate pivot columns, a tournament is performed based on a reduction operation, where at each node of the reduction tree b candidate columns are selected by using the strong rank revealing QR factorization. The idea of tournament pivoting has been first used to reduce communication in Gaussian elimination, an algorithm referred to as CALU.

We show that CARRQR reveals the numerical rank of a matrix in an analogous way to QR factorization with column pivoting (QRCP). Although the upper bound of a quantity involved in the characterization of a rank revealing factorization is worse for CARRQR than for QRCP, our numerical experiments on a set of challenging matrices show that this upper bound is very pessimistic, and CARRQR is an effective tool in revealing the rank in practical problems.

Our main motivation for introducing CARRQR is that it minimizes data transfer, modulo polylogarithmic factors, on both sequential and parallel machines, while previous factorizations as QRCP are communication sub-optimal and require asymptotically more communication than CARRQR. Hence CARRQR is expected to have a better performance on current and future computers, where commmunication is a major bottleneck that highly impacts the performance of an algorithm.

7.2. Algebraic preconditioners

Our work focused on the design of robust algebraic preconditioners and domain decomposition methods to accelerate the convergence of iterative methods.

In [5] we present a communication avoiding ILU0 preconditioner for solving large linear systems of equations by using iterative Krylov subspace methods. Recent research has focused on communication avoiding Krylov subspace methods based on so called s-step methods. However there is no communication avoiding preconditioner yet, and this represents a serious limitation of these methods. Our preconditioner allows to perform s iterations of the iterative method with no communication, through ghosting some of the input data and performing redundant computation. It thus reduces data movement by a factor of $3s$ between different levels of the memory hierarchy in a serial computation and between different processors in a parallel computation. To avoid communication, an alternating reordering algorithm is introduced for structured and unstructured matrices, that requires the input matrix to be ordered by using a graph partitioning technique such as kway or nested dissection. We show that the reordering does not affect the convergence rate of the ILU0 preconditioned system as compared to kway or nested dissection ordering, while it reduces data movement and should improve the expected time needed for convergence. In addition to communication avoiding Krylov subspace methods, our preconditioner can be used with classical methods such as GMRES or s-step methods to reduce communication.

7.3. A robust coarse space for Optimized Schwarz methods SORAS-GenEO-2

Optimized Schwarz methods (OSM) are very popular methods which were introduced by P.L. Lions for elliptic problems and Després for propagative wave phenomena. In [18], we have built a coarse space for which the convergence rate of the two-level method is guaranteed regardless of the regularity of the coefficients. We do this by introducing a symmetrized variant of the ORAS (Optimized Restricted Additive Schwarz) algorithm and by identifying the problematic modes using two different generalized eigenvalue problems instead of only one as for the ASM (Additive Schwarz method), BDD (balancing domain decomposition) or FETI (finite element tearing and interconnection) methods.

7.4. Time-dependent wave splitting and source separation

Starting from classical absorbing boundary conditions, we propose, in [17], a method for the separation of time-dependent scattered wave fields due to multiple sources or obstacles. In contrast to previous techniques, our method is local in space and time, deterministic, and also avoids a priori assumptions on the frequency spectrum of the signal. Numerical examples in two space dimensions illustrate the usefulness of wave splitting for time-dependent scattering problems.

7.5. Boundary integral formulations of wave scattering

We have continued to develop and further analyze new boundary integral formulation for wave scattering by complex objects.

In [13] we considered acoustic scattering of time-harmonic waves at objects composed of several homogeneous parts. Some of those may be impenetrable, giving rise to Dirichlet boundary conditions on their surfaces. We started from the second-kind boundary integral approach of [X. Claeys, and R. Hiptmair, and E. Spindler. A second-kind Galerkin boundary element method for scattering atcomposite objects. BIT Numerical Mathematics, 55(1):33-57, 2015] and extended it to this new setting. Based on so-called global multi-potentials, we derived variational second-kind boundary integral equations posed in $L^2(\Sigma)$, where Σ denotes the union of material interfaces. To suppress spurious resonances, we introduced a combined-field version (CFIE) of our new method. We conducted thorough numerical tests that highlighted the low and mesh-independent condition numbers of Galerkin matrices obtained with discontinuous piecewise polynomial boundary element spaces. They also confirmed competitive accuracy of the numerical solution in comparison with the widely used first-kind single-trace approach.

We spent much effort investigating the potentialities of multi-trace formulations in terms of domain decomposition. We considered multi-trace formulations in this perspective. Indeed Multi-Trace Formulations are based on a decomposition of the problem domain into subdomains, and thus domain decomposition solvers are of interest. The fully rigorous mathematical MTF can however be daunting for the non-specialist. In [12], we introduced MTFs on simple model problems using concepts familiar to researchers in domain decomposition. This allowed us to get a new understanding of MTFs and a natural block Jacobi iteration, for which we determined optimal relaxation parameters. We then showed how iterative multitrace formulation solvers are related to a well known domain decomposition method called optimal Schwarz method: a method which used Dirichlet to Neumann maps in the transmission condition. We finally showed that the insight gained from the simple model problem leads to remarkable identities for Calderón projectors and related operators, and the convergence results and optimal choice of the relaxation parameter we obtained is independent of the geometry, the space dimension of the problem, and the precise form of the spatial elliptic operator, like for optimal Schwarz methods. We confirmed this analysis with numerical experiments.

This work was extended in [10]. Considering pure transmission scattering problems in piecewise constant media, we derived an exact analytic formula for the spectrum of the corresponding local multi-trace boundary integral operators in the case where the geometrical configuration does not involve any junction point and all wave numbers equal. We deduced from this the essential spectrum in the case where wave numbers vary. Numerical evidences of these theoretical results were obtained in 2D.

Finally, in connection with boundary integral formulations, we extended the past work of [X. Claeys and R. Hiptmair, *Integral equations on multi-screens.* Integral Equations and Operator Theory, 77(2):167–197, 2013] where we had developed a framework for the analysis of boundary integral equations for acoustic scattering at so-called multi-screens, which are arbitrary arrangements of thin panels made of impenetrable material. In [3] we extended these considerations to boundary integral equations for electromagnetic scattering.

Viewing tangential multi-traces of vector fields from the perspective of quotient spaces we introduced the notion of single-traces and spaces of jumps. We also derived representation formulas and established key properties of the involved potentials and related boundary operators. Their coercivity were proved using a splitting of jump fields. Another new aspect emerged in the form of surface differential operators linking various trace spaces.

7.6. Asymptotic models for time harmonic wave propagation

Asymptotic models oriented toward more efficient numerical simulation methods have been investigated in three different directions.

In [8] we considered the Poisson equation in a domain with a small hole of size δ, and presented a simple numerical method, based on an asymptotic analysis, which allows to approximate robustly the far field of the solution as δ goes to zero without meshing the small hole. We proved the stability of the scheme and provide error estimates. This was confirmed with numerous numerical experiments illustrating the efficiency of the technique.

In [11] we considered a Laplace problem with Dirichlet boundary condition in a three dimensional domain containing an inclusion taking the form of a thin tube with small thickness. We proved convergence in operator norm of the resolvent of this problem as the thickness goes to 0, establishing that the perturbation on the resolvent induced by the inclusion is not greater than some (negative) power of the logarithm of the thickness. From this we deduced convergence of the eigenvalues of the perturbed operator toward the limit operator.

In [9] we investigated the eigenvalue problem $-\mathrm{div}(\sigma\nabla u) = \lambda u$ $(\mathcal{P})$ in a 2D domain Ω divided into two regions $\Omega_\pm$. We were interested in situations where σ takes positive values on Ω_+ and negative ones on Ω_-. Such problems appear in time harmonic electromagnetics in the modeling of plasmonic technologies. In a recent work [L. Chesnel, X. Claeys, and S.A. Nazarov. *A curious instability phenomenon for a rounded corner in presence of a negative material*. Asymp. Anal., 88(1):43–74, 2014], we had highlighted an unusual instability phenomenon for the source term problem associated with $(\mathcal{P})$: for certain configurations, when the interface between the subdomains $\Omega_\pm$ presents a rounded corner, the solution may depend critically on the value of the rounding parameter. In [9] we explained this property studying the eigenvalue problem $(\mathcal{P})$. We provided an asymptotic expansion of the eigenvalues and prove error estimates. We established an oscillatory behaviour of the eigenvalues as the rounding parameter of the corner tends to zero. This work was ended with numerical illustrations.

7.7. New results related to FreeFem++

In [6], we consider a model of soil water and nutrient transport with plant root uptake. The geometry of the plant root system is explicitly taken into account in the soil model. We first describe our modeling approach. Then, we introduce an adaptive mesh refinement procedure enabling us to accurately capture the geometry of the root system and small-scale phenomena in the rhizosphere. Finally, we present a domain decomposition technique for solving the problems arising from the soil model as well as some numerical results.

In [15], we study an interface transport scheme of a two-phase flow of an incompressible viscous immiscible fluid. The problem is discretized by the characteristics method in time and finite elements in space. The interface is captured by the Level-Set function. Appropriate boundary conditions for the problem of mould filling are investigated, a new natural boundary condition under pressure effect for the transport equation is proposed and an algorithm for computing the solution is presented. Finally, numerical experiments show and validate the effectiveness of the proposed scheme.

AVALON Project-Team

7. New Results

7.1. Energy Efficiency of Large Scale Distributed Systems

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

7.1.1. Energy efficient Core Networks

This work [8], [43] seeks to improve the energy efficiency of backbone networks by providing an intra-domain Software Defined Network (SDN) approach to selectively turn off a subset of links. To do this, we designed an energy-aware traffic engineering technique for reducing energy consumption in backbone networks. Energy-efficient traffic engineering was analysed in previous work, but none addressed implementation challenges of their solutions. We showed that ignoring to test the feasibility of techniques can lead to bad estimations and unstable solutions. We proposed the STREETE framework (SegmenT Routing based Energy Efficient Traffic Engineering) that represents an online method to switch some links off/on dynamically according to the network load. We have implemented a working prototype in the OMNET++ simulator. Networks are progressively using centralised architecture, and SDN is increasingly utilised in data centre networks. We believe that SDN may be extended to backbone networks. The implemented solution shows that SDN may also be a good means for reducing the energy consumption of network devices. Compared to previous work, in this work we used the SPRING protocol to improve the stability of energy-efficient traffic engineering solutions. To the best of our knowledge, this is the first work proposing the use of SPRING to improve the energy efficiency of backbone networks. The flexibility of this routing protocol is well suited to frequent route changes that happen when we switch links off and on. Moreover, this protocol can be easily applied to SDN solutions. Using simulations, we showed that as much as 44% of links can be switched off to save energy in real backbone networks. Even greedy techniques can easily approach the maximum reduction in the amount of energy consumed. In fact, the bottleneck in terms of energy efficiency in energy-aware traffic engineering is the connectivity constraint. We performed a stress test of our solution under rapidly increasing traffic patterns and showed that more work must be done in the domain of switching links back on: a field which has received little attention from the research community.

7.1.2. Energy proportionality in HPC systems

Energy savings are among the most important topics concerning Cloud and HPC infrastructures nowadays. Servers consume a large amount of energy, even when their computing power is not fully utilized. These static costs represent quite a concern, mostly because many datacenter managers are over-provisioning their infrastructures compared to the actual needs. This results in a high part of wasted power consumption. In this work [19], [47] , we proposed the BML ("Big, Medium, Little") infrastructure, composed of heterogeneous architectures, and a scheduling framework dealing with energy proportionality. We introduce heterogeneous power processors inside datacenters as a way to reduce energy consumption when processing variable workloads. Our framework brings an intelligent utilization of the infrastructure by dynamically executing applications on the architecture that suits their needs, while minimizing energy consumption. Our first validation process focuses on distributed stateless web servers scenario and we analyze the energy savings achieved through energy proportionality. This research activity is performed with the collaboration of Sepia Team (IRIT, Toulouse) through the co-advising of Violaine Villebonnet.

7.1.3. Energy-Aware Server Provisioning

Several approaches to reduce the power consump- tion of datacenters have been described in the literature, most of which aim to improve energy efficiency by trading off performance for reducing power consumption. However, these approaches do not always provide means for administrators and users to specify how they want to explore such trade-offs. This work [27] provides techniques for assigning jobs to distributed resources, exploring energy efficient resource provisioning. We use middleware-level mechanisms to adapt resource allocation according to energy-related events and user-defined rules. A proposed framework enables developers, users and system ad- ministrators to specify and explore energy efficiency and perfor- mance trade-offs without detailed knowledge of the underlying hardware platform. Evaluation of the proposed solution under three scheduling policies shows gains of 25% in energy-efficiency with minimal impact on the overall application performance. We also evaluate reactivity in the adaptive resource provisioning This approach has been applied in the Nuage research project [26].

7.1.4. Virtual Home Gateway

About 80-90% of the energy in today's wireline networks is consumed in the access network, , including about 10 to 30W per user being dissipated mostly by the customer premises equipment (CPE). Home gateway is a popular equipment deployed at the end of networks and supporting a set of heterogeneous services (data, phone, television, multimedia, security services). These gateways and associated services can be difficult to deploy and maintain for customers. These gateways are difficult to manage for network operators and consume a lot of energy. We explore the technical solutions to reduce the complexity and energy impact of such equipments by moving services to some external dedicated and shared facilities of network operators. This result is a joint work between Avalon team (J.P. Gelas, L. Lefevre) and Addis Abeba University (M. Tsibie and T. Assefa). This research has been demonstrated in the GreenTouch final celebration event in New York (June 2015).

7.2. MPI Application and Storage System Simulation

Participants: Frédéric Suter, Laurent Pouilloux.

7.2.1. Scalable Off-line Simulation of MPI Applications

Analyzing and understanding the performance behavior of parallel applications on parallel computing platforms is a long-standing concern in the High Performance Computing community. When the targeted platforms are not available, simulation is a reasonable approach to obtain objective performance indicators and explore various hypothetical scenarios. In the context of applications implemented with the Message Passing Interface, two simulation methods have been proposed, on-line simulation and off-line simulation, both with their own drawbacks and advantages.

We proposed in [9] an off-line simulation framework, i.e., one that simulates the execution of an application based on event traces obtained from an actual execution. The main novelty of this work, when compared to previously proposed off-line simulators, is that traces that drive the simulation can be acquired on large, distributed, heterogeneous, and non-dedicated platforms. As a result the scalability of trace acquisition is increased, which is achieved by enforcing that traces contain no time-related information. Moreover, our framework is based on an state-of-the-art scalable, fast, and validated simulation kernel.

Such off-line analysis faces scalability issues for acquiring, storing, or replaying large event traces. Then, in [10], we combined our framework with another, specialized in the production of compact traces, to capitalize on their respective strengths while alleviating several of their limitations. We showed that the combined framework affords levels of scalability that are beyond that achievable by either one of the two individual frameworks.

7.2.2. Simulation of Storage Elements

Storage is a essential component of distributed computing infrastructures, *i.e.,* clusters, grids, clouds, data centers, or supercomputers, to cope with the tremendous increase in scientific data production and the ever-growing need for data analysis and preservation. Understanding the performance of a storage subsystem or dimensioning it properly is an important concern for which simulation can help by allowing for fast, fully repeatable, and configurable experiments for arbitrary hypothetical scenarios. However, most simulation frameworks tailored for the study of distributed systems offer no or little abstractions or models of storage resources.

In [34], we detailed the extension of SimGrid with storage simulation capacities. We first defined the required abstractions and propose a new API to handle storage components and their contents in SimGrid-based simulators. Then we characterized the performance of the fundamental storage component that are disks and derive models of these resources. Finally we listed several concrete use cases of storage simulations in clusters, grids, clouds, and data centers for which the proposed extension would be beneficial.

7.3. MapReduce Computations on Hybrid Distributed Computations Infrastructures

Participants: Gilles Fedak, Julio Anjos, Anthony Simonet.

In this section we report on our efforts to provide MapReduce Computing environments on Hybrid infrastructures, i.e composed of Desktop Grids and Cloud computing environments.

Cloud computing has increasingly been used as a platform for running large business and data processing applications. Although clouds have become extremely popular, when it comes to data processing, their use incurs high costs. Conversely, Desktop Grids, have been used in a wide range of projects, and are able to take advantage of the large number of resources provided by volunteers, free of charge. Merging cloud computing and desktop grids into a hybrid infrastructure can provide a feasible low-cost solution for big data analysis. Although frameworks like MapReduce have been devised to exploit commodity hardware, their use in a hybrid infrastructure raise some challenges due to their large resource heterogeneity and high churn rate.

7.3.1. BIGhybrid - A Toolkit for Simulating MapReduce in Hybrid Infrastructures

In [20], we introduced BIGhybrid, a toolkit that is used to simulate MapReduce in hybrid environments. Its main goal is to provide a framework for developers and system designers that can enable them to address the issues of Hybrid MapReduce. In this paper, we described the framework which simulates the assembly of two existing middleware: BitDew- MapReduce for Desktop Grids and Hadoop-BlobSeer for Cloud Computing. The experimental results that are included in this work demonstrate the feasibility of our approach.

7.3.2. HybridMR: a New Approach for Hybrid MapReduce Combining Desktop Grid and Cloud Infrastructures

In [18], we proposed a novel MapReduce computation model in hybrid computing environment called HybridMR. Using this model, high performance cluster nodes and heterogeneous desktop PCs in Internet or Intranet can be integrated to form a hybrid computing environment. In this way, the computation and storage capability of large-scale desktop PCs can be fully utilized to process large-scale datasets. HybridMR relies on a hybrid distributed file system called HybridDFS, and a time-out method has been used in HybridDFS to prevent volatility of desktop PCs, and file replication mechanism is used to realize reliable storage. A new node priority-based fair scheduling (NPBFS) algorithm has been developed in HybridMR to achieve both data storage balance and job assignment balance by assigning each node a priority through quantifying CPU speed, memory size and I/O bandwidth. Performance evaluation results showed that the proposed hybrid computation model not only achieves reliable MapReduce computation, reduces task response time and improves the performance of MapReduce, but also reduces the computation cost and achieves a greener computing mode.

7.3.3. D^3 -MapReduce: Towards MapReduce for Distributed and Dynamic Data Sets

So far MapReduce has been mostly designed for batch processing of bulk data. The ambition of D^3-MapReduce, presented in [32], is to extend the MapReduce programming model and propose efficient implementation of this model to: i) cope with distributed data sets, i.e. that span over multiple distributed infrastructures or stored on network of loosely connected devices; ii) cope with dynamic data sets, i.e. which dynamically change over time or can be either incomplete or partially available. In this paper, we draw the path towards this ambitious goal. Our approach leverages Data Life Cycle as a key concept to provide MapReduce for distributed and dynamic data sets on heterogeneous and distributed infrastructures. We first reported on our attempts at implementing the MapReduce programming model for Hybrid Distributed Computing Infrastructures (Hybrid DCIs). We present the architecture of the prototype based on BitDew, a middleware for large scale data management, and Active Data, a programming model for data life cycle management. Second, we outlined the challenges in term of methodology and present our approaches based on simulation and emulation on the Grid'5000 experimental testbed. We conducted performance evaluations and compare our prototype with Hadoop, the industry reference MapReduce implementation. We presented our work in progress on dynamic data sets that has lead us to implement an incremental MapReduce framework. Finally, we discussed our achievements and outline the challenges that remain to be addressed before obtaining a complete D 3-MapReduce environment.

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

MapReduce offers an ease-of-use programming paradigm for processing large datasets. In our previous work, we have designed a MapReduce framework called BitDew-MapReduce for desktop grid and volunteer computing environment, that allows nonexpert users to run data-intensive MapReduce jobs on top of volunteer resources over the Internet. However, network distance and resource availability have great impact on MapReduce applications running over the Internet. To address this, an availability and network-aware MapReduce framework over the Internet is proposed in [38]. Simulation results show that the MapReduce job response time could be decreased by 27.15%, thanks to Naive Bayes Classifier-based availability prediction and landmark-based network estimation.

7.4. Managing Big Data Life Cycle

Participants: Gilles Fedak, Anthony Simonet.

7.4.1. Active Data - Enabling Smart Data Life Cycle Management for Large Distributed Scientific Data Sets

The Big Data challenge consists in managing, storing, analyzing and visualizing these huge and ever growing data sets to extract sense and knowledge. As the volume of data grows exponentially, the management of these data becomes more complex in proportion. A key point is to handle the complexity of the data life cycle, i.e. the various operations performed on data: transfer, archiving, replication, deletion, etc. Indeed, data-intensive applications span over a large variety of devices and e-infrastructures which implies that many systems are involved in data management and processing. In [17], we proposed Active Data, a programming model to automate and improve the expressiveness of data management applications. We first define the concept of data life cycle and introduce a formal model that allows to expose data life cycle across heterogeneous systems and infrastructures. The Active Data programming model allows code execution at each stage of the data life cycle: routines provided by programmers are executed when a set of events (creation, replication, transfer, deletion) happen to any data. We implement and evaluate the model with four use cases: a storage cache to Amazon-S3, a cooperative sensor network, an incremental implementation of the MapReduce programming model and automated data provenance tracking across heterogeneous systems. Altogether, these scenarios illustrate the adequateness of the model to program applications that manage distributed and dynamic data sets. We also show that applications that do not leverage on data life cycle can still benefit from Active Data to improve their performances.

7.4.2. Using Active Data to Provide Smart Data Surveillance to E-Science Users

Modern scientific experiments often involve multiple storage and computing platforms, software tools, and analysis scripts. The resulting heterogeneous environments make data management operations challenging, the significant number of events and the absence of data integration makes it difficult to track data provenance, manage sophisticated analysis processes, and recover from unexpected situations. Current approaches often require costly human intervention and are inherently error prone. The difficulties inherent in managing and manipulating such large and highly distributed datasets also limits automated sharing and collaboration. In [37], we study a real world e-Science application involving terabytes of data, using three different analysis and storage platforms, and a number of applications and analysis processes. We demonstrate that using a specialized data life cycle and programming model, Active Data, we can easily implement global progress monitoring, and sharing, recover from unexpected events, and automate a range of tasks.

7.4.3. SMART: An Application Framework for Real Time Big Data Analysis on Heterogeneous Cloud Environments.

The amount of data that human activities generate poses a challenge to current computer systems. Big data processing techniques are evolving to address this challenge, with analysis increasingly being performed using cloud-based systems. Emerging services, however, require additional enhancements in order to ensure their applicability to highly dynamic and heterogeneous environments and facilitate their use by Small & Medium-sized Enterprises (SMEs). Observing this landscape in emerging computing system development, this work presents Small & Medium-sized Enterprise Data Analytic in Real Time (SMART) for addressing some of the issues in providing compute service solutions for SMEs. SMART offers a framework for efficient development of Big Data analysis services suitable to small and medium-sized organizations, considering very heterogeneous data sources, from wireless sensor networks to data warehouses, focusing on service composability for a number of domains. In [62], we presented the basis of this proposal and preliminary results on exploring application deployment on hybrid infrastructure.

7.5. Desktop Grid Computing

Participants: Gilles Fedak, Anthony Simonet.

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

Assembling and simultaneously using different types of distributed computing infrastructures (DCI) like Grids and Clouds is an increasingly common situation. Because infrastructures are characterized by different attributes such as price, performance, trust, greenness, the task scheduling problem becomes more complex and challenging. In [15], we presented the design for a fault-tolerant and trust-aware scheduler, which allows to execute Bag-of-Tasks applications on elastic and hybrid DCI, following user-defined scheduling strategies. Our approach, named Promethee scheduler, combines a pull-based scheduler with multi-criteria Promethee decision making algorithm. Because multi-criteria scheduling leads to the multiplication of the possible scheduling strategies, we proposed SOFT, a methodology that allows to find the optimal scheduling strategies given a set of application requirements. The validation of this method is performed with a simulator that fully implements the Promethee scheduler and recreates an hybrid DCI environment including Internet Desktop Grid, Cloud and Best Effort Grid based on real failure traces. A set of experiments shows that the Promethee scheduler is able to maximize user satisfaction expressed accordingly to three distinct criteria: price, expected completion time and trust, while maximizing the infrastructure useful employment from the resources owner point of view. Finally, we present an optimization which bounds the computation time of the Promethee algorithm, making realistic the possible integration of the scheduler to a wide range of resource management software.

7.5.2. Synergy of Volunteer Measurements and Volunteer Computing for Effective Data Collecting, Processing, Simulating and Analyzing on a Worldwide Scale

The paper [31] concerns the hype idea of Citizen Science and the related paradigm shift: to go from the passive "volunteer computing" to other volunteer actions like "volunteer measurements" under guidance of scientists. They can be carried out by ordinary people with standard computing gadgets (smartphone, tablet, etc.) and the various standard sensors in them. Here the special attention is paid to the system of volunteer scientific measurements to study air showers caused by cosmic rays. The technical implementation is based on integration of data about registered night flashes (by radiometric software) in shielded camera chip, synchronized time and GPS-data in ordinary gadgets: to identify night air showers of elementary particles; to analyze the frequency and to map the distribution of air showers in the densely populated cities. The project currently includes the students of the National Technical University of Ukraine KPI, which are compactly located in Kyiv city and contribute their volunteer measurements. The technology would be very effective for other applications also, especially if it will be automated (e.g., on the basis of XtremWeb or/and BOINC technologies for distributed computing) and used in some small area with many volunteers, e.g. in local communities (Corporative/Community Crowd Computing).

7.5.3. Towards an Environment for doing Data Science that runs in Browsers

In [25], we proposed a path for doing Data Science using browsers as computing and data nodes. This novel idea is motivated by the cross-fertilized fields of desktop grid computing, data management in grids and clouds, Web technologies such as Nosql tools, models of interactions and programming models in grids, cloud and Web technologies. We propose a methodology for the modeling, analyzing, implemention and simulation of a prototype able to run a MapReduce job in browsers. This work allows to better understand how to envision the big picture of Data Science in the context of the Javascript language for programming the middleware, the interactions between components and browsers as the operating system. We explain what types of applications may be impacted by this novel approach and, from a general point of view how a formal modeling of the interactions serves as a general guidelines for the implementation. Formal modeling in our methodology is a necessary condition but it is not sufficient. We also make round-trips between the modeling and the Javascript or used tools to enrich the interaction model that is the key point, or to put more details into the implementation. It is the first time to the best of our knowledge that Data Science is operating in the context of browsers that exchange codes and data for solving computational and data intensive programs. Computational and data intensive terms should be understand according to the context of applications that we think to be suitable for our system.

7.5.4. E-Fast & CloudPower: Towards High Performance Technical Analysis for Small Investors.

About 80% of the financial market investors fail, the main reason for this being their poor investment decisions. Without advanced financial analysis tools and the knowledge to interpret the analysis, the investors can easily make irrational investment decisions. Moreover, in- vestors are challenged by the dynamism of the market and a relatively large number of indicators that must be computed. In this paper we propose E-Fast, an innovative approach for on-line technical analysis for helping small investors to obtain a greater efficiency on the market by increasing their knowledge. The E-Fast technical analysis platform proto- type relies on High Performance Computing (HPC), allowing to rapidly develop and extensively validate the most sophisticated finance analysis algorithms. In [36], we aim at demonstrating that the E-Fast im- plementation, based on the CloudPower HPC infrastructure, is able to provide small investors a realistic, low-cost and secure service that would otherwise be available only to the large financial institutions. We describe the architecture of our system and provide design insights. We present the results obtained with a real service implementation based on the Exponential Moving Average computational method, using CloudPower and Grid5000 for the computations' acceleration. We also elaborate a set of interesting challenges emerging from this work, as next steps towards high performance technical analysis for small investors.

7.6. HPC Component Model

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

7.6.1. 3D FFT and L^{2}C

We have completed the work started in 2014. To harness the computing power of supercomputers, HPC application algorithms have to be adapted to the underlying hardware. This is a costly and complex process which requires handling many algorithm variants. In [23], we studied the ability of the component model L^{2}C to express and handle the variability of HPC applications. The goal is to ease application adaptation. Analysis and experiments are done on a 3D-FFT use case. Results show that L^{2}C, and components in general, offer a generic and simple handling of 3D-FFT variants while obtaining performance close to well-known libraries

7.6.2. Multi-Stencil DSL in L^{2}C

As high performance architectures evolve continuously to be more powerful, such architectures also usually become more difficult to use efficiently. As a scientist is not a low level and high performance programming expert, Domain Specific Languages (DSLs) are a promising solution to automatically and efficiently write high performance codes. However, if DSLs ease programming for scientists, maintainability and portability issues are transferred from scientists to DSL de- signers. This work [44] has dealt with an approach to improve maintainability and programming productivity of DSLs through the generation of a component-based parallel runtime. To study it, we have designed a DSL for multi-stencil programs, that is evaluated on a real-case of shallow water equations implemented with L^{2}C.

7.6.3. Reconfigurable HPC component model

High-performance applications whose structure changes dynamically during execution are extremely complex to develop, maintain and adapt to new hardware. Such applications would greatly benefit from easy reuse and separation of concerns which are typical advantages of component models. Unfortunately, no existing component model is both HPC-ready (in terms of scalability and overhead) and able to easily handle dynamic reconfiguration. In [33], we aimed at addressing performance, scalability and programmability by separating locking and synchronization concerns from reconfiguration code. To this end, we propose directMOD, a component model which provides on one hand a flexible mechanism to lock subassemblies with a very small overhead and high scalability, and on the other hand a set of well-defined mechanisms to easily plug various independently-written reconfiguration components to lockable subassemblies. We evaluate both the model itself and a C++/MPI implementation called directL2C .

7.6.4. Towards a Task-Component Model

In [24], we propose a first model that aims at combining both component models and task based models such as StarPU. Component models bring many good software enginering properties such as code re-use while task based models seems to be very efficient to exploit recent hardware such as SMP, manycore, or GPGPUs. This work evaluates a proof-of-concepts only considering SMP nodes.

7.7. Security for Virtualization and Clouds

Participants: Eddy Caron, Arnaud Lefray.

7.7.1. Security and placement

We have proposed a solution for placement-based security and client-centric security. Even with perfect information flow control mechanisms, virtualized environments are still sensitive to silent information leakage, that is covert channels, due to shared hardware ressources. We have proposed a fine-grained placement based on the client's security properties to tackle this issue. The client submits an application i.e., a graph of VMs, and information flow rules defining the acceptable risk. Due to the lack of usable covert channel metric to qualify an acceptable risk, we have proposed a new information leakage metric. As covert channels exploit microarchitecture flaws, we have integrated the specificity of NUMA allocation schemes in our placement algorithm.

7.7.2. Security and logic language

Besides, the main issue with existing security languages is the ability to formally guarantee the required property. On the one hand, security policies described in a natural language have quite ambiguous semantics. On the other hand, a formal language or logic provides clear syntax and semantics. Moreover, existing mechanisms are dedicated to secure specific type of entities (e.g., VM, Service, Data, VNet). Therefore, the problem is to have a formal definition of security properties and proven procedures to transform the end-user's global security properties into multiple local properties enforceable by several local mechanisms. For these reasons, we proposed a logic language called IF-PLTL (Information Flow Past Linear Time Logic). Our logic is dedicated to controlling the propagation of information i.e., direct and indirect information flows. As these information flows cannot be obtained directly, we have explained their construction from low-level observable events. Security decisions are naturally expressed according to past actions. Accordingly, IF-PLTL is based on the past fragment of LTL. In addition to using IF-PLTL to transform properties, we have proposed a dynamic monitor that can enforce the full expressivity of IF-PLTL even if its complexity (in time and space) would incur a high overhead in practice.

7.8. Autonomic Middleware Deployment using Self-Stabilization

Participants: Eddy Caron, Maurice Faye.

Dynamic nature of distributed architecture is a major challenge to avail the benefits of distributed computing. An effective solution to deal with this dynamic nature is to implement a self-adaptive mechanism to sustain the distributed architecture. Self-adaptive systems can autonomously modify their behavior at run-time in response to changes in their environment. This capability may be included in the software systems at design time or later by external mechanisms. We have created a self-adaptive algorithm for the DIET middleware. Once the middleware is deployed, it can detect a set of events which indicate an unstable deployment state. When an event is detected, some instructions are executed to handle the event. We have designed a simulator to have a deeper insights of our proposed self-adaptive algorithm.

HIEPACS Project-Team

7. New Results

7.1. High-performance computing on next generation architectures

7.1.1. Soft error sensitivity of PCG and reliability of detection mechanisms

Soft errors can be defined as failures arising from several electricity fluctuations, cosmic particle effects on chip or any other unexpected problem while computations are in progress. If computational environment grows up to exascale, the rate of these types of error is likely to increase. These bit-flips may have a strong impact on iterative methods, that might diverge or converge to an unexpected final accuracy. Consequently, soft errors deserve to be examined in details especially in the perspective of extreme scale computing platforms. In this work, we investigate the combination of different numerical techniques to tackle the challenge of the detection. The first ingredient relies on checksum mechanisms, that are applied to secure the sparse matrix vector (SpMV) products. However, the checksum equalities are only valid in exact arithmetic while calculation are performed in finite precision. Another possibility is to monitor the residual deviation between the true and computed residual. Exploiting finite precision analysis of the round-off provides us with an upper bound on the residual norm deviation that can be used. Through intensive numerical experiments and statistical analysis we shown how round-off error analysis for the residual norm deviation can be an efficient and robust soft error detection criterion alternative to checksum approaches. This methology has also be applied to other variants of CG, namely the pipelined and chronopolus/gear versions.

This research effort was conduced in collaboration with colleagues S. Cools and W. Vanroose from the Applied Mathematics Group of Antwerp university within the framework of the EXA2CT project. In this context, we also studied the impact of soft errors on a variant of the algorithm designed in their group (so-called pipelined CG). This study allowed to highlight some numerical instability in the baseline version of this variant of CG in the presence of round-off errors and we jointly proposed a correction of it that led a new both scalable and stable variant (see Section 7.2.5).

We have also designed an self-recovering CG algorithm which detects large magnitued faults with ABFT and smoothes low and average magnitued faults with deviation-based criteria.

7.1.2. Resilience of parallel sparse hybrid solvers

As the computational power of high performance computing (HPC) systems continues to increase by using a huge number of CPU cores or specialized processing units, extreme-scale applications are increasingly prone to faults. Consequently, the HPC community has proposed many contributions to design resilient HPC applications. These contributions may be system-oriented, theoretical or numerical. In this study we consider an actual fully-featured parallel sparse hybrid (direct/iterative) linear solver, MaPHyS, and we propose numerical remedies to design a resilient version of the solver. The solver being hybrid, we focus in this study on the iterative solution step, which is often the dominant step in practice. We furthermore assume that a separate mechanism ensures fault detection and that a system layer provides support for setting back the environment (processes, ...) in a running state. The present manuscript therefore focuses on (and only on) strategies for recovering lost data *after* the fault has been detected (a separate concern beyond the scope of this study), *once* the system is restored (another separate concern not studied here). The numerical remedies we propose are twofold. Whenever possible, we exploit the natural data redundancy between processes from the solver to perform exact recovery through clever copies over processes. Otherwise, data that has been lost and no longer available on any process is recovered through a so-called interpolation-restart mechanism. This mechanism is derived from our earlier studies by carefully taking into account the properties of the target hybrid solver. These numerical remedies have been implemented in the MaPHyS parallel solver so that we can assess their efficiency on a large number of processing units (up to $12,288$ CPU cores) for solving large-scale real-life problems.

These contributions will be presented at the international conference HiPC [42].

7.1.3. Hierarchical DAG scheduling for hybrid distributed systems

Accelerator-enhanced computing platforms have drawn a lot of attention due to their massive peak computational capacity. Despite significant advances in the programming interfaces to such hybrid architectures, traditional programming paradigms struggle mapping the resulting multi-dimensional heterogeneity and the expression of algorithm parallelism, resulting in sub-optimal effective performance. Task-based programming paradigms have the capability to alleviate some of the programming challenges on distributed hybrid many-core architectures. In this work we take this concept a step further by showing that the potential of task-based programming paradigms can be greatly increased with minimal modification of the underlying runtime combined with the right algorithmic changes. We propose two novel recursive algorithmic variants for one-sided factorizations and describe the changes to the PaRSEC task-scheduling runtime to build a framework where the task granularity is dynamically adjusted to adapt the degree of available parallelism and kernel efficiency according to runtime conditions. Based on an extensive set of results we show that, with one-sided factorizations, i.e. Cholesky and QR, a carefully written algorithm, supported by an adaptive tasks-based runtime, is capable of reaching a degree of performance and scalability never achieved before in distributed hybrid environments.
These contributions will be presented at the international conference IPDPS 2015 [34] in Hyderabad.

7.1.3.1. Comparison of Static and Dynamic Resource Allocation Strategies for Matrix Multiplication

The tremendous increase in the size and heterogeneity of supercomputers makes it very difficult to predict the performance of a scheduling algorithm. In this context, relying on purely static scheduling and resource allocation strategies, that make scheduling and allocation decisions based on the dependency graph and the platform description, is expected to lead to large and unpredictable makespans whenever the behavior of the platform does not match the predictions. For this reason, the common practice in most runtime libraries is to rely on purely dynamic scheduling strategies, that make short-sighted scheduling decisions at runtime based on the estimations of the duration of the different tasks on the different available resources and on the state of the machine. In this work, we considered the special case of Matrix Multiplication, for which a number of static allocation algorithms to minimize the amount of communications have been proposed. Through a set of extensive simulations, we analyzed the behavior of static, dynamic, and hybrid strategies, and we assessed the possible benefits of introducing more static knowledge and allocation decisions in runtime libraries. These contributions have been presented at the international conference SBAC-PAD 2015.

7.1.3.2. Scheduling Trees of Malleable Tasks for Sparse Linear Algebra

Scientific workloads are often described as directed acyclic task graphs. In this paper, we focus on the multifrontal factorization of sparse matrices, whose task graph is structured as a tree of parallel tasks. Among the existing models for parallel tasks, the concept of *malleable* tasks is especially powerful as it allows each task to be processed on a time-varying number of processors. Following the model advocated by Prasanna and Musicus for matrix computations, we considered malleable tasks whose speedup is p^α, where p is the fractional share of processors on which a task executes, and α ($0 < \alpha \leq 1$) is a parameter which does not depend on the task. We first motivated the relevance of this model for our application with actual experiments on multicore platforms. Then, we studied the optimal allocation proposed by Prasanna and Musicus for makespan minimization using optimal control theory. We largely simplified their proofs by resorting only to pure scheduling arguments. Building on the insight gained thanks to these new proofs, we extended the study to distributed multicore platforms. There, a task cannot be distributed among several distributed nodes. In such a distributed setting (homogeneous or heterogeneous), we proved the NP-completeness of the corresponding scheduling problem, and proposed some approximation algorithms. We finally assessed the relevance of our approach by simulations on realistic trees. We showed that the average performance gain of our allocations with respect to existing solutions (that are thus unaware of the actual speedup functions) is up to 16% for $\alpha = 0.9$ (the value observed in the real experiments). These contributions have been presented at the international conference Europar 2015.

7.1.3.3. Task-based multifrontal QR solver for GPU-accelerated multicore architectures

Recent studies have shown the potential of task-based programming paradigms for implementing robust, scalable sparse direct solvers for modern computing platforms. Yet, designing task flows that efficiently exploit heterogeneous architectures remains highly challenging. In this work we first tackled the issue of data partitioning using a method suited for heterogeneous platforms. On the one hand, we designed task of sufficiently large granularity to obtain a good acceleration factor on GPU. On the other hand, we limited the size in order to both fit the GPU memory constraints and generate enough parallelism in the task graph. Secondly we handled the task scheduling with a strategy capable of taking into account workload and architecture heterogeneity at a reduced cost. Finally we proposed an original evaluation of the performance obtained in our solver on a test set of matrices. We showed that the proposed approach allows for processing extremely large input problems on GPU-accelerated platforms and that the overall performance is competitive with equivalent state of the art solvers designed and optimized for GPU-only use. These contributions have been presented at the international conference HiPC 2015 where they received the best paper award.

7.1.3.4. Fast and Accurate Simulation of Multithreaded Sparse Linear Algebra Solvers

The ever growing complexity and scale of parallel architectures imposes to rewrite classical monolithic HPC scientific applications and libraries as their portability and performance optimization only comes at a prohibitive cost. There is thus a recent and general trend in using instead a modular approach where numerical algorithms are written at a high level independently of the hardware architecture as Directed Acyclic Graphs (DAG) of tasks. A task-based runtime system then dynamically schedules the resulting DAG on the different computing resources, automatically taking care of data movement and taking into account the possible speed heterogeneity and variability. Evaluating the performance of such complex and dynamic systems is extremely challenging especially for irregular codes. In this work, we explained how we crafted a faithful simulation, both in terms of performance and memory usage, of the behavior of qr_mumps, a fully-featured sparse linear algebra library, on multi-core architectures. In our approach, the target high-end machines are calibrated only once to derive sound performance models. These models can then be used at will to quickly predict and study in a reproducible way the performance of such irregular and resource-demanding applications using solely a commodity laptop. These contributions have been presented at the international conference ICPADS 2015.

7.2. High performance solvers for large linear algebra problems

7.2.1. Divide and conquer symmetric tridiagonal eigensolver for multicore architectures

Computing eigenpairs of a symmetric matrix is a problem arising in many industrial applications, including quantum physics and finite-elements computation for automobiles. A classical approach is to reduce the matrix to tridiagonal form before computing eigenpairs of the tridiagonal matrix. Then, a back-transformation allows one to obtain the final solution. Parallelism issues of the reduction stage have already been tackled in different shared-memory libraries. In this work, we focus on solving the tridiagonal eigenproblem, and we describe a novel implementation of the Divide and Conquer algorithm. The algorithm is expressed as a sequential task-flow, scheduled in an out-of-order fashion by a dynamic runtime which allows the programmer to play with tasks granularity. The resulting implementation is between two and five times faster than the equivalent routine from the INTEL MKL library, and outperforms the best MRRR implementation for many matrices.
These contributions have be presented at the international conference IPDPS 2015 [32] in Hyderabad.

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

Solving sparse linear systems is a problem that arises in many scientific applications, and sparse direct solvers are a time consuming and key kernel to those applications or more advanced solvers such as hybrid direct-iterative solvers. That is why optimizing their performance on modern architectures is a crucial problem. The preprocessing steps of sparse direct solvers: ordering and symbolic factorization, are two major steps that lead to a reduced amount of computation and memory, and to a better task granularity to reach a good level of performance when using `BLAS` kernels. With the advent of GPUs, the granularity of the symbolic factorization

became more important than ever. In this work, we present a reordering strategy that increases the block granularity. This strategy relies on the symbolic factorization to refine the ordering produced by tools such as METIS or Scotch, and does not impact the number of operations required to solve the problem. We integrated this algorithm in the PaStiX solver and show a reduction of the number of off-diagonal blocks by two to three on a large spectrum of matrices. This improvement leads to an efficiency on GPUs raised by up to 40%. These contributions have be presented at the Sparse Days [51] in Saint-Girons.

7.2.3. On the use of $\mathcal{H}$-Matrix Arithmetic in PaStiX: a Preliminary Study

The objective is to investigate innovative lowrank approximations based on $\mathcal{H}$matrix variants for direct solver and Schur complements. The intent is to improve scalability of those components involved in preconditioners and hybrid solvers by reducing the computational and memory costs of the dense calculation. The quality of hybrid ordering algorithms combining topdown (such as nested dissection) and bottomup (such as minimum degree) ordering techniques in the context of sparse linear solvers will be investigated.

In this work, we describe a preliminary fast direct solver using HODLR library to compress large blocks appearing in the symbolic structure of the PaStiX sparse direct solver. We present our general strategy before analyzing the practical gains in terms of memory and floating point operations with respect to a theoretical study of the problem. Finally, we discuss ways to enhance the overall performance of the solver.

Some contributions have already been presented at the Workshop on Fast Solvers [52] in Toulouse. This work is a joint effort between Professor Darve's group at Stanford and the Inria HiePACS team within FASTLA.

7.2.4. Data sparse techniques for parallel hybrid solvers

In this work we describe how data sparse techniques exploiting $\mathcal{H}$-matrix calculations can be implemented in a parallel hybrid sparse linear solver based on an algebraic non overlapping domain decomposition approach.

Various graph-based clustering techniques to approximate the local Schur complements are investigated, with the aim of optimally complying with the interface structure of the local interfaces of the subdomains. We consider strong-hierarchical (sH) matrix arithmetic as efficient means for obtaining low rank approximations in terms of workload distribution as well as memory consumption. We also show how sH-arithmetic can be utilized to form an effective global preconditioner for the iterative phase of the hybrid solver. Numerical and parallel experiments are presented to evaluate the advantages and drawbacks of the different variants.

This work is a joint effort between Professor Darve's group at Stanford and the Inria HiePACS team within FASTLA. Some intermediate progresses have already been presented [38], [37]

7.2.5. Analysis of the rounding error accumulation in Conjugate Gradient to improve the maximal attainable accuracy of pipelined CG

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

This research effort was conduced in collaboration with colleagues S. Cools and W. Vanroose from the Applied Mathematics Group of Antwerp university within the framework of the EXA2CT project.

7.3. High performance Fast Multipole Method for N-body problems

7.3.1. Task-based Fast Multipole Method

Last year we have worked primarily on developing an efficient fast multipole method for heterogeneous architecture. Some of the accomplishments for this year include:

1. We have finalized the Uniform FMM (ufmm) based on polynomial interpolations combined with a hierarchical (data sparse) representation of a kernel matrix. The algorithm is close to the Black Box FMM by Fong and Darve developed with Chebyschev polynomials, however it uses an interpolation scheme based on an equispaced grid, which allows the use of FFT and consequently reduce both running time and memory footprint but has implications on accuracy and stability. The theory behind the Uniform FMM kernel is explained in a research report [63] along with numerical benchmarks on artificial test cases and presented in [44] . . This new kernel was extended to be used for dislocation kernel.
2. Concerning the Group-Tree approach, we have shown in past studies its advantages of the task-based FMM and how the group-tree is well suited for runtime systems. In fact, it improves the locality, but it also reduces the number of dependencies which is an important asset to decrease the runtime overhead. These prospective task-based FMM can solve problems on heterogeneous architecture as presented in [36]. Therefore, we have continued this work and created a robust group-tree that has been included in ScalFMM and which is now available to the community. This data structure is generic and can be used with the different ScalFMM kernels. Moreover, we have extended our work and implemented a distributed task-based FMM above StarPU. The description of the data structure and some experimental studies will be presented in February 2016 during PhD defense of B. Bérenger.
3. With the advent of complex modern architectures, the low-level paradigms long sufficient to build high performance computing (HPC) numerical codes have met their limits. Achieving efficiency, ensuring portability, while pre- serving programming tractability on such hardware prompted the HPC community to design new, higher level paradigms. Indeed, several robust runtime systems proposed recently have shown the benefit of task-based parallelism models in terms of performance portability on complex platforms, on top of which full-featured numerical libraries have been ported successfully. However, the common weakness of these projects is to deeply tie applications to specific expert-only runtime system APIs. The OPENMP specification, which aims at providing a common parallel programming means for shared-memory platforms, appears a good candidate to address this issue thanks to the latest task-based constructs introduced as part of its revision 4.0. The goal of this joint work with STORM team is to assess the effectiveness and limits of this support for designing a high-performance numerical library like `ScalFMM` library, which implements state-of-the-art fast multipole methods (FMM) algorithms and that we have considerably re-designed with respect to the most advanced features provided by OPENMP 4.0. We show that OPENMP 4.0 allows for significant performance improvements over previous OPENMP revisions on recent multicore processors. We furthermore propose extensions to the OPENMP 4.0 standard and show how they could enhance FMM performance. To assess our statement, we have implemented this support within the KLANG-OMP source-to-source compiler that translates OPENMP directives into calls to the `StarPU` task-based runtime system. This study shows that we can take advantage of the advanced capabilities of a fully-featured runtime system without resorting to a specific, native runtime port, hence bridging the gap between the OPENMP standard and the very high performance that was so far reserved to expert-only runtime system APIs.

7.3.2. Time-domain boundary element method

The Time-domain Boundary Element Method (TD-BEM) has not been widely studied but represents an interesting alternative to its frequency counterpart. Usually based on inefficient Sparse Matrix Vector-product (SpMV), we investigate other approaches in order to increase the sequential flop-rate.

The TD-BEM formulation we is naturally expressed using sparse-matrix vector product (SpMV). We describe how the Flop-rate can be improved using a so-called multi-vectors/vector product, and we provide an efficient implementation of this operation using vectorization. We have extended our TD-BEM solver to support NVidia GPUs, and we have looked at different blocking schemes and their respective implementations. We have created a new blocking storage which matches our operators and allows to obtain a high Flop-rate. In addition, we provide a balancing heuristic to divide the work between the CPUs and the GPUs dynamically. The results have been published in [20], and our solver is now able to work on distributed heterogeneous nodes.

Our TD-BEM solver is efficient, but it still has a quadratic complexity which might become a problem for large problems. This high complexity motivates the study of an FMM based TD-BEM solver with the objective of being more competitive as the problem size increases. Therefore, we have implemented an FMM-based solver but while the complexity should be lower than the matrix approach, it remains unclear from which problem size. Moreover, we show in [PhD defense of B. Bérenger] different results and point-out that the memory cost is much more expensive for the FMM approach compare to the matrix one. The method has been discussed in [43] among other ScalFMM applications.

All the implementations should be in high quality in the Software Engineering sense since the resulting library is going to be used by industrial applications.
This work is developed in the framework of Bérenger Bramas's PhD and contributes to the EADS-ASTRIUM, Inria, Conseil Régional initiative.

7.3.3. Randomized algorithms for covariance matrices

7.3.3.1. Covariance kernel matrices

Random projection based Low Rank Approximation (LRA) algorithms such as the randomized SVD produce approximate matrix factorizations in quadratic instead of cubic time in N (N being the matrix size). This complexity can be further improved if fast matrix multiplication is available. A paper explaining our recent advances in fast randomized LRA of covariance kernel matrices using FMM is available as a research report [63] and presented in [44] . In particular, the fast multipole acceleration of the randomized SVD allowed for generating Gaussian random fields on arbitrary grids in linear running time and memory requirements. The code is available in the open source C++ project FMR: https://gforge.inria.fr/projects/fmr, it relies heavily on the ScalFMM library for data structures and fast matrix multiplication.

7.3.3.2. New applications: Data Assimilation and Taxonomy

Many applications like data assimilation (e.g. Kalman Filtering or variational approaches) or biology (e.g. taxonomy) involve covariance matrices that are only known in algebraic form, as opposed to kernel matrices that can be explicitly build given a kernel function. In a joint project (called FastMDS) with Alain Franc (INRA, Inria PLEIADE) addressing fast methods for the classification of biological species (taxonomy) our randomized SVD algorithm was used in order to accelerate a MultiDimensionalScaling (MDS) algorithm. The MDS is a widely used method in machine learning and data analysis that aim at visualizing the information contained in a distance matrix. Our MDS algorithm is applied to DNA sequences coming from various sources (e.g. Leman's lake), it consists in forming an euclidian image of the sample by taking the square root of a covariance matrix computed from the distance matrix. The randomized SVD approach lead to promising results, since it allowed to treat up to 100.000 samples in a few seconds. Since the covariance matrix still needs to be loaded in memory, storage might become problematic for larger samples. Therefore we are now considering matrix-free methods in order to decrease the memory requirements but also hierarchical algorithms in order to compute the MDS in near-linear time. The following methods are currently under investigation:

- Random column selection based LRA methods such as the Nystrom method or blocked variant of the Nystrom method (BBF, see Wang, Darve, Mahoney).
- Random projection based LRA powered by general H2-methods.

All these techniques are considered since they apply well, when the relevant information is spread uniformly among the data, just like in our data sets.

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

7.4.1. Dynamic load balancing for massively parallel coupled codes

In the field of scientific computing, load balancing is a major issue that determines the performance of parallel applications. Nowadays, simulations of real-life problems are becoming more and more complex, involving numerous coupled codes, representing different models. In this context, reaching high performance can be a great challenge. In the PhD of Maria Predari (started in october 2013), we develop new graph partitioning techniques, called co-partitioning, that address the problem of load balancing for two coupled codes: the key idea is to perform a *coupling-aware* partitioning, instead of partitioning these codes independently, as it is usually done. However, our co-partitioning technique requires to use graph partitioning with *fixed vertices*, that raises serious issues with state-of-the-art software, that are classically based on the well-known recursive bisection paradigm (RB). Indeed, the RB method often fails to produce partitions of good quality. To overcome this issue, we propose a *new* direct k-way greedy graph growing algorithm, called KGGGP, that overcomes this issue and succeeds to produce partition with better quality than RB while respecting the constraint of fixed vertices. Experimental results compare KGGGP against state-of-the-art methods for graphs available from the popular *DIMACS'10* collection. This work will be presented in the 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP 2016).

7.5. Application Domains

7.5.1. Material physics

7.5.1.1. Molecular Vibrational Spectroscopy

Quantum chemistry eigenvalue problem is a big challenge in recent research. Here we are interested in solving eigenvalue problems coming from the molecular vibrational analysis. These problems are challenging because the size of the vibrational Hamiltonian matrix to be diagonalized is exponentially increasing with the size of the molecule we are studying. So, for molecules bigger than 10 atoms the actual existent algorithms suffer from a curse of dimensionality or computational time. We propose a new variational algorithm (namely residue-based adaptive vibrational configuration interaction) intended for the resolution of the vibrational Schrödinger equation. The main advantage of this approach is to efficiently reduce the dimension of the active space generated into the configuration interaction (CI) process. This adaptive algorithm is developed with the use of three correlated conditions i.e. a suitable starting space ; a criterion for convergence, and a procedure to expand the approximate space. The speed of the algorithm was increased with the use of a posteriori error estimator (residue) to select the most relevant direction to increase the space. Two examples have been selected for benchmark. In the case of Formalde hydemolecule (H_2CO) with a dimension space of 6, we mainly study the performance of RA-VCI algorithm: comparison with the variation-perturbation method, choice of the initial space, residual contributions. For Acetonitrile molecule (CH_3CN) with dimension space of 12 the active space computed by our algorithm is deivided by 20 compared to the computations sone by Avila et. al using the same potential energy surface. This work was presented in [54], [53].

7.5.1.2. Dislocations

7.5.1.2.1. Direct evaluation of the anisotropic elastic force field

The anisotropic elastic force field created by dislocations is not explicitly given, in fact it is only known in integral form using Green's or Stroh's formalism. The approach considered in OptiDis is based on Stroh's formalism, i.e. we compute the stress field using tensorial angular functions known as Stroh matrices. A benefit of using Stroh's formalism is that it only requires the evaluation of a single line integral for the force field and no integration for the stress field, while Green's formalism involve double and single line integral respectively. The evaluation of Stroh matrices in arbitrary directions is not affordable, therefore spherical harmonic expansions were considered in order to approximate the stress field efficiently. Until now the integration of the stress field on target dislocations was performed numerically using simple quadratures,

although the quadrature size required to evaluate the force field at a given precision may explode as segments get closer and computation may become untractable. In order to avoid this behaviour, we developed semi-analytical expressions of the force field based on the analytic integration of the expansions of the stress field (in spherical harmonics). This new method is an adaptation of Aubry et al. approach to Stroh's formalism, in the sense that it also provides optimized recursive formulae to efficiently evaluate these semi-analytic expressions. Numerous verifications and further improvements of the expressions are required before implementing it inside OptiDis.

7.5.1.2.2. Parallel dislocation dynamics simulation

We have focused on the improvements of our hybrid MPI-OpenMP parallelism of the OptiDis code. More precisely, we have continued the development of parallel algorithm to add/remove element in the cache-conscious data structure. This data structured combined with an octree manages efficiently large set of data (segments and nodes) during all the steps of the algorithm. Moreover, we have tuned and improved our hybrid MPI-OpenMP parallelism to run simulations with large number of radiation induced defects forming our dislocation network. To obtain a good scalability, we have introduced a better load balancing at thread level as well as process level. By combining efficient data structure and hybrid parallelism we obtained a speedup of 112 on 160 cores for a simulation of half a million of segments.

All this work was developped in the Phd of A. Etchevery.

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

7.5.2.1. MHD instabilities edge localized modes

The last contribution of Xavier Lacoste's thesis deals with the integration of our work in JOREK, a production controlled plasma fusion simulation code from CEA Cadarache. We described a generic finite element oriented distributed matrix assembly and solver management API. The goal of this API is to optimize and simplify the construction of a distributed matrix which, given as an input to PaStiX, can improve the memory scaling of the application. Experiments exhibit that using this API we could reduce the memory consumption by moving to a distributed matrix input and improve the performance of the factorized matrix assembly by reducing the volume of communication. All this study is related to PaStiX integration inside JOREK but the same API could be used to produce a distributed assembly for another solver or/and another finite elements based simulation code.

7.5.2.2. Turbulence of plasma particules inside a tokamak

Concerning the GYSELA global non-linear electrostatic code, the efforts during the period have concentrated on predicting memory requirement and on the gyroaverage operator.

The Gysela program uses a mesh of 5 dimensions of the phase space (3 dimensions in configuration space and 2 dimensions in velocity space). On the large cases, the memory consumption already reaches the limit of the available memory on the supercomputers used in production (Tier-1 and Tier-0 typically). Furthermore, to implement the next features of Gysela (e.g. adding kinetic electrons in addition to ions), the needs of memory will dramatically increase, the main unknown will represents hundreds of TB. In this context, two tools were created to analyze and decrease the memory consumption. The first one is a tool that plots the memory consumption of the code during a run. This tool helps the developer to localize where the memory peak is located. The second tool is a prediction tool to compute the peak memory in offline mode (for production use mainly). A post processing stage combined with some specific traces generated on purpose during runtime allow the analysis of the memory consumption. Low-level primitives are called to generate these traces and to model memory consumption : they are included in the libMTM library (Modeling and Tracing Memory). Thanks to this work on memory consumption modeling, we have decreased the memory peak of the Gysela code up to 50 % on a large case using 32,768 cores and memory scalability improvement has been shown using these tools up to 65k cores.

The main unknown of the Gysela is a distribution function that represents either the density of the guiding centers, either the density of the particles in a tokamak (depending of the location in the code). The switch between these two representations is done thanks to the gyroaverage operator. In the actual version of Gysela, the computation of this operator is achieved thanks to the so-called Padé approximation. In order to improve the precision of the gyroaveraging, a new implementation based on interpolation methods has been done (mainly by researchers from the Inria Tonus project-team and IPP Garching). We have performed the integration of this new implementation in Gysela and also some parallel benchmarks. However, the new gyroaverage operator is approximatively 10 times slower than the original one. Investigations and optimizations on this operator are still a work in progress.
This work has been carried on in the framework of Fabien Rozar's PhD in collaboration with CEA Cadarache (defended in November 2015). A new PhD (Nicolas Bouzat) has started in October 2015 and the scientific objectives of this work will be first to consolidate the parallel version of the gyroaverage operator, in particular by designing a complete MPI+OpenMP parallel version, and then to design new numerical methods for the gyroaverage, source and collision operators to deal with new physics in Gysela. The objective is to tackle kinetic electron configurations for more realistic simulations.

7.5.2.3. SN Cartesian solver for nuclear core simulation

High-fidelity nuclear power plant core simulations require solving the Boltzmann transport equation. In discrete ordinate methods, the most computationally demanding operation of this equation is the sweep operation. Considering the evolution of computer architectures, we propose in this work, as a first step toward heterogeneous distributed architectures, a hybrid parallel implementation of the sweep operation on top of the generic task-based runtime system: PaRSEC. Such an implementation targets three nested levels of parallelism: message passing, multi-threading, and vectorization. A theoretical performance model was designed to validate the approach and help the tuning of the multiple parameters involved in such an approach. The proposed parallel implementation of the Sweep achieves a sustained performance of 6.1 Tflop/s, corresponding to 33.9% of the peak performance of the targeted supercomputer. This implementation compares favorably with state-of-art solvers such as PARTISN; and it can therefore serve as a building block for a massively parallel version of the neutron transport solver DOMINO developed at EDF.
The main contribution has be presented at the international conference IPDPS 2015 [31] in Hyderabad.

7.5.2.4. 3D aerodynamics for unsteady problems with moving bodies

In the first part of our research work concerning the parallel aerodynamic code FLUSEPA, a first OpenMP-MPI version based on the previous one has been developped. By using an hybrid approach based on a domain decomposition, we achieved a faster version of the code and the temporal adaptive method used without bodies in relative motion has been tested successfully for real complex 3D-cases using up to 400 cores. Moreover, an asynchronous strategy for computing bodies in relative motion and mesh intersections has been developed and has been used for actual 3D-cases. A journal article (for JCP) to sum-up this part of the work is under redaction and a presentation at ISC at the "2nd International Workshop on High Performance Computing Simulation in Energy/Transport Domains" on July 2015 is scheduled.

This intermediate version exhibited synchronization problems for the aerodynamic solver due to the time integration used by the code. To tackle this issue, a task-based version over the runtime system StarPU is currently under development and evaluation. This year was mainly devoted to the realisation of this version. Task generation function have been designed in order to maximize asynchronism in execution. Those functions respect the data pattern access of the code and led to the refactorization of the actual kernels. A task-based version is now available for the aerodynamic solver and is available for both shared and distributed memory. This work has been presented as a poster during the SIAM CSE'15 conference and at the Parallel CFD'15 and HPCSET'15 conferences.

The next steps will be to validate the correction of this task-based version and to work on the performance of this new version on actual cases. Later, the task description should be extended to the motion and intersection operations.
This work is carried on in the framework of Jean-Marie Couteyen's PhD in collaboration with Airbus Defence and Space.

7.5.2.5. Spectral recycling strategies for the solution of nonlinear eigenproblems in thermoacoustics

In this work we consider the numerical solution of large nonlinear eigenvalue problems that arise in thermoacoustic simulations involved in the stability analysis of large combustion devices. We briefly introduce the physical modeling that leads to a nonlinear eigenvalue problem that is solved using a nonlinear fixed point iteration scheme. Each step of this nonlinear method requires the solution of a complex non-Hermitian linear eigenvalue problem. We review a set of state of the art eigensolvers and discuss strategies to recycle spectral information from one nonlinear step to the next. More precisely, we consider the Jacobi-Davidson algorithm, the Implicitly Restarted Arnoldi method, the Krylov-Schur solver and its block-variant, as well as the subspace iteration method with Chebyshev acceleration. On a small test example we study the relevance of the different approaches and illustrate on a large industrial test case the performance of the parallel solvers best suited to recycle spectral information for large scale thermoacoustic stability analysis.

The results of this work conducted in collaboration with S. Moreau (Sherbrooke University) and Y. Saad (University of Minnesota Twin-cities) are detailed in [22]

7.5.2.6. A conservative 2-D advection model towards large-scale parallel calculation

To exploit the possibilities of parallel computers, we designed a large-scale bidimensional atmospheric advection model named Pangolin. As the basis for a future chemistry-transport model, a finite-volume approach for advection was chosen to ensure mass preservation and to ease parallelization. To overcome the pole restriction on time steps for a regular latitude–longitude grid, Pangolin uses a quasi-area-preserving reduced latitude–longitude grid. The features of the regular grid are exploited to reduce the memory footprint and enable effective parallel performances. In addition, a custom domain decomposition algorithm is presented. To assess the validity of the advection scheme, its results are compared with state-of-the-art models on algebraic test cases. Finally, parallel performances are shown in terms of strong scaling and confirm the efficient scalability up to a few hundred cores

The results of this work are detailed in [21].

KERDATA Project-Team

7. New Results

7.1. Efficient data management for hybrid and multi-site clouds

7.1.1. JetStream: enabling high-throughput live event streaming on multi-site clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Scientific and commercial applications operate nowadays on tens of cloud datacenters around the globe, following similar patterns: they aggregate monitoring or sensor data, assess the QoS or run global data mining queries based on inter-site event stream processing. Enabling fast data transfers across geographically distributed sites allows such applications to manage the continuous streams of events in real time and quickly react to changes. However, traditional event processing engines often consider data resources as second-class citizens and support access to data only as a side-effect of computation (i.e. they are not concerned by the transfer of events from their source to the processing site). This is an efficient approach as long as the processing is executed in a single cluster where nodes are interconnected by low latency networks. In a distributed environment, consisting of multiple datacenters, with orders of magnitude differences in capabilities and connected by a WAN, this will undoubtedly lead to significant latency and performance variations.

This is namely the challenge we addressed this year by proposing JetStream [15], a high performance batch-based streaming middleware for efficient transfers of events between cloud datacenters. JetStream is able to self-adapt to the streaming conditions by modeling and monitoring a set of context parameters. It further aggregates the available bandwidth by enabling multi-route streaming across cloud sites, while at the same time optimizing resource utilization and increasing cost efficiency. The prototype was validated on tens of nodes from US and Europe datacenters of the Windows Azure cloud with synthetic benchmarks and a real-life application monitoring the ALICE experiment at CERN. The results show a $3\times$ increase of the transfer rate using the adaptive multi-route streaming, compared to state of the art solutions.

7.1.2. Multi-site metadata management for geographically distributed cloud workflows

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

With their globally distributed datacenters, clouds now provide an opportunity to run complex large-scale applications on dynamically provisioned, networked and federated infrastructures. However, there is a lack of tools supporting data-intensive applications (e.g. scientific workflows) on virtualized IaaS or PaaS systems across geographically distributed sites. As a relevant example, data-intensive scientific workflows struggle in leveraging such distributed cloud platforms. For instance, scientific workflows which handle many small files can easily saturate state-of-the-art distributed filesystems based on centralized metadata servers (e.g., HDFS, PVFS).

In [22], we explore several alternative design strategies to efficiently support the execution of existing workflow engines across multi-site clouds, by reducing the cost of metadata operations. These strategies leverage workflow semantics in a 2-level metadata partitioning hierarchy that combines distribution and replication. The system was validated on the Microsoft Azure cloud across 4 EU and US datacenters. The experiments were conducted on 128 nodes using synthetic benchmarks and real-life applications. We observe as much as 28% gain in execution time for a parallel, geo-distributed real-world application (Montage) and up to 50% for a metadata-intensive synthetic benchmark, compared to a baseline centralized configuration.

7.1.3. Understanding the performance of Big Data platforms in hybrid and multi-site clouds

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

Recently, hybrid multi-site big data analytics (that combines on-premise with off-premise resources) has gained increasing popularity as a tool to process large amounts of data on-demand, without additional capital investment to increase the size of a single datacenter. However, making the most out of hybrid setups for big data analytics is challenging because on-premise resources can communicate with off-premise resources at significantly lower throughput and higher latency. Understanding the impact of this aspect is not trivial, especially in the context of modern big data analytics frameworks that introduce complex communication patterns and are optimized to overlap communication with computation in order to hide data transfer latencies. This year we started to work on a study that aims to identify and explain this impact in relationship to the known behavior on a single cloud.

A first step towards this goal consisted of analysing a representative big data workload on a hybrid Spark setup [24]. Unlike previous experience that emphasized low end-impact of network communications in Spark, we found significant overhead in the shuffle phase when the bandwidth between the on-premise and off-premise resources is sufficiently small. We plan to continue this study by investigating additional parameters at a finer grain and adding new platforms, like Apache Flink.

7.2. Optimizing Map-Reduce

7.2.1. Chronos: failure-aware scheduling in shared Hadoop clusters

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop emerged as the de facto state-of-the-art system for MapReduce-based data analytics. The reliability of Hadoop systems depends in part on how well they handle failures. Currently, Hadoop handles machine failures by re-executing all the tasks of the failed machines (i.e., executing recovery tasks). Unfortunately, this elegant solution is entirely entrusted to the core of Hadoop and hidden from Hadoop schedulers. The unawareness of failures therefore may prevent Hadoop schedulers from operating correctly towards meeting their objectives (e.g., fairness, job priority) and can significantly impact the performance of MapReduce applications.

In [23], we propose Chronos, a failure-aware scheduling strategy that enables an early yet smart action for fast failure recovery while operating within a specific scheduler objective. Chronos takes an early action rather than waiting an uncertain amount of time to get a free slot (thanks to our preemption technique). Chronos embraces a smart selection algorithm that returns a list of tasks that need to be preempted in order to free the necessary slots to launch recovery tasks immediately. This selection considers three criteria: the progress scores of running tasks, the scheduling objectives, and the recovery tasks input data locations. In order to make room for recovery tasks rather than waiting an uncertain amount of time, a natural solution is to kill running tasks in order to create free slots. Although killing tasks can free the slots easily, it wastes the work performed by the killed tasks. Therefore, we present the design and implementation of a novel work-conserving preemption technique that allows pausing and resuming both map and reduce tasks without resource wasting and with little overhead.

We demonstrate the utility of Chronos by combining it with two state-of-the-art Hadoop schedulers: Fifo and Fair schedulers. The experimental results show that Chronos achieves almost optimal data locality for the recovery tasks and reduces the job completion times by up to 55% over state-of-the-art schedulers. Moreover, Chronos recovers to a correct scheduling behavior after failure detection within only a couple of seconds.

7.2.2. On the usability of shortest remaining time first policy in shared Hadoop clusters

Participants: Nathanaël Cheriere, Shadi Ibrahim.

A practical problem facing the Hadoop community is how to reduce job makespans by reducing job waiting times and execution times. Previous Hadoop schedulers have focused on improving job execution times, by improving data locality but not considering job waiting times. Even worse, enforcing data locality according to the job input sizes can be inefficient: it can lead to long waiting times for small yet short jobs when sharing the cluster with jobs with smaller input sizes but higher execution complexity.

We have introduced hSRTF [16], an adaption of the well-known Shortest Remaining Time First scheduler (i.e., SRTF) in shared Hadoop clusters. hSRTF embraces a simple model to estimate the remaining time of a job and a preemption primitive (i.e., kill) to free the resources when needed. We have implemented hSRTF and performed extensive evaluations with Hadoop on the Grid'5000 testbed. The results show that hSRTF can significantly reduce the waiting times of small jobs and therefore improves their make-spans, but at the cost of a relatively small increase in the make-spans of large jobs. For instance, a time-based proportional share mode of hSRTF (i.e., hSRTF-Pr) speeds up small jobs by (on average) 45% and 26% while introducing a performance degradation for large jobs by (on average) 10% and 0.2% compared to Fifo and Fair schedulers, respectively.

7.2.3. A Performance evaluation of Hadoop's schedulers under failures

Participants: Shadi Ibrahim, Gabriel Antoniu.

Recently, Hadoop has not only been used for running single batch jobs but it has also been optimized to simultaneously support the execution of multiple jobs belonging to multiple concurrent users. Several schedulers (i.e., Fifo, Fair, and Capacity schedulers) have been proposed to optimize locality executions of tasks but do not consider failures, although, evidence in the literature shows that faults do occur and can probably result in performance problems.

In [19], we have designed a set of experiments to evaluate the performance of Hadoop under failure when applying several schedulers (i.e., explore the conflict between job scheduling, exposing locality executions, and failures). Our results reveal several drawbacks of current Hadoop's mechanism in prioritizing failed tasks. By trying to launch failed tasks as soon as possible regardless of locality, it significantly increases the execution time of jobs with failed tasks, due to two reasons: 1) available resources might not be freed up as quickly as expected and 2) failed tasks might be re-executed on machines with no data on it, introducing extra cost for data transferring through network, which is normally the most scarce resource in today's datacenters.

Our preliminary study with Hadoop not only helps us to understand the interplay between fault-tolerance and job scheduling, but also offers useful insights into optimizing the current schedulers to be more efficient in case of failures.

7.2.4. Kvasir: empowering Hadoop with knowledge

Participants: Nathanaël Cheriere, Shadi Ibrahim.

Most of Hadoop schedulers are based on homogeneity hypotheses about the jobs and the nodes and therefore strongly rely on the location of the input data when scheduling tasks. However, our study revealed that Hadoop is a highly dynamic environment (e.g., variation in task duration within a job and across different jobs). Even worse, clouds are multi-tenant environments which in turn introduce more heterogeneity and dynamicity in Hadoop clusters. As a result, relying on static knowledge (i.e. data location) may lead to wrong scheduling decisions.

We have developed a new scheduling framework for Hadoop, named Kvasir. Kvasir aims to provide an up-to-date knowledge that reflects the dynamicity of the environment while being light-weight and performance-oriented. The utility of Kvasir is demonstrated by the implementation of several schedulers including Fifo, Fair, and SRTF schedulers.

7.3. Energy-aware data management in clouds and HPC

7.3.1. On understanding the energy impact of speculative execution in Hadoop

Participants: Tien Dat Phan, Shadi Ibrahim, Gabriel Antoniu, Luc Bougé.

Hadoop emerged as an important system for large-scale data analysis. Speculative execution is a key feature in Hadoop that is extensively leveraged in clouds: it is used to mask slow tasks (i.e., stragglers) — resulted from resource contention and heterogeneity in clouds — by launching speculative task copies on other machines. However, speculative execution is not cost-free and may result in performance degradation and extra resource and energy consumption. While prior literature has been dedicated to improving stragglers detection to cope with the inevitable heterogeneity in clouds, little work is focusing on understanding the implications of speculative execution on the performance and energy consumption in Hadoop cluster.

In [21], we have designed a set of experiments to evaluate the impact of speculative execution on the performance and energy consumption of Hadoop in homogeneous and heterogeneous environments. Our studies reveal that speculative execution may sometimes reduce, sometimes increase the energy consumption of Hadoop clusters. This strongly depends on the reduction in the execution time of MapReduce applications and on the extra power consumption introduced by speculative execution. Moreover, we show that the extra power consumption varies among applications and is contributed to by three main factors: the duration of speculative tasks, the idle time, and the allocation of speculative tasks. To the best of our knowledge, our work provides the first deep look into the energy efficiency of speculative execution in Hadoop.

7.3.2. On the energy footprint of I/O management in Exascale HPC systems

Participants: Orçun Yildiz, Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

The advent of unprecedentedly scalable yet energy hungry Exascale supercomputers poses a major challenge in sustaining a high performance-per-watt ratio. With I/O management acquiring a crucial role in supporting scientific simulations, various I/O management approaches have been proposed to achieve high performance and scalability. However, the details of how these approaches affect energy consumption have not been studied yet.

Therefore, we have explored how much energy a supercomputer consumes while running scientific simulations when adopting various I/O management approaches. In particular, we closely examined three radically different I/O schemes including time partitioning, dedicated cores, and dedicated nodes. To do so, we implemented the three approaches within the Damaris I/O middleware and performed extensive experiments with one of the target HPC applications of the Blue Waters sustained-petaflop supercomputer project: the CM1 atmospheric model.

Our experimental results obtained on the French Grid'5000 platform highlighted the differences among these three approaches and illustrate in which way various configurations of the application and of the system can impact performance and energy consumption. Considering that choosing the most energy-efficient approach for a particular simulation on a particular machine can be a daunting task, we provided a model to estimate the energy consumption of a simulation under different I/O approaches. Our proposed model gives hints to pre-select the most energy-efficient I/O approach for a particular simulation on a particular HPC system and therefore provides a step towards energy-efficient HPC simulations in Exascale systems.

We validated the accuracy of our proposed model using a real-life HPC application (CM1) and two different clusters provisioned on the Grid'5000 testbed. The estimated energy consumptions are within 5.7% of the measured ones for all I/O approaches.

7.3.3. Exploring energy-consistency trade-offs in cloud storage systems and beyond

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

Apache Cassandra is an open-source cloud storage system that offers multiple types of operation-level consistency including eventual consistency with multiple levels of guarantees and strong consistency. It is being used by many datacenter applications (e.g., Facebook and AppScale). Most existing research efforts have been dedicated to exploring trade-offs such as: consistency vs. performance, consistency vs. latency and consistency vs. monetary cost. In contrast, a little work is focusing on the consistency vs. energy trade-off. As power bills have become a substantial part of the monetary cost for operating a datacenter, we aim to provide a clearer understanding of the interplay between consistency and energy consumption.

In [17], a series of experiments have been conducted to explore the implication of different factors on the energy consumption in Cassandra. Our experiments have revealed a noticeable variation in the energy consumption depending on the consistency level. Furthermore, for a given consistency level, the energy consumption of Cassandra varies with the access pattern and the load exhibited by the application. This further analysis indicated that the uneven distribution of the load amongst different nodes also impacts the energy consumption in Cassandra. Finally, we experimentally compared the impact of four storage configuration and data partitioning policies on the energy consumption in Cassandra: interestingly, we achieve 23% energy saving when assigning 50% of the nodes to the hot pool for the applications with moderate ratio of reads and writes, while applying eventual (quorum) consistency.

This study points to opportunities for future research on consistency-energy trade-offs and offers useful insight into designing energy-efficient techniques for cloud storage systems. This work was done in collaboration with Houssem-Eddine Chihoub (LIG lab, Grenoble) and María Pérez (UPM, Madrid).

Recently, we have been looking at in-memory storage systems. In particular, we are investigating the current replication schemes, data placement strategies and consistency models which are used in in-memory storage systems. Next, an empirical study will be performed to analyze the potential impact of the aforementioned issues on energy consumption. At this point, we are working with RAMCloud.

7.3.4. Governing energy consumption in Hadoop through CPU frequency scaling: an analysis

Participants: Tien Dat Phan, Shadi Ibrahim, Gabriel Antoniu.

In [12], we studied the impact of different existing DVFS (*Dynamic Voltage and Frequency Scaling*) governors (i.e., performance, powersave, on-demand, conservative and userspace) on Hadoop's performance and power efficiency. Interestingly, our experimental results reported not only a noticeable variation of the power consumption and performance with different applications and under different governors, but also demonstrate the opportunity to achieve a better tradeoff between performance and power consumption.

The primary contributions of this work are as follows: (1) it provides an overview of the state-of-the-art techniques for energy-efficiency in Hadoop; (2) it discusses and demonstrates the need for exploiting DVFS techniques for energy reduction in Hadoop; (3) it experimentally demonstrates that MapReduce applications experience variations in performance and power consumption under different CPU frequencies and also under different governors. A micro-analysis section is provided to explain this variation and its cause; (4) it illustrates in practice how the behavior of different governors influences the execution of MapReduce applications and how it shapes the performance of the entire cluster; (5) it also brings out the differences between these governors and CPU frequencies and shows that they are not only sub-optimal for different applications but also sub-optimal for different stages of MapReduce execution; (6) it demonstrates that achieving better energy efficiency in Hadoop cannot be done simply by tuning the governor parameters, nor through a naive coarse-grained tuning of the CPU frequencies or the governors according to the running phase (i.e., map phase or reduce phase).

7.4. Scalable I/Os: visualization and processing

7.4.1. Modeling and predicting I/O patterns of large-scale simulations

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

The increasing gap between the computation performance of post-petascale machines and the performance of their I/O subsystem has motivated many I/O optimizations including prefetching, caching, and scheduling. In order to further improve these techniques, modeling and predicting spatial and temporal I/O patterns of HPC applications as they run has become crucial. Our work in this context focuses on Omnisc'IO, an approach that builds a grammar-based model of the I/O behavior of HPC applications and uses it to predict when future I/O operations will occur, and where and how much data will be accessed. To infer grammars, Omnisc'IO is based on StarSequitur, a novel algorithm extending Nevill-Manning's Sequitur algorithm [11]. Omnisc'IO is transparently integrated into the POSIX and MPI I/O stacks and does not require any modification in applications or higher-level I/O libraries. It works without any prior knowledge of the application and converges to accurate predictions of any N future I/O operations within a couple of iterations. Its implementation is efficient in both computation time and memory footprint.

7.4.2. In situ analysis and visualization workflows

Participants: Matthieu Dorier, Lokman Rahmani, Gabriel Antoniu.

In situ visualization has been proposed in the past few years to couple running simulations with parallel visualization and analysis tools. While many parallel visualization tools now provide in situ visualization capabilities, the trend has been to feed such tools with what previously was large amounts of unprocessed output data and let them render everything at the highest possible resolution. This leads to an increased run time of simulations that still have to complete within a fixed-length job allocation. In this work, we tackle the challenge of enabling in situ visualization under performance constraints. Our approach shuffles data across processes according to its content and filters out part of it in order to feed a visualization pipeline with only a reorganized subset of the data produced by the simulation. Our framework monitors its own performance and reconfigures itself dynamically to achieve the best possible visual fidelity within predefined performance constraints. Experiments on the Blue Waters supercomputer with the CM1 simulation show that our approach enables a $5\times$ speedup and is able to meet performance constraints.

7.5. Scalable storage for data-intensive applications

7.5.1. OverFlow: multi-site aware Big Data management for scientific workflows on clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

The global deployment of cloud datacenters is enabling large-scale scientific workflows to improve performance and deliver fast responses. This unprecedented geographical distribution of the computation is doubled by an increase in the scale of the data handled by such applications, bringing new challenges related to the efficient data management across sites. High throughput, low latencies or cost-related trade-offs are just a few concerns for both cloud providers and users when it comes to handling data across datacenters. Existing solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemes. In turn, workflow engines need to improvise substitutes, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability.

In [14], we introduced OverFlow, a uniform data-management system for scientific workflows running across geographically distributed sites, aiming to reap economic benefits from this geo-diversity. Our solution is environment-aware, as it monitors and models the global cloud infrastructure, offering high and predictable data-handling performance for transfer cost and time, within and across sites. OverFlow proposes a set of pluggable services, grouped in a data-scientist cloud kit. They provide the applications with the possibility to monitor the underlying infrastructure, to exploit smart data compression, deduplication and geo-replication, to evaluate data-management costs, to set a tradeoff between money and time, and optimize the transfer strategy accordingly. The system was validated on the Microsoft Azure cloud across its 6 EU and US datacenters. The experiments were conducted on hundreds of nodes using synthetic benchmarks and real-life bio-informatics applications (A-Brain, BLAST). The results show that our system is able to model the cloud performance accurately and to leverage this for efficient data dissemination, being able to reduce the monetary costs and transfer time by up to 3 times.

7.5.2. Efficient transactional storage for data-intensive applications

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

As the computational power used by large-scale applications increases, the amount of data they need to manipulate tends to increase as well. A wide range of such applications require robust and flexible storage support for atomic, durable and concurrent transactions. Historically, databases have provided the *de facto* solution to transactional data management, but they have forced applications to drop control over data layout and access mechanisms, while remaining unable to meet the scale requirements of Big Data. More recently, key-value stores have been introduced to address these issues. However, this solution does not provide transactions, or only restricted transaction support, constraining users to carefully coordinate access to data in order to avoid race conditions, partial writes, overwrites, and other hard problems that cause erratic behavior.

We argue that there is a gap between existing storage solutions and application requirements that limits the design of transaction-oriented data-intensive applications. We have started working on a prototype of a massively parallel distributed transactional blob storage system, aiming to fill this gap.

MESCAL Project-Team

6. New Results

6.1. Reproducible Research

In the field of large-scale distributed systems, experimentation is particularly difficult. The studied systems are complex, often non-deterministic and unreliable, software is plagued with bugs, whereas the experiment workflows are unclear and hard to reproduce. In [11], we provide an extensive list of features offered by general-purpose experiment management tools dedicated to distributed systems research on real platforms. We then use it to assess existing solutions and compare them, outlining possible future paths for improvements.

In [20], we address the question of developing a lightweight and effective workflow for conducting experimental research on modern parallel computer systems in a reproducible way. Our workflow simply builds on two well-known tools (Org-mode and Git) and enables us to address issues such as provenance tracking, experimental setup reconstruction, replicable analysis. Although this workflow is perfectible and cannot be seen as a final solution, we have been using git for two years now and we have recently published a fully reproducible article, which demonstrates the effectiveness of our proposal.

6.2. Performance Characterization and Optimization of IOs

In high-performance computing environments, parallel file systems provide a shared storage infrastructure to applications. In the situation where multiple applications access this shared infrastructure concurrently, their performance can be impaired because of interference. In [22], we improve improve performance by alleviating interference effects through a smart I/O scheduler scheduler that organizes and optimizes the applications' requests and adjusts the access pattern to the device characteristics. We apply machine learning techniques to automatically select the best scheduling algorithm for each situation. Our approach improves performance by up to 75

In [33], we present a new storage device profiling tool that characterizes the sequential to random throughput ratio for reads and writes of different sizes. As we explained previously, several optimizations aim at adapting applications' access patterns in order to generate contiguous accesses for improved performance when accessing storage devices like hard disks. However, when considering other storage options like RAID arrays and SSDs, the access time ratio between contiguous and non-contiguous accesses may not compensate for these optimizations' cost. In this scenario, the information provided by our tool could be used to dynamically decide if optimizations are beneficial for performance, which is why we took a particular attention to obtain accurate information in a minimal benchmarking time.

6.3. Application of Game Theory and Distributed Optimization to Wireless Networks

In wireless networks, channel conditions and user quality of service (QoS) requirements vary, often quite arbitrarily, with time (e.g. due to user mobility, fading, etc.) and users only have a very limited information about their environment. In such context optimizing transmission while taking power consumption into account is extremely challenging. We apply game theory technique to MIMO wireless network using OFDM or OFDMA where multi-path channels can be handled efficiently

In [25], [9], we show that distributed power allocation in heterogeneous OFDMA cognitive radio networks can be modeled as a game where each user equipment in the network engages in a non-cooperative game and allocates its available transmit power over subcarriers to maximize its individual utility. The corresponding equilibrium (Debreu, an extension of Nash Equilibrium) can be characterized with fractional programming and we provide sufficient conditions for computing such equilibria as fixed points of a water-filling best response operator. Using such approach can however be quite slow and is very sensitive to delay and information uncertainty (it may not converge). Therefore, we explain in [17] how signal covariance matrices in Gaussian MIMO multiple access channel can be learnt in presence of imperfect (and possibly delayed) feedback. The algorithm we propose is based on on the method of matrix exponential learning (MXL) and it has the same information and computation requirements as distributed water-filling. However our algorithm converge much faster even for large numbers of users and/or antennas per user and in the presence of user update asynchrony, random delays and/or ergodically changing channel conditions. Yet, since the system may evolve over time in an unpredictable fashion (e.g. due to changes in the wireless medium or the users' QoS requirements), static solution concepts (such as Nash equilibrium) may be no longer relevant and users must adapt to changes in the environment "on the fly", without being able to predict the system's evolution ahead of time. Hence, we focus on the concept of no-regret : policies that perform at least as well as the best fixed transmit profile in hindsight. In [31] and [41], we provide a formulation of power control as an online optimization problem and we show that the FM dynamics lead to no regret in this dynamic context. In [40] we apply this approach energy efficient transmission in MIMO-OFDM systems and we show through numerical simulations that, in realistic network environments even under rapidly changing channel conditions, users can track their individually optimum transmit profile, achieving gains of up to 600in energy efficiency over uniform power allocation policies.

We also apply this technique to multi-carrier cognitive radio systems. Such systems allow opportunistic secondary users (SUs) to access portions of the spectrum that are unused by the network's licensed primary users (PUs), provided that the induced interference does not compromise the PUs' performance guarantees. In [14], we introduce a flexible spectrum access pricing schemes such that the corresponding Nash equilibrium is unique under very mild assumptions and satisfies the performance constraints. In addition, we derive a dynamic power allocation policy that converges to equilibrium within a few iterations (even for large numbers of users) and that relies only on local—and possibly imperfect—signal-to-interference-and-noise ratio measurements. In [24], we draw on exponential learning techniques to design an algorithm that is able to adapt to system changes "on the fly", i.e. such that the proposed transmit policy leads to no regret even under rapidly changing network conditions.

6.4. General Results in Game Theory

Our work on game theory is often motivated by applications to wireless networks but can often have a more general application.

In [38], motivated by applications to multi-antenna wireless networks, we propose a distributed and asynchronous algorithm for stochastic semidefinite programming This algorithm is a stochastic approximation of a continous-time matrix exponential scheme regularized by the addition of an entropy-like term to the problem's objective function. We show that the resulting algorithm converges almost surely to an (ϵ)-approximation of the optimal solution requiring only an unbiased estimate of the gradient of the problem's stochastic objective.

As explained in the previous section, classical Nash equilibrium concepts become irrelevant in situations where the environment evolves over time. In [15], we study one of the main concept of online learning and sequential decision problem known as regret minimization. Our objective is to provide a quick overview and a comprehensive introduction to online learning and game theory.

In practice, it is rarely reasonable to assume that players have access to the strategy of the others and implementing a best response can thus become cumbersome. Replicator dynamics is a fundamental approach in evolutionary game theory in which players adjust their strategies based on their actions' cumulative payoffs over time – specifically, by playing mixed strategies that maximize their expected cumulative payoff.

- In [19], we investigate the impact of payoff shocks on the evolution of large populations of myopic players that employ simple strategy revision protocols such as the "imitation of success". In the

noiseless case, this process is governed by the standard (deterministic) replicator dynamics; in the presence of noise however, the induced stochastic dynamics are different from previous versions of the stochastic replicator dynamics (such as the aggregate-shocks model of Fudenberg and Harris, 1992). In this context, we show that strict equilibria are always stochastically asymptotically stable, irrespective of the magnitude of the shocks; on the other hand, in the high-noise regime, non-equilibrium states may also become stochastically asymptotically stable and dominated strategies may survive in perpetuity (they become extinct if the noise is low). Such behavior is eliminated if players are less myopic and revise their strategies based on their cumulative payoffs. In this case, we obtain a second order stochastic dynamical system whose attracting states coincide with the game's strict equilibria and where dominated strategies become extinct (a.s.), no matter the noise level.

- In [13], we study a new class of continuous-time learning dynamics consisting of a replicator-like drift adjusted by a penalty term that renders the boundary of the game's strategy space repelling. These penalty-regulated dynamics are equivalent to players keeping an exponentially discounted aggregate of their ongoing payoffs and then using a smooth best response to pick an action based on these performance scores. Building on the duality with evolutionary game theory, we design a discrete-time, payoff-based learning algorithm that converges to (arbitrarily precise) approximations of Nash equilibria in potential games. Moreover, the algorithm remains robust in the presence of stochastic perturbations and observation errors, and it does not require any synchronization between players, which is a very important property when applying such technique to traffic engineering.
- In [18], we investigate an other class of reinforcement learning dynamics in which the players strategy adjustment is regularized with a strongly convex penalty term. In contrast to the class of penalty functions used to define smooth best responses in models of stochastic fictitious play, the regularizers used in this paper need not be infinitely steep at the boundary of the simplex. Dropping this requirement gives rise to an important dichotomy between steep and non-steep cases. In this general setting, our main results extend several properties of the replicator dynamics such as the elimination of dominated strategies, the asymptotic stability of strict Nash equilibria and the convergence of time-averaged trajectories to interior Nash equilibria in zero-sum games.
- In [37], we study a general class of game-theoretic learning dynamics in the presence of random payoff disturbances and observation noise, and we provide a unified framework that extends several rationality properties of the (stochastic) replicator dynamics and other game dynamics. In the unilateral case, we show that the stochastic dynamics under study lead to no regret, irrespective of the noise level. In the multi-player case, we find that dominated strategies become extinct (a.s.) and strict Nash equilibria remain stochastically asymptotically stable – again, independently of the perturbations' magnitude. Finally, we establish an averaging principle for 2-player games and we show that the empirical distribution of play converges to Nash equilibrium in zero-sum games under any noise level.

6.5. Simulation

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

- Indeed, multi-core architectures comprising several GPUs have become mainstream but obtaining the maximum performance of such heterogeneous machines is challenging as it requires to carefully offload computations and manage data movements between the different processing units. The most promising and successful approaches so far build on task-based runtimes that abstract the machine and rely on opportunistic scheduling algorithms. As a consequence, the problem gets shifted to choosing the task granularity, task graph structure, and optimizing the scheduling strategies. Trying different combinations of these different alternatives is also itself a challenge. Indeed, getting accurate measurements requires reserving the target system for the whole duration of experiments. Furthermore, observations are limited to the few available systems at hand and may be difficult

to generalize. In [21], we show how we crafted a coarse-grain hybrid simulation/emulation of StarPU, a dynamic runtime for hybrid architectures, over SimGrid. This approach allows to obtain performance predictions of classical dense linear algebra kernels accurate within a few percents and in a matter of seconds, which allows both runtime and application designers to quickly decide which optimization to enable or whether it is worth investing in higher-end GPUs or not. Additionally, it allows to conduct robust and extensive scheduling studies in a controlled environment whose characteristics are very close to real platforms while having reproducible behavior. In [30], we have extended this approach to the simulation of a multithreaded multifrontal QR solver of sparse matrices: QR-MUMPS. In our approach, the target high-end machines are calibrated only once to derive sound performance models. These models can then be used at will to quickly predict and study in a reproducible way the performance of such irregular and resource-demanding applications using solely a commodity laptop. Our approach also allows to study the memory consumption along time, which is a critical factor for such applications.

- Beside the inherent heterogeneity of distributed computing infrastructures, storage is also a essential component to cope with the tremendous increase in scientific data production and the ever-growing need for data analysis and preservation. Understanding the performance of a storage subsystem or dimensioning it properly is an important concern for which simulation can help. In [29], we detail how we have extended SimGrid with storage simulation capacities and we list several concrete use cases of storage simulations in clusters, grids, clouds, and data centers for which the proposed extension would be beneficial.

$\Psi^2$is a simulation software of markovian models that is able to provide a perfect sampling of the stationary distribution. In [12], we consider open Jackson networks with losses with mixed finite and infinite queues and analyze the efficiency of sampling from their exact stationary distribution. We show that perfect sampling is possible, although the underlying Markov chain may have an infinite state space. The main idea is to use a Jackson network with infinite buffers (that has a product form stationary distribution) to bound the number of initial conditions to be considered in the coupling from the past scheme. We also provide bounds on the sampling time of this new perfect sampling algorithm for acyclic or hyper-stable networks. These bounds show that the new algorithm is considerably more efficient than existing perfect samplers even in the case where all queues are finite. We illustrate this efficiency through numerical experiments. We also extend our approach to variable service times and non-monotone networks such as queueing networks with negative customers.

6.6. Asymptotic Models

Analyzing a set of n stochastic entities interacting with each others can be particularly difficult but the *mean field approximation* is a very effective technique to characterize the probability distribution of such systems when the number of entities n grows very large. The limit system is generally deterministic and characterized by a differential equation that is more amenable to analysis and optimization. Such approximation however typically requires that the dynamics of the entities depend only on their state (the state space of each object does not scale with n the number of objects) but neither on their identity nor on their spatial location.

- In [28], we analyze a family of list-based cache replacement algorithms. We present explicit expressions for the cache content distribution and miss probability under some assumptions and we develop an algorithm with a time complexity that is polynomial in the cache size and linear in the number of items to compute the exact miss probability. We further introduce a mean field model to approximate the transient behavior of the miss probability and prove that this model becomes exact as the cache size and number of items tends to infinity. We show that the set of ODEs associated to the mean field model has a unique fixed point that can be used to approximate the miss probability in case the exact computation becomes too time consuming. Using this approximation, we provide guidelines on how to select a replacement algorithm within the family considered such that a good trade-off is achieved between the cache reactivity and its steady-state hit probability
- For distributed systems where /locality/ is essential in the dynamics the mean-field approach requires to resort to discretization of space into a finite number of cells to fit in the classical framework.

Such approach not only scales badly but also requires that spatial interactions are weak. One of the tool to tackle this difficult problem comes from statistical physics and is popular in biology: pair approximation. In [26], we successfully apply this approach to the "Power of Two Choice" load balancing paradigm: each incoming task is allocated to the least loaded of two servers picked at random among a collection of n servers. We study the power of two-choice in a setting where the two servers are not picked independently at random but are connected by an edge in an underlying graph. Our problem is motivated by systems in which choices are geometrically constrained (e.g., a bike-sharing system). We study a dynamic setting in which jobs leave the system after being served by a server to which is was allocated. Our focus is when each server has few neighbors (typically 2 to 4) for which an mean-field approximation is not accurate. We build the pair-approximation equations and show that they describe accurately the steady-state of the system. Our results show that, even in a graph of degree 2, choosing between two neighboring improve dramatically the performance compared to a random allocation.

- In [8], we consider a queueing system composed of a dispatcher that routes deterministically jobs to a set of non-observable queues working in parallel. In this setting, the fundamental problem is which policy should the dispatcher implement to minimize the stationary mean waiting time of the incoming jobs. We present a structural property that holds in the classic scaling of the system where the network demand (arrival rate of jobs) grows proportionally with the number of queues. Assuming that each queue of type r is replicated k times, we consider a set of policies that are periodic with period $k \sum_r p_r$ and such that exactly p_r jobs are sent in a period to each queue of type r. When $k \to \infty$, our main result shows that all the policies in this set are equivalent, in the sense that they yield the same mean stationary waiting time, and optimal, in the sense that no other policy having the same aggregate arrival rate to all queues of a given type can do better in minimizing the stationary mean waiting time. Furthermore, the limiting mean waiting time achieved by our policies is a convex function of the arrival rate in each queue, which facilitates the development of a further optimization aimed at solving the fundamental problem above for large systems.

6.7. Trace and Statistical Analysis

Although we often use Markovian approaches to model large scale distributed system, these probabilistic tools can also be used to lay the foundation of statistical analysis of traces of real systems.

- In [36], we explain how we apply statistical statistical modelling and statistical inference of the ANR GEOMEDIA corpus, that is a collection of international RSS news feeds. Central to this project, RSS news feeds are viewed as a representation of the information in geopolitical space. As such they allow us to study media events of global extent and how they affect international relations. Here we propose hidden Markov models (HMM) as an adequate modelling framework to study the evolution of media events in time. This set of models respect the characteristic properties of the data, such as temporal dependencies and correlations between feeds. Its specific structure corresponds well to our conceptualisation of media attention and media events. We specify the general model structure that we use for modelling an ensemble of RSS news feeds. Finally, we apply the proposed models to a case study dedicated to the analysis of the media attention for the Ebola epidemic which spread through West Africa in 2014.
- The use of stochastic formalisms, such as Stochastic Automata Networks (SAN), can be very useful for statistical prediction and behavior analysis. Once well fitted, such formalisms can generate probabilities about a target reality. These probabilities can be seen as a statistical approach of knowledge discovery. However, the building process of models for real world problems is time consuming even for experienced modelers. Furthermore, it is often necessary to be a domain specialist to create a model. In [34], we present a new method to automatically learn simple SAN models directly from a data source. This method is encapsulated in a tool called SAN GEnerator (SANGE). Through examples we show how this new model fitting method is powerful and relatively easy to use, which can grant access to a much broader community to such powerful modeling formalisms.

- In [32], we have presented our recent results on macroscopic analysis of huge traces of parallel/distributed applications. To identify a *macroscopic phenomenon* over large traces, one needs to change the representation scale and to aggregate data both in time, space and application structure through meaningful operators to propose *multi-scale visualizations*. The question is then to know the quantity of information lost by such scaling to be able to correctly interpret them. The principles underlying this approach are based on information theory since the conditional entropy of an aggregation indicates the quantity of information loss when data are aggregated. This approach has been integrated in the Framesoc framework [35].
- In [27], We study the problem of making forecasts about the future availability of bicycles in stations of a bike-sharing system (BSS). This is relevant in order to make recommendations guaranteeing that the probability that a user will be able to make a journey is sufficiently high. To this end, we use probabilistic predictions obtained from a queuing theoretical time-inhomogeneous model of a BSS. The model is parametrized and successfully validated using historical data from the Vélib ' BSS of the City of Paris. We develop a critique of the standard root-mean-square-error (RMSE), commonly adopted in the bike-sharing research as an index of the prediction accuracy, because it does not account for the stochasticity inherent in the real system. Instead we introduce a new metric based on scoring rules. We evaluate the average score of our model against classical predictors used in the literature. We show that these are outperformed by our model for prediction horizons of up to a few hours. We also discuss that, in general, measuring the current number of available bikes is only relevant for prediction horizons of up to few hours.

MOAIS Project-Team

5. New Results

5.1. TABARNAC: Tools for Analyzing Behavior of Applications Running on NUMA Architecture.

In modern parallel architectures, memory accesses represent a common bottleneck. We develop TABARNAC, a tool for analyzing the memory behavior of parallel applications with a focus on NUMA architectures. TABARNAC provides a new visualization of the memory access behavior, focusing on the distribution of accesses by thread and by structure. Such visualization allows the developer to easily understand why performance issues occur. Using TABARNAC, we explain why some applications do not benefit from data and thread mapping. Moreover, we propose several code modifications to improve the memory access behavior of several parallel applications [29].

5.2. Computing the Rank Profile Matrix

We propose the definition of a new matrix invariant, the rank profile matrix, summarizing all information on the row and column rank profiles of all the leading sub-matrices. We also explore the conditions for a Gaussian elimination algorithm to compute all or part of this invariant, through the corresponding PLUQ decomposition [12].

5.3. Parallel Algebraic Linear Algebra Dedicated Interface

We propose a domain specific language based on C/C++ macros, PALADIn (Parallel Algebraic Linear Algebra Dedicated Interface) [15]. This domain specific language allows the user to write C++ code and benefits from sequential and parallel executions on shared memory architectures. With a unique syntax, the user can switch between different parallel runtime systems such as OpenMP, TBB and xKaapi. This interface provides data and task parallelism and has been used for recursion-based parallelization of exact dense linear algebra routines[7].

5.4. Communication models insights meet simulations

It is well-known that taking into account communications while scheduling jobs in large scale parallel computing platforms is a crucial issue. In modern hierarchical platforms, communication times are highly different when occurring inside a cluster or between clusters. Thus, allocating the jobs taking into account locality constraints is a key factor for reaching good performances. However, several theoretical results prove that imposing such constraints reduces the solution space and thus, possibly degrades the performances. In practice, such constraints simplify implementations and most often lead to better results. Our aim in this work is to bridge theoretical and practical intuitions, and check the differences between constrained and unconstrained schedules (namely with respect to locality and node contiguity) through simulations. We have developed a generic tool, using SimGrid as the base simulator, enabling interactions with external batch schedulers to evaluate their scheduling policies. The results confirm that insights gained through theoretical models are ill-suited to current architectures and should be reevaluated [13].

5.5. Adaptive Resource and Job Management for Limited Power Consumption

The last decades have been characterized by an evergrowing requirement in terms of computing and storage resources. This tendency has recently put the pressure on the ability to efficiently manage the power required to operate the huge amount of electrical components associated with state-of-the-art high performance computing systems. The power consumption of a supercomputer needs to be adjusted based on varying power budget or electricity availabilities. As a consequence, Resource and Job Management Systems have to be adequately adapted in order to efficiently schedule jobs with optimized performance while limiting power usage whenever needed. We introduce in this paper a new scheduling strategy that can adapt the executed workload to a limited power budget. The originality of this approach relies upon a combination of speed scaling and node shutdown techniques for power reductions. It is implemented into the widely used resource and job management system SLURM. Finally, it is validated through large scale emulations using real production workload traces of the supercomputer Curie [17].

5.6. Lessons Learned from Building In Situ Coupling Frameworks

Over the past few years, the increasing amounts of data produced by large-scale simulations have motivated a shift from traditional offline data analysis to in situ analysis and visualization. In situ processing began as the coupling of a parallel simulation with an analysis or visualization library, motivated primarily by avoiding the high cost of accessing storage. Going beyond this simple pairwise tight coupling, complex analysis workflows today are graphs with one or more data sources and several interconnected analysis components. In this paper, we review four tools that we have developed to address the challenges of coupling simulations with visualization packages or analysis workflows: Damaris, Decaf, FlowVR and Swift. This self-critical inquiry aims to shed light not only on their potential, but most importantly on the forthcoming software challenges that these and other in situ analysis and visualization frameworks will face in order to move toward exascale [11]. Besides, focusing on asynchronous In Situ Processing with Gromacs, we have exhibited how to take Advantage of GPUs [25].

5.7. Design and analysis of scheduling strategies for multi-CPU and multi-GPU architectures

In [8], we present a comparison of scheduling strategies for heterogeneous multi-CPU and multi-GPU architectures. We designed and evaluated four scheduling strategies on top of XKaapi runtime: work stealing, data-aware work stealing, locality-aware work stealing, and Heterogeneous Earliest-Finish-Time (HEFT). On a heterogeneous architecture with 12 CPUs and 8 GPUs, we analysed our scheduling strategies with four benchmarks: a BLAS-1 AXPY vector operation, a Jacobi 2D iterative computation, and two linear algebra algorithms Cholesky and LU. We conclude that the use of work stealing may be efficient if task annotations are given along with a data locality strategy. Furthermore, our experimental results suggests that HEFT scheduling performs better on applications with very regular computations and low data locality.

ROMA Project-Team

7. New Results

7.1. Scheduling computational workflows on failure-prone platforms

Participants: Guillaume Aupy, Anne Benoit, Henri Casanova [University of Hawaii], Yves Robert.

We study the scheduling of computational workflows on compute resources that experience exponentially distributed failures. When a failure occurs, rollback and recovery is used to resume the execution from the last checkpointed state. The scheduling problem is to minimize the expected execution time by deciding in which order to execute the tasks in the workflow and whether to checkpoint or not checkpoint a task after it completes. We give a polynomial-time algorithm for fork graphs and show that the problem is NP-complete with join graphs. Our main result is a polynomial-time algorithm to compute the execution time of a workflow with specified to-be-checkpointed tasks. Using this algorithm as a basis, we propose efficient heuristics for solving the scheduling problem. We evaluate these heuristics for representative workflow configurations.

This work has been published in the 17th Workshop on Advances in Parallel and Distributed Computational Models [20].

7.2. Efficient checkpoint/verification patterns

Participants: Anne Benoit, Saurabh K. Raina [Jaypee Institute of Information Technology], Yves Robert.

Errors have become a critical problem for high performance computing. Checkpointing protocols are often used for error recovery after fail-stop failures. However, silent errors cannot be ignored, and their peculiarity is that such errors are identified only when the corrupted data is activated. To cope with silent errors, we need a verification mechanism to check whether the application state is correct. Checkpoints should be supplemented with verifications to detect silent errors. When a verification is successful, only the last checkpoint needs to be kept in memory because it is known to be correct.

In this work, we analytically determine the best balance of verifications and checkpoints so as to optimize platform throughput. We introduce a balanced algorithm using a pattern with p checkpoints and q verifications, which regularly interleaves both checkpoints and verifications across same-size computational chunks. We show how to compute the waste of an arbitrary pattern, and we prove that the balanced algorithm is optimal when the platform MTBF (Mean Time Between Failures) is large in front of the other parameters (checkpointing, verification and recovery costs). We conduct several simulations to show the gain achieved by this balanced algorithm for well-chosen values of p and q, compared to the base algorithm that always perform a verification just before taking a checkpoint ($p = q = 1$), and we exhibit gains of up to 19%.

This work has been published in the International Journal of High Performance Computing Applications [8].

7.3. Assessing the impact of partial verifications against silent data corruptions

Participants: Aurélien Cavelan, Saurabh K. Raina [Jaypee Institute of Information Technology], Yves Robert, Hongyang Sun.

Silent errors, or silent data corruptions, constitute a major threat on very large scale platforms. When a silent error strikes, it is not detected immediately but only after some delay, which prevents the use of pure periodic checkpointing approaches devised for fail-stop errors. Instead, checkpointing must be coupled with some verification mechanism to guarantee that corrupted data will never be written into the checkpoint file. Such a guaranteed verification mechanism typically incurs a high cost. In this work, we assess the impact of using partial verification mechanisms in addition to a guaranteed verification. The main objective is to investigate to which extent it is worthwhile to use some light cost but less accurate verifications in the middle of a periodic computing pattern, which ends with a guaranteed verification right before each checkpoint. Introducing partial verifications dramatically complicates the analysis, but we are able to analytically determine the optimal computing pattern (up to the first-order approximation), including the optimal length of the pattern, the optimal number of partial verifications, as well as their optimal positions inside the pattern. Performance evaluations based on a wide range of parameters confirm the benefit of using partial verifications under certain scenarios, when compared to the baseline algorithm that uses only guaranteed verifications.

This work has been published in the proceedings of ICPP'15 [22].

7.4. Which Verification for Soft Error Detection?

Participants: Leonardo Bautista-Gomez [Argonne National Laboratory], Anne Benoit, Aurélien Cavelan, Saurabh K. Raina [Jaypee Institute of Information Technology], Yves Robert, Hongyang Sun.

This work is an extension of the work described in Section 7.4 to cope with imperfect verifications. Many methods are available to detect silent errors in high-performance computing (HPC) applications. Each comes with a given cost and recall (fraction of all errors that are actually detected). The main contribution of this work is to characterize the optimal computational pattern for an application: which detector(s) to use, how many detectors of each type to use, together with the length of the work segment that precedes each of them. We conduct a comprehensive complexity analysis of this optimization problem, showing NP-completeness and designing an FPTAS (Fully Polynomial-Time Approximation Scheme). On the practical side, we provide a greedy algorithm whose performance is shown to be close to the optimal for a realistic set of evaluation scenarios.

This work has been published in the proceedings of HiPC'15 [21].

7.5. Composing resilience techniques: ABFT, periodic and incremental checkpointing

Participants: George Bosilca [University of Tennessee, Knoxville], Aurélien Bouteiller [University of Tennessee, Knoxville], Thomas Hérault [University of Tennessee, Knoxville], Yves Robert, Jack Dongarra [University of Tennessee, Knoxville].

Algorithm Based Fault Tolerant (ABFT) approaches promise unparalleled scalability and performance in failure-prone environments. Thanks to recent advances in the understanding of the involved mechanisms, a growing number of important algorithms (including all widely used factorizations) have been proven ABFT-capable. In the context of larger applications, these algorithms provide a temporal section of the execution, where the data is protected by its own intrinsic properties, and can therefore be algorithmically recomputed without the need of checkpoints. However, while typical scientific applications spend a significant fraction of their execution time in library calls that can be ABFT-protected, they interleave sections that are difficult or even impossible to protect with ABFT. As a consequence, the only practical fault-tolerance approach for these applications is checkpoint/restart. In this work, we propose a model to investigate the efficiency of a composite protocol, that alternates between ABFT and checkpoint/restart for the effective protection of an iterative application composed of ABFT-aware and ABFT-unaware sections. We also consider an incremental checkpointing composite approach in which the algorithmic knowledge is leveraged by a novel optimal dynamic programming to compute checkpoint dates. We validate these models using a simulator. The model and simulator show that the composite approach drastically increases the performance delivered by an execution platform, especially at scale, by providing the means to increase the interval between checkpoints while simultaneously decreasing the volume of each checkpoint.

This work has been published in the International Journal of Networking and Computing [9].

7.6. Voltage Overscaling Algorithms for Energy-Efficient Workflow Computations With Timing Errors

Participants: Aurélien Cavelan, Yves Robert, Hongyang Sun, Frédéric Vivien.

We proposed a software-based approach using dynamic voltage overscaling to reduce the energy consumption of HPC applications. This technique aggressively lowers the supply voltage below nominal voltage, which introduces timing errors, and we used Algorithm-Based Fault-Tolerance (ABFT) to provide fault tolerance for matrix operations. We introduced a formal model, and we designed optimal polynomial-time solutions, to execute a linear chain of tasks. Evaluation results obtained for matrix multiplication demonstrated that our approach indeed leads to significant energy savings, compared to the standard algorithm that always operates at nominal voltage.

This work has been published in the proceedings of the 5th Workshop on Fault Tolerance for HPC at eXtreme Scale [23].

7.7. Approximation algorithms for energy, reliability and makespan optimization problems

Participants: Guillaume Aupy, Anne Benoit.

We consider the problem of scheduling an application on a parallel computational platform. The application is a particular task graph, either a linear chain of tasks, or a set of independent tasks. The platform is made of identical processors, whose speed can be dynamically modified. It is also subject to failures: if a processor is slowed down to decrease the energy consumption, it has a higher chance to fail. Therefore, the scheduling problem requires us to re-execute or replicate tasks (i.e., execute twice the same task, either on the same processor, or on two distinct processors), in order to increase the reliability. It is a tri-criteria problem: the goal is to minimize the energy consumption, while enforcing a bound on the total execution time (the makespan), and a constraint on the reliability of each task.

Our main contribution is to propose approximation algorithms for linear chains of tasks and independent tasks. For linear chains, we design a fully polynomial-time approximation scheme. However, we show that there exists no constant factor approximation algorithm for independent tasks, unless P=NP, and we propose in this case an approximation algorithm with a relaxation on the makespan constraint.

This work has been published in the Parallel Processing Letters [4].

7.8. Co-scheduling algorithms for high-throughput workload execution

Participants: Guillaume Aupy, Manu Shantharam [San Diego Supercomputer Center], Anne Benoit, Yves Robert, Padma Raghavan [Penn State University].

This work investigates co-scheduling algorithms for processing a set of parallel applications. Instead of executing each application one by one, using a maximum degree of parallelism for each of them, we aim at scheduling several applications concurrently. We partition the original application set into a series of packs, which are executed one by one. A pack comprises several applications, each of them with an assigned number of processors, with the constraint that the total number of processors assigned within a pack does not exceed the maximum number of available processors. The objective is to determine a partition into packs, and an assignment of processors to applications, that minimize the sum of the execution times of the packs.

We thoroughly study the complexity of this optimization problem, and propose several heuristics that exhibit very good performance on a variety of workloads, whose application execution times model profiles of parallel scientific codes. We show that co-scheduling leads to to faster workload completion time and to faster response times on average (hence increasing system throughput and saving energy), for significant benefits over traditional scheduling from both the user and system perspectives.

This work has been published in the Journal of Scheduling [6].

7.9. Scheduling the I/O of HPC Applications Under Congestion

Participants: Ana Gainaru [University of Illinois at Urbana Champaign], Guillaume Aupy, Anne Benoit, Franck Cappello, Yves Robert.

A significant percentage of the computing capacity of large-scale platforms is wasted due to interferences incurred by multiple applications that access a shared parallel file system concurrently. One solution to handling I/O bursts in large-scale HPC systems is to absorb them at an intermediate storage layer consisting of burst buffers. However, our analysis of the Argonne's Mira system shows that burst buffers cannot prevent congestion at all times. As a consequence, I/O performance is dramatically degraded, showing in some cases a decrease in I/O throughput of 67%.

In this work, we analyze the effects of interference on application I/O bandwidth, and propose several scheduling techniques to mitigate congestion. We focus on typical HPC applications, which have a periodic pattern consisting of some amount of computation followed by some volume of I/O to be transferred. We show through extensive experiments that our global I/O scheduler is able to reduce the effects of congestion, even on systems where burst buffers are used, and can increase the overall system throughput up to 56%. We also show that it outperforms current Mira I/O schedulers, even for non-periodic applications.

This work has been published in IPDPS'15 [26].

7.10. Scheduling trees of malleable tasks for sparse linear algebra

Participants: Abdou Guermouche [Univ. Bordeaux/Inria Bordeaux Sud-Ouest], Loris Marchal, Bertrand Simon, Oliver Sinnen [Univ. Auckland/New Zealand], Frédéric Vivien.

Scientific workloads are often described by directed acyclic task graphs. This is in particular the case for multifrontal factorization of sparse matrices —the focus of this work— whose task graph is structured as a tree of parallel tasks. Prasanna and Musicus [84], [85] advocated using the concept of *malleable* tasks to model parallel tasks involved in matrix computations. In this powerful model each task is processed on a time-varying number of processors. Following Prasanna and Musicus, we consider malleable tasks whose speedup is p^α, where p is the fractional share of processors on which a task executes, and α ($0 < \alpha \leq 1$) is a task-independent parameter. Firstly, we use actual experiments on multicore platforms to motivate the relevance of this model for our application. Then, we study the optimal time-minimizing allocation proposed by Prasanna and Musicus using optimal control theory. We greatly simplify their proofs by resorting only to pure scheduling arguments. Building on the insight gained thanks to these new proofs, we extend the study to distributed (homogeneous or heterogeneous) multicore platforms. We prove the NP-completeness of the corresponding scheduling problem, and we then propose some approximation algorithms [28].

In a second step, we studied a simplified speed-up function for malleable tasks, corresponding to perfect parallelism for a number of processors below a given threshold. The threshold depends on the task. We proved that scheduling independent chains of malleable tasks under this model is NP-complete. We study the performance of a classical allocation policy which is agnostic of the threshold and a simple greedy heuristic, and proved that both are 2-approximation algorithms, even if in practice, the latter often outperforms the former.

7.11. Parallel scheduling of task trees with limited memory

Participants: Clément Brasseur [ENS Lyon], Guillaume Aupy, Loris Marchal.

Scientific workloads are often described by directed acyclic task graphs. This is in particular the case for multifrontal factorization of sparse matrices —the focus of this work— whose task graph is structured as a tree of parallel tasks. When processing this tree on a multicore machine, we have to find a tradeoff between task parallelism and memory usage. In this context, Agullo et al. [62] proposed an activation scheme which follows a postorder traversal and books the memory needed for the task. This strategy has a low complexity and thus has been implemented in the lightweight runtime system StarPU [65], but may lead to excessive memory booking, which limits the task parallelism. In this work, we proposed a new booking strategy that books exactly what is necessary for a task, given what is already booked by its predecessors in the tree. We have shown by extensive simulations on realistic trees that this leads to better task parallelism and reduces the overall processing time.

7.12. Locality of Map tasks in MapReduce computations

Participants: Olivier Beaumont [Inria Bordeaux Sud-Ouest], Loris Marchal.

In data parallel system such as MapReduce, large data files are distributed among the storage attached to computing nodes, and the computation is afterwards allocated close to the data whenever it is possible. Several parameters may affect the locality of the data, and thus the amount of data that needs to be communicated during the computation: the possible replication of the data when it is distributed on the platform, and the load-balancing mechanism that transmits new data to node which have exhausted their own data. In this work, we have proposed a simple analytical model to estimate the amount of data transfer of various scenarios for the Map phase of MapReduce computations and we have validated this model using simulations.

7.13. Improving multifrontal methods by means of block low-rank representations

Participants: Patrick Amestoy [INPT-IRIT, Université of Toulouse], Cleve Ashcraft [LSTC], Olivier Boiteau [EDF], Alfredo Buttari [CNRS-IRIT, Université of Toulouse], Jean-Yves L'Excellent, Clément Weisbecker [INPT-IRIT, now at LSTC].

Matrices coming from elliptic Partial Differential Equations (PDEs) have been shown to have a low-rank property: well defined off-diagonal blocks of their Schur complements can be approximated by low-rank products. Given a suitable ordering of the matrix which gives the blocks a geometrical meaning, such approximations can be computed using an SVD or a rank-revealing QR factorization. The resulting representation offers a substantial reduction of the memory requirement and gives efficient ways to perform many of the basic dense linear algebra operations.

Several strategies, mostly based on hierarchical formats, have been proposed to exploit this property. We study a simple, non-hierarchical, low-rank format called Block Low-Rank (BLR), and explain how it can be used to reduce the memory footprint and the complexity of sparse direct solvers based on the multifrontal method. We present experimental results on matrices coming from elliptic PDEs and from various other applications. We show that even if BLR based factorizations are asymptotically less efficient than hierarchical approaches, they still deliver considerable gains. The BLR format is compatible with numerical pivoting, and its simplicity and flexibility make it easy to use in the context of a general purpose, algebraic solver. This work has been published in the SIAM Journal on Scientific Computing [2].

7.14. Parallel Computation of a subset of entries of the inverse

Participants: Patrick Amestoy [INPT-IRIT, Université of Toulouse], Iain Duff [RAL and CERFACS], Jean-Yves L'Excellent, François-Henry Rouet.

We consider the computation in parallel of several entries of the inverse of a large sparse matrix. We assume that the matrix has already been factorized by a direct method and that the factors are distributed. Entries are efficiently computed by exploiting sparsity of the right-hand sides and the solution vectors in the triangular solution phase. We demonstrate that in this setting, parallelism and computational efficiency are two contrasting objectives. We develop an efficient approach and show its efficiency on a general purpose parallel multifrontal solver. This work has been published in the SIAM Journal on Scientific Computing [3].

7.15. Efficient 3D frequency-domain seismic modeling with a parallel block low-rank (BLR) direct solver

Participants: Patrick Amestoy [INPT-IRIT, University of Toulouse], Romain Brossier [ISTerre, University of Grenoble-Alpes], Alfredo Buttari [CNRS-IRIT, University of Toulouse], Jean-Yves L'Excellent, Théo Mary [UPS-IRIT, University of Toulouse], Ludovic Métivier [ISTerre-JK-CNRS], Alain Miniussi [Geoazur-CNRS-UNSA], Stéphane Operto [Geoazur-CNRS-UNSA], Alessandra Ribodetti [Geoazur-CNRS-UNSA], Jean Virieux [ISTerre-UJF, University of Grenoble-Alpes], Clément Weisbecker [INPT-IRIT, now at LSTC].

Three-dimensional frequency-domain full waveform inversion (FWI) of fixed-spread data can be efficiently performed in the visco-acoustic approximation when seismic modeling is based on a sparse direct solver. Based on the work in [3] and its extension to a parallel environment, we studied the application of a parallel algebraic Block Low-Rank (BLR) multifrontal solver providing an approximate solution of the time-harmonic wave equation with a reduced operation count, memory demand, and volume of communication relative to the full-rank solver. We analyzed the parallel efficiency and the accuracy of the solver with a realistic FWI case [19]. The application of this parallel BLR solver to a real data case from the North Sea for full waveform inversion of ocean-bottom cable data was also presented in [18], where a multiscale frequency-domain FWI is applied by successive inversions of 11 discrete frequencies in the 3.5Hz-10Hz frequency band. The velocity model built by FWI reveals short-scale features such as channels, scrapes left by drifting icebergs, fractures and deep reflectors below the reservoir level, alhough the presence of gas in the overburden. The quality of the FWI results is controlled by time-domain modeling and source wavelet estimation. This work was done in the context of an on-going collaboration with the Seiscope consortium (https://seiscope2.obs.ujf-grenoble.fr/?lang=en?).

7.16. Approximation algorithms for bipartite matching on multicore architectures

Participants: Fanny Dufossé [DOLPHIN/Inria Lille - Nord Europe], Kamer Kaya [BMI, The Ohio State Univ., USA], Bora Uçar.

We proposed [13] two heuristics for the bipartite matching problem that are amenable to shared-memory parallelization. The first heuristic is very intriguing from a parallelization perspective. It has no significant algorithmic synchronization overhead and no conflict resolution is needed across threads. We showed that this heuristic has an approximation ratio of around 0.632 under some common conditions. The second heuristic was designed to obtain a larger matching by employing the well-known Karp-Sipser heuristic on a judiciously chosen subgraph of the original graph. We showed that the Karp-Sipser heuristic always finds a maximum cardinality matching in the chosen subgraph. Although the Karp-Sipser heuristic is hard to parallelize for general graphs, we exploited the structure of the selected subgraphs to propose a specialized implementation which demonstrates very good scalability. We proved that this second heuristic has an approximation guarantee of around 0.866 under the same conditions as in the first algorithm. We discussed parallel implementations of the proposed heuristics on a multicore architecture. Experimental results, for demonstrating speed-ups and verifying the theoretical results in practice, were also provided.

7.17. Hypergraph partitioning for multiple communication cost metrics

Participants: Mehmet Deveci [BMI, The Ohio State Univ., USA], Kamer Kaya [BMI, The Ohio State Univ., USA], Umit V. Çatalyürek [BMI, The Ohio State Univ., USA], Bora Uçar.

We investigated [12] hypergraph partitioning-based methods for efficient parallelization of communicating tasks. A good partitioning method should divide the load among the processors as evenly as possible and minimize the inter-processor communication overhead. The total communication volume is the most popular communication overhead metric which is reduced by the existing state-of-the-art hypergraph partitioners. However, other metrics such as the total number of messages, the maximum amount of data transferred by a processor, or a combination of them are equally, if not more, important. Existing hypergraph-based

solutions use a two phase approach to minimize such metrics where in each phase, they minimize a different metric, sometimes at the expense of others. We proposed a one-phase approach where all the communication cost metrics can be effectively minimized in a multi-objective setting and reductions can be achieved for all metrics together. For an accurate modeling of the maximum volume and the number of messages sent and received by a processor, we proposed the use of directed hypergraphs. The directions on hyperedges necessitate revisiting the standard partitioning heuristics. We did so and proposed a multi-objective, multi-level hypergraph partitioner. The partitioner takes various prioritized communication metrics into account, and optimizes all of them together in the same phase. Compared to the state-of-the-art methods which only minimize the total communication volume, we showed on a large number of problem instances that the new method produced better partitions in terms of several communication metrics.

7.18. Comments on the hierarchically structured bin packing problem

Participants: Thomas Lambert [Inria Bordeaux Sud-Ouest], Loris Marchal, Bora Uçar.

We studied [16] the hierarchically structured bin packing problem. In this problem, the items to be packed into bins are at the leaves of a tree. The objective of the packing is to minimize the total number of bins into which the descendants of an internal node are packed, summed over all internal nodes. We investigated an existing algorithm and made a correction to the analysis of its approximation ratio. Further results regarding the structure of an optimal solution and a strengthened inapproximability result were given.

7.19. Semi-two-dimensional partitioning for parallel sparse matrix-vector multiplication

Participants: Enver Kayaaslan, Cevdet Aykanat [Bilkent Univ., Turkey], Bora Uçar.

We proposed [31] a novel sparse matrix partitioning scheme, called semi-two-dimensional (s2D), for efficient parallelization of sparse matrix-vector multiply (SpMV) operations on distributed memory systems. In s2D, matrix nonzeros are more flexibly distributed among processors than one dimensional (rowwise or columnwise) partitioning schemes. Yet, there is a constraint which renders s2D less flexible than two-dimensional (nonzero based) partitioning schemes. The constraint is enforced to confine all communication operations in a single phase, as in 1D partition, in a parallel SpMV operation. In a positive view, s2D thus can be seen as being close to 2D partitions in terms of flexibility, and being close to 1D partitions in terms of computation/communication organization. We described two methods that take partitions on the input and output vectors of SpMV and produce s2D partitions while reducing the total communication volume. The first method obtains optimal total communication volume, while the second one heuristically reduces this quantity and takes computational load balance into account. We demonstrated that the proposed partitioning method improves the performance of parallel SpMV operations both in theory and practice with respect to 1D and 2D partitionings.

7.20. Combining backward and forward recovery to cope with silent errors in iterative solvers

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

We proposed combining checkpointing and verification for coping with silent errors in iterative solvers. We used algorithm based fault tolerance for error detection and error correction, allowing a forward recovery (and no rollback nor re-execution) when a single error is detected. We introduced an abstract performance model to compute the performance of all schemes, and we instantiated it using the Conjugate Gradient (CG) algorithm. Finally, we validate our new approach through a set of simulations both in normal and preconditioned CG [48], [25], [47].

7.21. Load-balanced local time stepping for large-scale wave propagation

Participants: Max Rietmann [Univ. Lugano, CH], Daniel Peter [Univ. Lugano, CH], Olaf Schenk [Univ. Lugano, CH], Bora Uçar, Marcus J. Grote [Univ. Basel, CH].

In complex acoustic or elastic media, finite element meshes often require regions of refinement to honor external or internal topography, or small-scale features. These localized smaller elements create a bottleneck for explicit time-stepping schemes due to the Courant-Friedrichs-Lewy stability condition. Recently developed local time stepping (LTS) algorithms reduce the impact of these small elements by locally adapting the time-step size to the size of the element. The recursive, multi-level nature of our LTS scheme introduces an additional challenge, as standard partitioning schemes create a strong load imbalance across processors. We examined [33] the use of multi-constraint graph and hypergraph partitioning tools to achieve effective, load-balanced parallelization. We implemented LTS-Newmark in the seismology code SPECFEM3D and compared performance and scalability between different partitioning tools on CPU and GPU clusters using examples from computational seismology.

7.22. Fast and high quality topology-aware task mapping

Participants: Mehmet Deveci [BMI, The Ohio State Univ., USA], Kamer Kaya [BMI, The Ohio State Univ., USA], Umit V. Çatalyürek [BMI, The Ohio State Univ., USA], Bora Uçar.

Considering the large number of processors and the size of the interconnection networks on exascale-capable supercomputers, mapping concurrently executable and communicating tasks of an application is a complex problem that needs to be dealt with care. For parallel applications, the communication overhead can be a significant bottleneck on scalability. Topology-aware task-mapping methods that map the tasks to the processors (i.e., cores) by exploiting the underlying network information are very effective to avoid, or at worst bend, this limitation. We proposed [24] novel, efficient, and effective task mapping algorithms employing a graph model. The experiments showed that the methods are faster than the existing approaches proposed for the same task, and on 4096 processors, the algorithms improved the communication hops and link contentions by 16% and 32%, respectively, on the average. In addition, they improved the average execution time of a parallel SpMV kernel and a communication-only application by 9% and 14%, respectively.

7.23. Distributed memory tensor computations

Participants: Oguz Kaya, Bora Uçar.

There are two prominent tensor decomposition formulations. CANDECOMP/PARAFAC (CP) formulation approximates a tensor as a sum of rank-one tensors. *Tucker* formulation approximates a tensor with a core tensor multiplied by a matrix along each mode. Both of these formulations have uses in applications. The most common algorithms for both decompositions are based on the alternating least squares method. The algorithms of this type are iterative, where the computational core of an iteration is a special operation operation between an N-mode tensor and N matrices. These key operations are called the matricized tensor times Khatri-Rao product (MTTKRP) in the CP-ALS case, and the n-mode product in the Tucker decomposition case. We have investigated efficient parallelizations of full fledged algorithms for obtaining these two decompositions in distributed memory systems [30], [51] with a special focus on the mentioned key operations. In both studies, hypergraphs are used for computational load balancing and communication cost reduction. We are currently finalizing our last touches on the Tucker decomposition algorithms [51] to submit it to a conference. We are also working towards a unified view of the parallelization of the two algorithms. This work with its whole extend is carried out in the context of the thesis of Oguz Kaya.

7.24. Bridging the gap between performance and bounds of Cholesky factorization on heterogeneous platforms

Participants: Emmanuel Agullo [Inria Bordeaux Sud-Ouest], Olivier Beaumont [Inria Bordeaux Sud-Ouest], Lionel Eyraud-Dubois [Inria Bordeaux Sud-Ouest], Julien Herrmann, Suraj Kumar [Inria Bordeaux Sud-Ouest], Loris Marchal, Samuel Thibault [Inria Bordeaux Sud-Ouest].

In this work, we consider the problem of allocating and scheduling dense linear application on fully heterogeneous platforms made of CPUs and GPUs. More specifically, we focus on the Cholesky factorization since it exhibits the main features of such problems. Indeed, the relative performance of CPU and GPU highly depends on the sub-routine: GPUs are for instance much more efficient to process regular kernels such as matrix-matrix multiplications rather than more irregular kernels such as matrix factorization. In this context, one solution consists in relying on dynamic scheduling and resource allocation mechanisms such as the ones provided by PaRSEC or StarPU. We analyze the performance of dynamic schedulers based on both actual executions and simulations, and we investigate how adding static rules based on an offline analysis of the problem to their decision process can indeed improve their performance, up to reaching some improved theoretical performance bounds which we introduce [17].

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

Participants: Julien Herrmann, George Bosilca [University of Tennessee, Knoxville], Thomas Hérault [University of Tennessee, Knoxville], Loris Marchal, Yves Robert, Jack Dongarra [University of Tennessee, Knoxville].

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

This work has been published in the Parallel Computing journal [15].

7.26. STS-k: A Multi-level Sparse Triangular Solution Scheme for NUMA Multicores

Participants: Humayun Kabir [Penn State University], Joshua Booth [Sandia National Laboratories], Guillaume Aupy, Anne Benoit, Yves Robert, Padma Raghavan [Penn State University].

We consider techniques to improve the performance of parallel sparse triangular solution on non-uniform memory architecture multicores by extending earlier coloring and level set schemes for single-core multiprocessors. We develop STS-k, where k represents a small number of transformations for latency reduction from increased spatial and temporal locality of data accesses. We propose a graph model of data reuse to inform the development of STS-k and to prove that computing an optimal cost schedule is NP-complete. We observe significant speed-ups with STS-3 on 32-core Intel Westmere-EX and 24-core AMD 'MagnyCours' processors. Incremental gains solely from the 3-level transformations in STS-3 for a fixed ordering, correspond to reductions in execution times by factors of 1.4 (Intel) and 1.5 (AMD) for level sets and 2 (Intel) and 2.2 (AMD) for coloring. On average, execution times are reduced by a factor of 6 (Intel) and 4 (AMD) for STS-3 with coloring compared to a reference implementation using level sets.

This work has been published in SC'15 [29].

7.27. Mono-parametric Tiling

Participants: Guillaume Iooss [Inria/ENS-Lyon/UCBL/CNRS], Sanjay Rajopadhye [Colorado State University], Christophe Alias, Yun Zou [Colorado State University].

Tiling is a crucial program transformation with many benefits: it improves locality, exposes parallelism, allows for adjusting the ops-to-bytes balance of codes, and can be applied at multiple levels. Allowing tile sizes to be symbolic parameters at compile time has many benefits, including efficient autotuning, and run-time adaptability to system variations. For polyhedral programs, parametric tiling in its full generality is known to be non-linear, breaking the mathematical closure properties of the polyhedral model. Most compilation tools therefore either avoid it by only performing fixed size tiling, or apply it in only the final, code generation step. Both strategies have limitations.

We first introduce mono-parametric partitioning, a restricted parametric, tiling-like transformation which can be used to express a tiling. We show that, despite being parametric, it is a polyhedral transformation. We first prove that applying mono-parametric partitioning (i) to a polyhedron yields a union of polyhedra, and (ii) to an affine function produces a piecewise-affine function. We then use these properties to show how to partition an entire polyhedral program, including one with reductions. Next, we generalize this transformation to tiles with arbitrary tile shapes that can tesselate the iteration space (e.g., hexagonal, trapezoidal, etc). We show how mono-parametric tiling can be applied at multiple levels, and enables a wide range of polyhedral analyses and transformations to be applied.

This work has been published as an Inria research report [49] and will be submitted to a journal.

7.28. Data-aware Process Networks

Participants: Christophe Alias, Alexandru Plesco [XtremLogic SAS].

High-level circuit synthesis (HLS, high-level synthesis) consists in compiling a C-like high-level program to a circuit. The circuit must be as efficient as possible while using properly the resources (energy, memory, FPGA building blocks, etc). Thought many progresses were achieved on the low aspects of circuit generation (pipeline, place/route), the front-end aspects (parallelism, communications) are still rudimentary compared to the state-of-the-art techniques in the HPC community.

We introduce the Data-aware Process Networks (DPN), a new parallel execution model adapted to the hardware constraints of high-level synthesis, where the data transferts are made explicit. We show that the DPN model is consistant in the meaning where any translation of a sequential program produces an equivalent DPN without deadlocks. Finally, we show how to compile a sequential program to a DPN and how to optimize the input/output and the parallelism.

This work was published as an Inria research report [63] and will be submitted to a journal.

7.29. Termination of C programs

Participants: Laure Gonnord, David Monniaux [CNRS/VERIMAG], Gabriel Radanne [Univ Paris 7/ PPS].

We designed a complete method for synthesizing lexicographic linear ranking functions (and thus proving termination), supported by inductive invariants, in the case where the transition relation of the program includes disjunctions and existentials (large block encoding of control flow).

Previous work would either synthesize a ranking function at every basic block head, not just loop headers, which reduces the scope of programs that may be proved to be terminating, or expand large block transitions including tests into (exponentially many) elementary transitions, prior to computing the ranking function, resulting in a very large global constraint system. In contrast, our algorithm incrementally refines a global linear constraint system according to extremal counterexamples: only constraints that exclude spurious solutions are included.

Experiments with our tool Termite 6.5 show marked performance and scalability improvements compared to other systems.

This work has been published in the proceedings of PLDI'15 [27].

7.30. Analysing C programs with arrays

Participants: Laure Gonnord, David Monniaux [CNRS/VERIMAG].

Automatically verifying safety properties of programs is hard, and it is even harder if the program acts upon arrays or other forms of maps. Many approaches exist for verifying programs operating upon Boolean and integer values (e.g. abstract interpretation, counterexample-guided abstraction refinement using interpolants), but transposing them to array properties has been fraught with difficulties.

In contrast to most preceding approaches, we do not introduce a new abstract domain or a new interpolation procedure for arrays. Instead, we generate an abstraction as a scalar problem and feed it to a preexisting solver. The intuition is that if there is a proof of safety of the program, it is likely that it can be expressed by elementary steps between properties involving only a small (tunable) number N of cells from the array.

Our transformed problem is expressed using Horn clauses over scalar variables, a common format with clear and unambiguous logical semantics, for which there exist several solvers. In contrast, solvers directly operating over Horn clauses with arrays are still very immature.

An important characteristic of our encoding is that it creates a nonlinear Horn problem, with tree unfoldings, contrary to the linear problems obtained by flatly encoding the control-graph structure. Our encoding thus cannot be expressed by encoding into another control-flow graph problem, and truly leverages the Horn clause format.

Experiments with our prototype VAPHOR show that this approach can prove automatically the functional correctness of several classical examples of the literature, including *selection sort*, *bubble sort*, *insertion sort*, as well as examples from previous articles on array analysis.

This work has been published as a research report [53] and is currently under submission.

7.31. Symbolic Range Analysis of Pointers in C programs

Participants: Maroua Maalej, Vitor Paisante [Univ. Mineas Gerais, Brasil], Laure Gonnord, Fernando Pereira [Univ. Mineas Gerais, Brasil], Vitor Paisante [Univ. Mineas Gerais, Brasil].

Alias analysis is one of the most fundamental techniques that compilers use to optimize languages with pointers. However, in spite of all the attention that this topic has received, the current state-of-the-art approaches inside compilers still face challenges regarding precision and speed. In particular, pointer arithmetic, a key feature in C and C++, is yet to be handled satisfactorily. We designed a new alias analysis algorithm to solve this problem. The key insight of our approach is to combine alias analysis with symbolic range analysis. This combination lets us disambiguate fields within arrays and structs, effectively achieving more precision than traditional algorithms. To validate our technique, we have implemented it on top of the LLVM compiler. Tests on a vast suite of benchmarks show that we can disambiguate several kinds of C idioms that current state-of-the-art analyses cannot deal with. In particular, we can disambiguate 1.35x more queries than the alias analysis currently available in LLVM. Furthermore, our analysis is very fast: we can go over one million assembly instructions in 10 seconds.

This work has been published at CGO'16 [32].

An extended version of the related work has also been published as an Inria research report [52] and will be the basis of a journal submission.

STORM Team

7. New Results

7.1. Memory Management and Distributed Computing with StarPU

Task-based programming models manage to abstract away much of the architecture complexity while efficiently meeting the performance challenge, even at large scale. While computation scheduling has been well studied, the dynamic management of memory resource subscription inside such run-times has however been little explored, despite the fact that the lookahead, anticipative capabilities offered by the decoupled task submission/task execution steps may sometime come with a high memory cost, especially in distributed context where buffers for receiving incoming contributions have to be accounted for. We therefore studied the cooperation between a task-based application code and a run-time system engine to control the memory subscription levels throughout the execution. We showed that the task paradigm allows to control the memory footprint of the application by throttling the task submission flow rate, striking a compromise between the performance benefits of anticipative task submission and the resulting memory consumption.

7.2. Simulation Execution with StarPU and SimGrid

The combination of StarPU and SimGrid allows to fast, accurate, and reproducible simulation the execution of task-based HPC applications.

This has proved to be very useful for theoretical research on scheduling heuristics [10]. It notably allowed to modify the simulated platform in order to easily observe and understand which parts of the platform (bandwidth, computation power) cause a bottleneck. It also allowed to remove some parts of the problem, such as the cost of data transfers, to simplify the problem and be able to deeply study scheduling solutions and compare them with optimum solutions in a simple environment before tackling the complete platform.

We have also extended the modelization of computation nodes, to take into account the PCI hierarchy of the system. This allows to get a more accurate simulation of systems which have dedicated channels between GPUs.

Last but not least, we have started to extend the StarPU+Simgrid combination to StarPU+MPI+Simgrid, to simulate the execution of HPC applications on *clusters* of heterogeneous systems. The preliminary results seem to show good accuracy. This will allow to easily study how applications scale, and study for instance how network performance have impact on it.

7.3. Scheduling heuristics for dense linear algebra

In the context of Suraj KUMAR's PhD thesis, we are studying the scheduling of the Cholesky factorization on heterogeneous systems.

We have started to introduce communication costs into the constraint programming. Since this increases resolution time a lot, we had to optimize the expression of the data transfers to simplify the resolution. We modified the StarPU runtime system to be able to inject not only a static schedule of tasks, but also a static schedule of data transfers. This allows to inject the schedule optimized by constraint programming into real executions.

We have also shown how static schedules and dynamic scheduling strategies compare on heterogeneous platforms, and notably in the context of varying task execution time can typically be a problem for static scheduling. Static schedules have actually proven to be robust against variation in execution time. We have also studied injecting static information into dynamic schedulers, which improves the resulting performance with little offline analysis.

7.4. Out-of-core support for task graphs

In the context of the Hi-BOX project, Airbus factorizes very large compressed matrices, which can not fit in the main memory, and most of the data thus have to be temporarily transferred to the disk, and loaded on-demand. We have thus extended the StarPU out-of-core support to the case of compressed matrices, and improved the eviction heuristics, so as to transfer data to the disk in advance.

7.5. Parallel Tasks within StarPU

One of the biggest challenge raised by the design of high performance task-based applications on top of heterogeneous accelerator-based machines lies in choosing the adequate granularity of tasks. Indeed, GPUs generally exhibit better performance when executing kernels featuring numerous threads whereas regular CPU cores typically reach their peak performance with fine grain tasks working on a reduced memory footprint. As a consequence, task-based applications running on such heterogeneous platforms have to adopt an intermediate granularity, chosen as a trade-off between coarse-grain and fine-grain tasks. We have tackle this granularity problem via resource aggregation : our approach consists in reducing the performance gap between accelerators and single cores by scheduling parallel tasks over cluster of CPUs. For this purpose, we have extended the concept of scheduling context, which we introduced in a previous work, so that the main runtime system only sees virtual computing resources on which it can schedule parallel tasks (e.g. BLAS kernels). The execution of tasks inside such clusters can virtually rely on any thread-based runtime system, and does not interfere with the outer scheduler. As a proof of concept we allow the interoperability of StarPU and OpenMP to co-manage task parallelism. We showed that our approach is able to outperform the magma, dplasma and chameleon state-of-the-art dense linear algebra libraries when dealing with matrices of small and medium size.

7.6. Running Compliant OpenMP Applications on top of StarPU with the Klang-Omp Compiler

Several robust runtime systems proposed recently have shown the benefits of task-based parallelism models. However, the common weakness of these supports is to tie applications to specific APIs. The OpenMP specification aims at providing a common parallel programming means for shared-memory platforms. It appears a good candidate to address this issue. We assessed the effectiveness and limits of this approach on the ScalFMM library developed by Inria HiePACS team, implementing fast multipole methods (FMM) algorithms. We showed that OpenMP dependent tasks allow for significant performance improvements over OpenMP loops and independent tasks for this application. We also demonstrated that suitable, targeted language extensions can further improve performances by a important margin in some cases.

7.7. Task-based Seismology Simulation on top of StarPU

Understanding three-dimensional seismic wave propagation in complex media is still one of the main challenges of quantitative seismology. Because of its simplicity and numerical efficiency, the finite-differences method is one of the standard techniques implemented to consider the elastodynamics equation. Additionally, this class of modelling heavily relies on parallel architectures in order to tackle large scale geometries including a detailed description of the physics. Last decade, significant efforts have been devoted towards efficient implementation of the finite-differences methods on emerging architectures. These contributions have demonstrated their efficiency leading to robust industrial applications. The growing representation of heterogeneous architectures combining general purpose multicore platforms and accelerators leads to re-design current parallel application. We thus considered the StarPU task-based runtime system in order to harness the power of heterogeneous CPU+GPU computing nodes. Preliminary results demonstrate significant speedups in comparison with the best implementation suitable for homogeneous cores.

7.8. Interfacing the MPC Parallel Framework with StarPU

CEA has developped a framework named MPC that transforms MPI+OpenMP applications into a lightweight thread-based program which can flexibly and efficiently exploit multicore architectures. StarPU, on its side, is mainly dedicated to schedule coarse grain tasks over accelerators, and is less suited to fine grain task scheduling. We have thus initiated a software interoperability effort between StarPU and MPC. The first step was to implement a new StarPU task scheduling strategy based on a NUMA-aware adaptative task granularity according to the target device (GPU or CPU). We observed significative performance gains for a Cholesky application in comparison to an eager strategy, thanks to the NUMA-aware aspect. However more work is still needed with respect to task decomposition as it implies data partitioning during the execution. We are also working on a variable granularity task programming interface in order to simplify the developer's coding effort. Finally, we develop a mechanism in StarPU to isolate some parts of the computing platform for another runtime. We used nested *scheduling contexts* to redirect some tasks to a scheduling component that StarPU may or may not control. The idea is to associate a scheduling subcontext to a runtime, for instance MPC, that would access to a dedicated set of computing resources for executing parallel kernels.

7.9. A Domain Specific Framework for Executing Stencil Kernels on Accelerated Platforms with SYCL

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

7.10. Combining Code Generation and Template Specialization Techniques in the P-EDGE Generic Polar Error Correction Code Framework

Error Correction Code decoding algorithms for consumer products such as *Internet of Things* (IoT) devices are usually implemented as dedicated hardware circuits. As processors are becoming increasingly powerful and energy efficient, there is now a strong desire to perform this processing in software to reduce production costs and time to market. The recently introduced family of Successive Cancellation decoders for Polar codes has been shown in several research works to efficiently leverage the ubiquitous SIMD units in modern CPUs, while offering strong potentials for a wide range of optimizations. Together with the IMS Laboratory, we designed the P-EDGE framework which combines a specialized skeleton generator and a building blocks library routines to provide a generic, extensible Polar code exploration workbench. It enables ECC code designers to easily experiments with combinations of existing and new optimizations, while delivering performance close to state-of-art decoders.

7.11. Binary Kernel Analysis, Hinting and Transformation for SIMDization

SIMD processor units have become ubiquitous. Using SIMD instructions is the key for performance for many applications. Modern compilers have made immense progress in generating efficient SIMD code. However, they still may fail or SIMDize poorly, due to conservativeness, source complexity or missing

capabilities. When SIMDization fails, programmers are left with little clues about the root causes and actions to be taken. Our proposed guided SIMDization framework builds on the assembly-code quality assessment toolkit MAQAO to analyzes binaries for possible SIMDization hindrances. It proposes improvement strategies and readily quantifies their impact, using *in vivo* evaluations of suggested transformation. Thanks to our framework, the programmer gets clear directions and quantified expectations on how to improve his/her code SIMDizability. We show results of our technique on TSVC benchmark.

7.12. Dynamic Granularity Adaptation of OpenCL Kernels on Heterogeneous Multi-device Systems

On-going work as part of the PhD of P. Huchant aims to transparently execute an OpenCL kernel, and further a complete task graph, on an heterogeneous multi-device system. We propose methods to split an OpenCL kernel at compile time and adapt its granularity dynamically to ensure load balance. If the kernel is executed multiple times, we propose to determine its granularity by using a linear program whose constraints are built from performance measurements collected during the first invocations of the kernel with predefined granularities. Splitting the execution of one kernel into different executions does not require additional information from the user, therefore increasing the level of portability of OpenCL codes. First experiments show the interest of our approach.

TADAAM Team

6. New Results

6.1. TreeMatch Development

This year we have modified the TreeMatch API in order to enable better integration inside application with higher-level abstractions more precise semantic. We also introduced the "over subscribing" features that allow to map more than one process on a given processing unit. We also added new metrics to measure the performance of the proposed placement. We now have three metrics: The sum of the communication cost, the maximum of the communication cost and the hope-byte.

6.2. Affinity Abstraction

This year we worked on the affinity abstraction. Often, the affinity between two processes or threads is measured by the a matrix where a high entry represent a high affinity. Such example of matrices are the number of messages and the number of bytes exchanged between processing units. However, such matrix hide many characteristics of the application such as computation/communication overlap, network contention, etc. First, we have developed a new OpenMPI PML module to gather communication matrix of a running application. Then, we have conducted an extensive study of the minighost application to understand how such communication matrix actually measure the affinity between processes. On this application it appears that the size metric better matches the performances and that the performance of process placement is highly correlated to the proportion of communication in the application.

6.3. Locality for Application Using Locks on Clusters of Multicore Platforms

The aim of this post-doc work is to study the locality for applications based on read-write locks on clusters of multi-core platforms. We focused on the implementation of the video tracking application [25] using the Ordered Read Write Locks (ORWL) [20] model of programming on multi-cores architecture. For several uses, such as, human-computer interaction, security or traffic control, the tracking application aim to locate multiple moving objects over time using a camera. Its processing can be a time consuming process due to the amount of data that is contained in high definition video which leads to decrease the throughput. To overcome this problem it is possible to speed up the processing by exploiting task parallelism of ORWL model. Indeed, the model proposes abstractions of the decomposition in parallel parts (tasks), the synchronization of and the communication between threads. However, we noted some problems which decrease the parallelism scaling thus we introduced different optimizations: stream multiplexing, multiple buffering, etc. We are now working on parallelizing long-running tasks.

6.4. Topology Aware Malleability of MPI programs

Current parallel environments aggregate large numbers of computational resources with a high rate of change in their availability and load conditions. In order to obtain the best performance in this type of infrastructures, parallel applications must be able to adapt to these changing conditions.

In collaboration with Universidade da Coruña, Spain, we have worked on automatically and transparently adapt MPI applications to available resources is proposed. The solution relies on application-level migration approach, incorporating a new scheduling algorithm, based on TreeMatch and Hwloc, to obtain well balanced nodes while preserving performance as much as possible.

The experimental evaluation shows successful and efficient operation, with an overhead of less than 1 second for the proposed scheduling algorithm, and of only a few seconds for the complete reconfiguration, which will be negligible in large applications with a realistic reconfiguration frequency.

6.5. Topology Aware Load Balancing

Charm++ is a message-passing based programming environment that uses an object-oriented approach. However, where MPI considers processes in its model, Charm++ uses finer-grain migratable objects called chares. Brought together with an adaptive runtime system, Charm++ allows to perform dynamic load balancing considering the CPU load of each chare. Our work on data locality and process placement lead us to add the benefits of our TreeMatch algorithm in a load balancing solution. Thus we developed few months ago a topology-aware load-balancer in Charm++ using TreeMatch to reduce the communication costs. During the last months, we significantly improved this load-balancer and its scalability. Particularly, our load balancing algorithm is now hierarchical and distributed. To validate this approach, we have begun to carry out experiments with a cosmological application on the Blue Waters supercomputer. The results will be published soon.

6.6. Topology Aware Resource Management

SLURM [24] is a Resource and Job Management System, a middleware in charge of delivering computing power to applications in HPC systems. Our goal is to take in account in SLURM placement process hardware topology as well as application communication pattern. We have a new selection option for the cons_res plugin in SLURM. In this case the usually BestFit algorithm used to choose nodes is replaced by TreeMatch to find the best placement among the free nodes list in light of a given application communication matrix.

We updated this plugin based on SLURM 2.6.5 for last version SLURM 15.08. To decrease the overhead due to our algorithm we also implemented an alternative to use a subtree of the global topology. We ran experiments to compare these different solutions using our plugin with or without subtree and the current algorithm topology-aware in SLURM.

6.7. Topology Aware Performance Monitoring

While system's scale is growing exponentially, memory hierarchy is getting larger, at various levels. Hence optimizing applications to reach an optimal usage of a machine may involve a large spectrum of performance metrics interacting at different level of the system's hierarchy. Memory bound applications showing irregular patterns lead to locality issues. Addressing those issues and getting a good schedule on complex systems is a NP hard problem and can therefore only be solved with heuristics. Although powerful algorithms using the most intuitive heuristics such as communications path reduction and/or cache contention reduction may show good results on some cases, there are still room for improvements in this direction so much the configuration of applications, systems, software stack vary and impact the execution time.

In order to step in this direction we developped a highly extendible tool to gather asynchronously performance data from different sources. This information is then aggregated into different topology objects (cache, node, processing unit, ...) in order to give a synthetic and topology aware information to drive optimization.

In brief the tool works this way: The user provide a description file with arithmetic expression of performance counters(defined into performance data plugins), and topology objects where to map the expression. A pair (expression,object) defines a monitor which will sample performance data and stored them into an history. Then others monitors can be defined as a combination of the previous. For instance we can attach a process and record on each core its L3 cache miss counter, and then add each of those monitor into an upper monitor located on the L3 cache. Several aggregation functions are already available but we aim to provide several statistical function to extend the possibility of data interpretation. Such functions allow to aggregate results in a meaningful way. Then we add a locality insight using lstopo tool from hwloc to draw the results on a topology. This has been published in [12]

6.8. Memory Hierarchy Aware Roofline Model

The increasing complexity of computer architectures, makes challenging to fully exploit computer systems' capabilities. The cost of tuning applications on such machines can raise quickly. Therefore, linking the information about a machine performance bounds and applications performance results respectively to those bounds can help finding the bottlenecks and motivate code optimization.

In 2009 the Roofline model [23] throws those bases by ploting on a 2 dimensional diagram, application performance (GFlop/s) and arithmetic intensity (Flop/Byte) with respect to the main memory bandwidth (GByte/s) and peak floating point performance (GFlop/s). In 2014 the model extended by Alexandar Illic, take into account the data movement inside the cache hierarchy to provide a finer analyse by showing application's performance results with respect to the differents cache bandwidths.

With the cooperation of the Cache Aware Roofline Model authors, we have worked on extending this model to the whole memory hierarchy at NUMA scale in order to drive optimisations on next generation processors embeding different memory technologies and different memory configurations like Intel's KNL does.

While we are designing a tool based on hwloc and micro-kernels to empirically extract and validate machines bottlenecks, we also want to show with real NUMA applications that the model may be extended to such hierarchy levels, still providing insightful representation.

6.9. Topology Management and Standardization

We continued to work on the diffusion of our software and ideas in existing programming interfaces and standards tailored for HPC and parallel computing. In particular, we did integrate our TreeMatch algorithm in the Open MPI implementation of the Message Passing Interface, so as to provide enhanced Virtual Topology routines in MPI allowing the user to effectively create parallel applications taking into account both their behaviour and the characteristics of the underlying hardware. Our code is available in the master repository and should be available in an Open MPI distribution at some point in the next year (2016).

We also drafted and submitted a proposal to modify the MPI interface so that information regarding the underlying physical topology could be made available at the MPI application level. We plan to push our ideas during the next year so that our proposal can eventually make its way into the MPI standard.

6.10. Modeling Next-Generation Memory Architectures

We initiated a research topic on modeling next generation memory architectures that will mix different kinds of memories. Indeed the arrival of high-bandwidth and non-volatile memories cause computing cores to have different local memory banks with different characteristics.

The hwloc software 5.1 is being extended in collaboration with CPU vendors such as Intel and AMD to better represent these new memory technologies. We are working with Bull in the context of Nicolas Denoyelle's PhD on developing abstractions for deciding where to allocate the application buffers.

6.11. Modeling Affinity of Multithreaded Applications

With the increasing complexity and scale of multi-core processors, optimizing thread placement becomes more and more challenging. Our goal is to better understand which characteristics of a multi-threaded application can have an impact on a placement decision for a given architecture. To this end, we analyze the performance of a set of applications under different placement strategies and we try to relate the obtained results to characteristics of the applications such as the data footprint of each thread, the amount of data shared between threads, or the reuse distance.

To collect information about the characteristics of multi-threaded applications, we developed a set of tools based on the PIN dynamic binary instrumentation tools. PIN allows us to get information about all instructions executed and memory location accessed by each thread of an application during its execution, and this without modifying the source code of the application.

We used our PIN-based tools to study a representative set of applications taken from two well-known benchmark suites, namely the Mantevo benchmark suites (HPC applications based on OpenMP) and the Parsec benchmark suites (general-purpose applications based on pthreads). Analyzing the results of all our tests is an ongoing work.

6.12. Thread placement and threads policy on a multicore machine with NUMA effects.

Threads placement on multicore machine with NUMA effects is inevitable to have better performances. Threads must bind on cores to avoid thread migration and to have better cache locality. MPI non-blocking collectives can generate progress threads to complete communications. These additional threads can disturb computational threads. That is why we have implemented several thread placement algorithms into the MPC framework [22]. These algorithms allow to dedicate resources only for progress threads. Thus computational threads are not disturb. We test them with our own benchmarks which test all the MPI non-blocking collectives to compare the performances with different thread placement. We observe an improvement when resources are dedicated to progress threads and take NUMA effects into account.

We want to include a mechanism into MPC to specify thread kinds (MPI, OpenMP,...). These mechanism will allow the MPC scheduler to take threads specificity into account to improve the scheduling policy. Our goal is to increase runtime performances considering each type specific needs. We have begun to implement this mechanism.

Several MPC framework bugs have been corrected, thus we contribute to its stability.

6.13. Multithreaded Communications

To program clusters of multicores, hybrid models mixing MPI+threads, and in particular MPI+OpenMP are gaining popularity. This imposes new requirements on communication libraries, such as the need for MPI_THREAD_MULTIPLE level of multi-threading support. Moreover, the high number of cores brings new opportunities to parallelize communication libraries, so as to have proper background progression of communication and communication/computation overlap.

We have proposed PIOMan [11], a generic framework to be used by MPI implementations, that brings seamless asynchronous progression of communication by opportunistically using available cores. It uses system threads and thus is composable with any runtime system used for multithreading. Through various benchmarks, we demonstrated that our pioman-based MPI implementation exhibits very good properties regarding overlap, progression, and multithreading, and outperforms state-of-art MPI implementations.

6.14. RDMA-based Communications

High-performance network hardware is nowadays dominated by RDMA-oriented technologies. The software stack is moving too towards Remote Memory Access. However, most communication libraries stil use send/receive paradigm as a common denominator. We have proposed to study a software stack for networks the is based on remote memory access from the hardware up to the enduser API, where RDMA is first class citizen and not a compatibility layer. It is expected to obtain better performance, better scalability with regard to number of communication flows or threads, and better asynchronous progression, while optimization strategies on the packet flows such as aggregation as proposed in NewMadeleine are still possible. Work has begun as a Masters thesis [18] and continues as Romain Prou Ph.D. thesis.

6.15. Network Modeling

Netloc is a tool for hwloc [1] to find the topology of a supercomputer. For that, it discovers all the networks by exploring them by using tools specifying to the network type. The exploration gives all the machines and all the switches, with all the links between them. We improved netloc by adding the visualization of the topologies discovered. The visualization is dynamic and the user can interact with it, to get some information about the machines, the switches or the link such as the physical address, the hostname or the speed of the link. In order to be able to do optimizations that can be helping process placement, we started to class the different topologies. For now, we only handle Clos networks [21] and we are able to transform them into fat trees. The categorization in classes permits to have a clean graph and then interact with graph partitioners.

To have a complete tool, we need to handle all major classes of topologies such as meshes, torus or hypercubes. When the graph partitioning will be integrated with tools such as SCOTCH, we will be able to find a good mapping for the processes of a job. It could also helps the resource scheduling to optimize the resource sharing between jobs. The visualization can be improved by showing the architecture information retrieved by hwloc for each machine. We can complete the visualization by giving more information especially when the original graph was transformed to simplify it, as we did to Clos networks to obtain fat trees.

6.16. Scalable mapping onto (disconnected) parts of regular target architectures

Since its inception, SCOTCH allows one to map graphs onto so-called "algorithmically-defined" target architectures. They are regular architectures such as hypercube, multi-dimensional grids and tori, butterfly networks, etc., whose characteristics are defined by subroutines which are part of the SCOTCH library. However, on today's large-scale computer systems, software jobs do not usually run on all of the machine, but on a set of nodes assigned by the batch scheduler. Consequently, one should be able to map a process graph onto (possibly disconnected) parts of an algorithmically-defined target architecture, which was not an available feature. Only "decomposition-defined" architectures (another way to represent target architectures in SCOTCH) supported this feature, but are not scalable above a few hundred processing elements.

In order to allow SCOTCH to provide mappings onto parts of an algorithmically-defined target architecture, a new meta target architecture, called "`sub`", has been created. The `sub` architecture allows one to restrict a regular algorithmically-defined target architecture to a subset of its vertices. Instead of using a top-down approach to build a description of the target architecture, through a recursive bipartitioning algorithm, our new algorithm uses a bottom-up approach, based on recursive matching and coarsening of neighboring vertices, much like for graph coarsening. The clustering tree is pruned of branches that lead to parts of the machine that are not allowed mapping targets. Distance between subdomains is computed using the distance function of the underlying algorithmically-defined target architecture. Preliminary results have been presented at a SIAM CS&E conference workshop [14], and a beta-version of the upcoming release 6.0.5 of SCOTCH has been shipped to early testers at Lawrence Livermore National Laboratory.

6.17. Multi-Level Parallelism in a CFD code

Code_Saturne [19] is an industrial and open source Computational Fluid Dynamics software. Developed at EDF R&D, it solves the Navier-Stokes equations for 2D, 2D-axisymmetric and 3D flows, steady or unsteady, laminar or turbulent, incompressible or weakly dilatable, isothermal or not, with scalars transport if required.

Our goal is to evaluate different ways of improving and preparing this application for the future HPC architectures. We strengthened our application knowledge by using various instrumentation tools and provided a small topology instrumentation library. As instrumentation of a full code can be a tedious thing, we provided a mini application on which to perform our future experiments. We have run experiments to determine the potential gain of topology awareness on our code by using the graph mapping solutions of PT-SCOTCH. We have also run experiments on ghost cells numbering to see the impact of their locations on cache misses.

ASCOLA Project-Team

6. New Results

6.1. Highlights of the year

Nicolas Tabareau has been awarded a starting grant from the European Research Council (ERC), the most prestigious type of research projects of the European Union for young researchers. From 2015–2020 he will pursue research on "CoqHoTT: Coq for Homotopy Type Theory."

In the domain of resource management notably for Cloud infrastructures, the team has produced several very visible results. These include contributions to popular and new simulation tools and platforms [17], [27], [28] as well as new techniques for the energy-efficient execution of Cloud applications [15].

On the topics of software composition and programming languages, the team has, among others, two remarkable results: a new notion of effect capabilities and corresponding monadic analysis techniques [14] as well as the first comprehensive survey of domain-specific aspect languages [13].

6.2. Programming Languages

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

6.2.1. Formal Methods, logics and type theory

This year we have proposed "Gradual Certified Programming" as a bridge between type-based expressive proofs and programming languages, have extended previous type theories by new homotopy-based means, and have introduced "effect capabilities" to control monad-based effects in Haskell.

6.2.1.1. Gradual Certified Programming in Coq

Expressive static typing disciplines are a powerful way to achieve high-quality software. However, the adoption cost of such techniques should not be under-estimated. Just like gradual typing allows for a smooth transition from dynamically-typed to statically-typed programs, it seems desirable to support a gradual path to certified programming. We have explored gradual certified programming in Coq [33], providing the possibility to postpone the proofs of selected properties, and to check "at runtime" whether the properties actually hold. Casts can be integrated with the implicit coercion mechanism of Coq to support implicit cast insertion à la gradual typing. Additionally, when extracting Coq functions to mainstream languages, our encoding of casts supports lifting assumed properties into runtime checks. Much to our surprise, it is not necessary to extend Coq in any way to support gradual certified programming. A simple mix of type classes and axioms makes it possible to bring gradual certified programming to Coq in a straightforward manner.

6.2.1.2. Homotopy Hypothesis in Type Theory

In classical homotopy theory, the homotopy hypothesis asserts that the fundamental omega-groupoid construction induces an equivalence between topological spaces and weak omega-groupoids. In the light of Voevodsky's univalent foundations program, which puts forward an interpretation of types as topological spaces, we have considered the question of transposing the homotopy hypothesis to type theory [16]. Indeed such a transposition could stand as a new approach to specifying higher inductive types. Since the formalization of general weak omega-groupoids in type theory is a difficult task, we have only taken a first step towards this goal, which consists in exploring a shortcut through strict omega-categories. The first outcome is a satisfactory type-theoretic notion of strict omega-category, which has hsets of cells in all dimensions. For this notion, defining the 'fundamental strict omega-category' of a type seems out of reach. The second outcome is an 'incoherently strict' notion of type-theoretic omega-category, which admits arbitrary types of cells in all dimensions. These are the 'wild' omega-categories of the title. They allow the definition of a 'fundamental wild omega-category' map, which leads to our (partial) homotopy hypothesis for type theory (stating an adjunction, not an equivalence). All of our results have been formalized in the Coq proof assistant. Our formalization makes systematic use of the machinery of coinductive types.

6.2.1.3. Effect Capabilities For Haskell

Computational effects complicate the tasks of reasoning about and maintaining software, due to the many kinds of interferences that can occur. While different proposals have been formulated to alleviate the fragility and burden of dealing with specific effects, such as state or exceptions, there is no prevalent robust mechanism that addresses the general interference issue. Building upon the idea of capability-based security, we have proposed effect capabilities [14] as an effective and flexible manner to control monadic effects and their interferences. Capabilities can be selectively shared between modules to establish secure effect-centric coordination. We have further refined capabilities with type-based permission lattices to allow fine-grained decomposition of authority. An implementation of effect capabilities in Haskell has been done, using type classes to establish a way to statically share capabilities between modules, as well as to check proper access permissions to effects at compile time.

6.2.1.4. Correct Refactoring Tools

Most integrated development environments provide refactoring tools. However, these tools are often unreliable. As a consequence, developers have to test their code after applying an automatic refactoring.

Refactoring tools for industrial languages are difficult to test and verify. We have developed a refactoring operation for C programs (renaming of global variables) for which we have proved that it preserves the set of possible behaviors of the transformed programs [39]. That proof of correctness relies on the operational semantics provided by CompCert C in Coq. We have also proved some properties of the transformation which are used to establish properties of a composed refactoring operations.

6.2.2. Language Mechanisms

This year we have contributed new results on domain-specific aspect languages, concurrent event-based programming, model transformations as well as the relationship between functional and constraint programming.

Furthermore, we have proposed language support for the definition and enforcement of security properties, in particular related to the accountability of service-based systems, see Sec. 6.3 .

6.2.2.1. Domain-Specific Aspect Languages

Domain-Specific Aspect Languages (DSALs) are Domain-Specific Languages (DSLs) designed to express crosscutting concerns. Compared to DSLs, their aspectual nature greatly amplifies the language design space. In the context of the Associate Team RAPIDS/REAL, we have structured this space in order to shed light on and compare the different domain-specific approaches to deal with crosscutting concerns [13]. We have reported on a corpus of 36 DSALs covering the space, discussed a set of design considerations and provided a taxonomy of DSAL implementation approaches. This work serves as a frame of reference to DSAL and DSL researchers, enabling further advances in the field, and to developers as a guide for DSAL implementations.

6.2.2.2. Concurrent Event-Based Programming

The advanced concurrency abstractions provided by the Join calculus overcome the drawbacks of low-level concurrent programming techniques. However, with current approaches, the coordination logic involved in complex coordination schemas is still fragmented. In [11], Jurgen Van Ham presents JEScala, a language that captures coordination schemas in a more expressive and modular way by leveraging a seamless integration of an advanced event system with join abstractions. The implementation of joins-based state machines is discussed with alternative faster implementations made possible through a domain specific language. Event monitors are introduced as a way of synchronizing event handling and building concurrent event-based applications from sequential event-based parts.

6.2.2.3. Model Lazy Transformation

The Object Constraint Language (OCL) is a central component in modeling and transformation languages such as the Unified Modeling Language (UML), the Meta Object Facility (MOF), and Query View Transformation (QVT). OCL is standardized as a strict functional language. We have proposed a lazy evaluation strategy for OCL [36]. This lazy evaluation semantics is beneficial in some model-driven engineering scenarios for speeding up the evaluation times for very large models, simplifying expressions on models by using infinite

data structures (e.g., infinite models) and increasing the reusability of OCL libraries. We have implemented the approach on the ATL virtual machine EMFTVM. This is a joint work with the Inria team Atlanmod.

6.2.2.4. Composition Mechanisms for Constraints Generalization

Structural time series (pattern for sequences of values) can be described with numerous automata-based constraints. In [12], we describe a large family of constraints for structural time series by means of function composition. We formalize the patterns using finite transducers. Based on that description, we automatically synthesize automata with accumulators, as well as constraint checkers. The description scheme not only unifies the structure of the existing 30 time-series constraints, but also leads to over 600 new constraints, with more than 100,000 lines of synthesized code. This is a joint work with the Inria team Tasc.

6.3. Software Composition

Participants: Walid Benghrabit, Ronan-Alexandre Cherrueau, Rémi Douence, Hervé Grall, Jean-Claude Royer, Mario Südholt.

6.3.1. Constructive Security

Nowadays we are witnessing the wide-spread use of cloud services. As a result, more and more end-users (individuals and businesses) are using these services for achieving their electronic transactions (shopping, administrative procedures, B2B transactions, etc.). In such scenarios, personal data is generally flowing between several entities and end-users need (i) to be aware of the management, processing, storage and retention of personal data, and (ii) to have necessary means to hold service providers accountable for the usage of their data. Usual preventive security mechanisms are not adequate in a world where personal data can be exchanged on-line between different parties and/or stored at multiple jurisdictions. Accountability becomes a necessary principle for the trustworthiness of open computer systems. It regards the responsibility and liability for the data handling performed by a computer system on behalf of an organization. In case of misconduct (e.g. security breaches, personal data leak, etc.), accountability should imply remediation and redress actions, as in the real life.

In 2015, we have contributed two main results: first, techniques for the logic-based definition, analysis and verification of accountability properties; second, a new framework for the compositional definition of privacy-properties and their type-based enforcement.

6.3.1.1. Logic-based accountability properties

We have proposed a framework for the representation of accountability policies [37]. This framework comes with two novel accountability policy languages; the Abstract Accountability Language (AAL), which is devoted to the representation of preferences/obligations in an human readable fashion, and a concrete one for the mapping to concrete enforceable policies. Our efforts have focused on a formal foundation for the AAL language and some applications.

We have also introduced an approach to assist the design of accountable applications [21]. In particular, we consider an application's abstract component design and we introduce a logical approach allowing various static verification. This approach offers effective means to early check the design and the behavior of an application and its offered/required services. We motivate our work with a realistic use case coming from the A4Cloud project and validate our proposal with experiments using the TSPASS theorem prover. This prover is competitive with other model-checkers and sat solvers and we gain a more abstract approach than with our previous experiment with a model-checker. It makes also easier the link with end users, for instance privacy officers.

To give a formal foundation of the AAL language we define a translation into first-order temporal logic [20]. We introduce a formula to interpret accountability and a natural criterion to achieve the accountability compliance of two clauses. We continue to apply it to an health care system taking into account data privacy features, data transfers and location processing. We demonstrate few heuristics to speed up the resolution time and to assist in conflict detection. Tool support (AccLab) has been provided to support editing, checking and proving AAL clauses.

6.3.1.2. Composition of Privacy-Enforcement Techniques

Today's large-scale computations, e.g., in the Cloud, are subject to a multitude of risks concerning the divulging and ownership of private data. Privacy risks are mainly addressed using a large variety of encryption-based techniques. We have proposed a compositional approach for the declarative and correct composition of privacy-preserving applications in the Cloud [22], [38]. Our approach provides language support for the compositional definition of encryption-based and fragmentation-based privacy-preserving algorithms. This language comes equipped with a set of laws that allows us to verify privacy properties. We have provided implementation support in Scala that ensures certain privacy properties by construction using advanced features of Scala's type system.

6.3.2. Modular systems

6.3.2.1. Modularity for Javascript Interpreters.

With an initial motivation based on the security of web applications written in JavaScript, we have provided new techniques for the instrumentation of an interpreter for a dynamic analysis as a crosscutting concern [31]. We have defined the instrumentation problem — an extension to the expression problem with a focus on modifying interpreters. We have then shown how we can instrument an interpreter for a simple language using only the bare language features provided by JavaScript.

6.4. Cloud applications and infrastructures

Participants: Frederico Alvares, Simon Dupont, Md Sabbir Hasan, Adrien Lebre, Thomas Ledoux, Jonathan Lejeune, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Mario Südholt.

In 2015, we have provided solutions for Cloud-based and distributed programming, virtual environments and data centers.

6.4.1. Cloud and distributed programming

6.4.1.1. Cloud elasticity

Cloud Computing has provided important new means for the capacity management of resources. The elasticity and the economy of scale are the intrinsic elements that differentiate it from traditional computing paradigm.

A good capacity planning method is a necessary factor but not sufficient to fully exploit Cloud elasticity. In [26], we propose innovative policies for resource management to achieve the optimal balance between capacity and quality of Cloud services The main idea is to finely control the scalability and the termination of virtual machines with respect to several criteria such as the lifecycle of the instances (*e.g.* initialization time) or their cost. The approach was evaluated on an Amazon EC2 cluster. Experimental results illustrate the soundness of the proposed approach and the impact of scalability/termination resource policies: a cost saving of as much as 30% can be achieved with a minimal number of violations, as small as 1%.

In order to improve Cloud elasticity, we advocate that the software layer can take part in the elasticity process as the overhead of software reconfiguration can be usually considered negligible compared to infrastructural costs. Thanks to this extra level of elasticity, we are able to define cloud reconfigurations that enact elasticity in both the software and infrastructure layers. In [23], we present an autonomic approach to manage cloud elasticity in a cross-layered manner. First, we enhance cloud elasticity with the software elasticity model. Then, we describe how our autonomic cloud elasticity model relies on the dynamic selection of elasticity tactics. We present an experimental analysis of a subset of those elasticity tactics under different scenarios in order to provide insights on strategies that could drive the autonomic selection of the proper tactics to be applied.

6.4.1.2. Service-level agreement for the Cloud

Quality-of-service and SLA guarantees are among the major challenges of cloud-based services. In [18], we first present a new cloud model called SLAaaS — SLA aware Service. SLAaaS considers QoS levels and SLA as first class citizens of cloud-based services. This model is orthogonal to other SaaS, PaaS, and IaaS cloud models, and may apply to any of them. More specifically, we make three contributions: (i) we provide a domain-specific language that allows to define SLA constraints in cloud services; (ii) we present a general control-theoretic approach for managing cloud service SLA; (iii) we apply our approach to MapReduce, locking, and e-commerce services.

6.4.1.3. Distributed multi-resource allocation

Generalized distributed mutual exclusion algorithms allow processes to concurrently access a set of shared resources. However, they must ensure an exclusive access to each resource. In order to avoid deadlocks, many of them are based on the strong assumption of a prior knowledge about conflicts between processes' requests. Some other approaches, which do not require such a knowledge, exploit broadcast mechanisms or a global lock, degrading message complexity and synchronization cost. We propose in [29] [41] a new solution for shared resources allocation which reduces the communication between non-conflicting processes without a prior knowledge of processes conflicts. Performance evaluation results show that our solution improves resource use rate by a factor up to 20 compared to a global lock based algorithm.

6.4.2. Virtualization and data centers

In 2015, we have produced results and tools for the simulation of large-scale distributed algorithms, notably VM scheduling algorithms, have contributed new abstractions for storage systems and have devised new means for the introspection of Cloud infrastructures.

6.4.2.1. SimGrid / VMPlaceS

We have developed VMPlaceS [28], a framework providing programming support for the definition of VM placement algorithms, execution support for their simulation at large scales, as well as new means for their trace-based analysis. VMPlaceS enables, in particular, the investigation of placement algorithms in the context of numerous and diverse real-world scenarios. To illustrate relevance of such a tool, we evaluated three different classes of virtualization environments: centralized, hierarchical and fully distributed placement algorithms. We showed that VMPlaceS facilitates the implementation and evaluation of variants of placement algorithms. The corresponding experiments have provided the first systematic results comparing these algorithms in environments including up to one thousand of nodes and ten thousands of VMs in most cases.

While such a number is already valuable and although we finalized the virtualization abstractions in SimGrid [17], we are in touch with the core developers in order to improve the code of VMPlaceS with the ultimate objective of addressing infrastructures up to 100K physical machines and 1 Millions virtual machines over a period of one day.

The current version of VMPlaceS is available on a public git repository :http://beyondtheclouds.github.io/VMPlaceS/.

6.4.2.2. Storage abstractions within the SimGrid framework

With the recent data deluge, storage is becoming the most important resource to master in modern computing infrastructures. Dimensioning and assessing the performance of storage systems are challenges for which simulation constitutes a sound approach. Unfortunately, only a few existing simulators of large scale distributed computing systems go beyond providing merely a notion of storage capacity. In 2015, we contributed to the SimGrid efforts toward the simulation of such systems [27]. Concretely, we characterized the performance behavior of several types of disks to derive a first model of storage resource. This model has been integrated within the SimGrid framework available under the LGPL license (http://simgrid.gforge.inria.fr).

6.4.2.3. Cloud Introspection

Cloud Computing has become a new technical and economic model for many IT companies. By virtualizing services, it allows for a more flexible management of datacenters capacities. However, its elasticity and its flexibility led to the explosion of virtual environments to manage. It's common for a system administrator to manage several hundreds or thousands virtual machines. Without appropriate tools, this administration task may be impossible to achieve.

We purpose in [32] a decision support tool to detect virtual machines with atypical behavior. Virtual machines whose behavior is different from other VMs running in the data center are tagged as atypicals. Our analysis tool is based on a specific partitioning algorithm which identifies VM behaviors. This tool has been validated in production environments and is used by several companies.

To collect finer metrics (for security, energy management etc.), VM introspection an agent can be installed in a VM to intrusively supervise it or the hypervisor can be used to non-intrusively recover the introspection metrics. In the case of intrusive introspection, the agent installed on the VM operating system will retrieve a set of information related to the operating system operation. However, the installation of an agent in the virtual machine increases the cost of deploying the virtual machine and its resource consumption. The Virtual Machine Introspection (VMI) at the hypervisor level (non intrusively) offer a complete, consistent and untainted view of the VM state. This solution allows an isolation of the VMI mechanism from the guest OS, while allowing monitoring and modifying any state of the VM.

We have also provided a comprehensive summary on VM introspection techniques [25]. Existing VMI techniques are analyzed with respect to their approach to closing the "semantic gap" between the (low level) information provided by the hypervisor and the input to the security analysis.

Finally, we have introduced an extension to LibVmi to detect and monitor a process resource consumption inside a VM from the hypervisor [34]. This extension monitor process cpu and ram ressources without probe. This extension can detect abusive cpu resource usage and atypical ram utilization. This fine monitoring system can be used in many context (security, power consumption, fault tolerance).

6.5. Green IT

Participants: Simon Dupont, Md Sabbir Hasan, Thomas Ledoux, Jonathan Lejeune, Guillaume Le Louët, Jean-Marc Menaud.

In 2015, we have provided new models and solutions for the energy-optimal execution of cloud applications in data centers.

6.5.1. Renewable energy

With the emergence of the Future Internet and the emergence of new IT models such as cloud computing, the usage of data centers (DC) and consequently their power consumption increase dramatically. Besides the ecological impact, the energy consumption is a predominant criteria for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale distributed systems.

6.5.1.1. Renewable energy for data centers

We have presented the EPOC project which focuses on optimizing the energy consumption of mono-site DCs connected to the regular electrical grid and to renewable energy sources [19]. A first challenge in this context consists in developing a (for users) transparent distributed system that enables energy-proportional computations from the system to service-oriented levels. The second challenge addresses the corresponding energy issues through collaborative measurements and energy-optimizing actions inside infrastructure-software stack, more precisely between applications and resource management systems. This approach must manage Service Level Agreement (SLA) constraints by striving for the best trade-off between energy cost (from the regular electric grid), its availability (from renewable energy sources), and service degradation (from application reconfiguration issues to job suspension ones). The third challenge embarks pursues energy efficient optical networks as key enablers of the future internet and cloud-networking service deployment through the convergence of optical infrastructure with the upper network layers.

The second challenge is more precisely describe in [30]. In this paper we present PIKA, a framework aiming at reducing the Brownian energy consumption (ie. from non renewable energy sources), and improving the usage of renewable energy for mono-site data centers. PIKA exploits jobs with slack periods, and executes and suspends them depending on the available renewable energy supply. By consolidating the virtual machines (VMs) on the physical servers, PIKA adjusts the number of powered-on servers in order for the overall energy consumption to match the renewable energy supply. Using simulations driven by real-world workloads and solar power traces, we demonstrate that PIKA consumes 41% less Brownian energy and increases 35.3% renewable energy integration ratio in comparison with the baseline algorithm from the literature.

6.5.1.2. Energy monotoring

We have designed SensorScript, a Business-Oriented Domain-Specific Language for Sensor Networks [24], [35]. In smart grids, or more generally the Internet of Things, many research work has been performed on the whole chain, from communication sensors to big data management, through communication middlewares. Few of this work have addressed the problem of gathered data access. In fact, being able, as a system administrator, to manipulate and gather data collected from a set of sensors in a simple and efficient way represents an essential need.

To address this issue, the solution we considered consists of a multi-context modeling for raw data, in the form of a multi-tree: a directed acyclic graph consisting of multiple intricate trees, each of them describing a hierarchy corresponding to a given use context. The objectives are to provide not only a means to rationalize users needs before writing queries, but also to offer a domain-specific language (DSL) which takes advantage of the multi-tree modeling to simplify the experience of pre-identified users that query data.

6.5.1.3. Green SLA and virtualization of green energy

The demand for energy-efficient services is increasing considerably as people are getting more environmentally-conscious in order to build a sustainable society. The main challenge for Cloud providers is to manage Green SLA (Service Level Agreement) constraints for their customers while satisfying their business objectives, such as maximizing profits by lowering expenditure for so-called green (renewable) energy. Since, Green SLA needs to be proposed based on the presence of green energy, the intermittent nature of renewable sources makes it difficult to be achieved. In response, we propose a scheme for green energy management based on three contributions [15]: i) we introduce the concept of virtualization of green energy to address the uncertainty of green energy availability, ii) we extend the Cloud Service Level Agreement (CSLA) language to support Green SLA by introducing two new threshold parameters and iii) we introduce algorithms for Green SLA which leverage the concept of virtualization of green energy to provide interval-specific Green SLA. We have conducted experiments with real workload profiles from PlanetLab and server power model from SPECpower to demonstrate that Green SLA can be successfully established and satisfied without incurring higher cost.

DIVERSE Project-Team

7. New Results

7.1. Results on Software Language Engineering

7.1.1. Modular and Reusable Development of DSLs

Domain-Specific Languages (DSLs) are now developed for a wide variety of domains to address specific concerns in the development of complex systems. When engineering new DSLs, it is likely that previous efforts spent on the development of other languages could be leveraged, especially when their domains overlap. However, legacy DSLs may not fit exactly the end user requirements and thus require further extension, restriction, or specialization. While current language workbenches provide import mechanisms, they usually lack an explicit support for such customizations of imported artifacts. We propose an approach for building DSLs by safely assembling and customizing legacy DSLs artifacts. This approach is based on typing relations that provide a reasoning layer for manipulating DSLs while ensuring type safety. On top of this reasoning layer, we provide an algebra of operators for extending, restricting, and assembling separate DSL artifacts. We implemented the typing relations and algebra into the Melange meta-language [30], [29], [73].

7.1.2. Executable Domain-Specific Modeling Languages (xDSMLs)

Executable Domain-Specific Modeling Languages (xDSMLs) open many possibilities for performing early verification and validation (V&V) of systems. Dynamic V&V approaches rely on execution traces, which represent the evolution of models during their execution. In order to construct traces, generic trace metamodels can be used. Yet, regarding trace manipulations, they lack both efficiency because of their sequential structure, and usability because of their gap to the xDSML. We contributed a generative approach that defines a rich and domain-specific trace metamodel enabling the construction of execution traces for models conforming to a given xDSML [24]. We also contributed a partly generic omniscient debugger supported by generated domain-specific trace management facilities [49].

The emergence of modern concurrent systems calls for xDSMLs where concurrency is of paramount importance. Such xDSMLs are intended to propose constructs with rich concurrency semantics, which allow system designers to precisely define and analyze system behaviors. In [34], we introduce a concurrent executable metamodeling approach, which supports a modular definition of the execution semantics, including the concurrency model, the semantic rules, and a well-defined and expressive communication protocol between them. In [28], we present MoCCML, a dedicated meta-language for formally specifying the concurrency concern within the definition of a DSL. The concurrency constraints can reflect the knowledge in a particular domain, but also the constraints of a particular platform. MoCCML comes with a complete language workbench to help a DSL designer in the definition of the concurrency directly within the concepts of theDSL itself, and a generic workbench to simulate and analyze any model conforming to this DSL. MoCCML is illustrated on the definition of an lightweight extension of SDF (SynchronousData Flow).

7.1.3. Globalization of Domain-Specific Modeling Languages

The development of modern complex software-intensive systems often involves the use of multiple DSMLs that capture different system aspects. Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages, that is, the use of multiple modeling languages to support coordinated development of diverse aspects of a system.

In a book published in 2015 [66], a number of articles describe the vision and the way globalized DSMLs currently assist integrated DSML support teams working on systems that span many domains and concerns to determine how their work on a particular aspect influences work on other aspects. Globalized DSMLs offer support for communicating relevant information, and for coordinating development activities and associated technologies within and across teams, in addition to providing support for imposing control over development artifacts produced by multiple teams. DSMLs can be used to support socio-technical coordination by providing the means for stakeholders to bridge the gap between how they perceive a problem and its solution, and the programming technologies used to implement a solution. They also support coordination of work across multiple teams. DSMLs developed in an independent manner to meet the specific needs of domain experts have an associated framework that regulates interactions needed to support collaboration and work coordination across different system domains. The book includes [63], [65], [64], [62] with authors from the DIVERSE team.

In [43], we propose a Behavioral Coordination Operator Language (B-COOL) to reify coordination patterns between specific domains by using coordination operators between the Domain-Specific Modeling Languages used in these domains. Those operators are then used to automate the coordination of models conforming to these languages. We illustrate the use of B-COOL with the definition of coordination operators between timed finite state machines and activity diagrams.

The GEMOC Studio (http://gemoc.org/studio) is an eclipse package that contains components for building and composing executable Domain-Specific Modeling Languages (DSMLs). The GEMOC Studio complements Melange to formally define in a modular way the concurrency model of executable DSMLs, and provides analysis and coordination facilities based on the concurrency model. It also integrates all the contributions presented in this document related to model execution, animation, debugging and trace management. The GEMOC studio has been the overall winner of the transformation tool contest 2015 on Model Execution [52].

7.1.4. An analysis of metamodeling practices for MOF and OCL

The definition of a metamodel that precisely captures domain knowledge for effective know-how capitalization is a challenging task. A major obstacle for domain experts who want to build a metamodel is that they must master two radically different languages: an object-oriented, MOF-compliant, modeling language to capture the domain structure and first order logic (the Object Constraint Language) for the definition of well-formedness rules. However, there are no guidelines to assist the conjunct usage of both paradigms, and few tools support it. Consequently, we observe that most metamodels have only an object-oriented domain structure, leading to inaccurate metamodels. In [21], we perform the first empirical study, which analyzes the current state of practice in metamodels that actually use logical expressions to constrain the structure. We analyze 33 metamodels including 995 rules coming from industry, academia and the Object Management Group, to understand how metamodelers articulate both languages. We implement a set of metrics in the OCLMetrics tool to evaluate the complexity of both parts, as well as the coupling between both. We observe that all metamodels tend to have a small, core subset of concepts, which are constrained by most of the rules, in general the rules are loosely coupled to the structure and we identify the set of OCL constructs actually used in rules.

7.1.5. Model Slicers

Among model comprehension tools, model slicers are tools that extract a subset of model elements, for a specific purpose. We propose the Kompren language to model and generate model slicers for any DSL (*e.g.* modeling for software development or for civil engineering) and for different purposes (*e.g.* monitoring and model comprehension). We detail the semantics of the Kompren language and of the model slicer generator. This provides a set of expected properties about the slices that are extracted by the different forms of the slicer [18]. We show how the use of Kompren, a domain-specific language for defining model slicers, can ease the development of such interactive visualization features [19].

In Model Driven Development (MDD), it is important to ensure that a model conforms to the invariants defined in the metamodel. General-purpose rigorous analysis tools that check invariants are likely to perform the analysis over the entire metamodel and model. Since modern day software is exceedingly complex, the

size of the model together with the metamodel can be very large. Consequently, invariant checking can take a very long time.To this end, we introduce model slicing within the invariant checking process, and use a slicing technique to reduce the size of the inputs in order to make invariant checking of large models feasible with existing tools [22], [42].

7.1.6. Bridging the gap between scientific models and engineering models with MDE

The complex problems that computational science addresses are more and more benefiting from the progress of computing facilities (e.g., simulators, librairies, accessible languages). Nevertheless, the actual solutions call for several improvements. Among those, we address in [25] the needs for leveraging on knowledge and expertise by focusing on Domain-Specific Modeling Languages application. In this vision paper we illustrate, through concrete experiments, how the last DSML research help getting closer the problem and implementation spaces.

Various disciplines use models for different purposes. While engineering models, including software engineering models, are often developed to guide the construction of a non- existent system, scientific models, in contrast, are created to better understand a natural phenomenon (i.e., an already existing system). An engineering model may incorporate scientific models to build a system. Both engineering and scientific models have been used to support sustainability, but largely in a loosely-coupled fashion, independently developed and maintained from each other. Due to the inherent complex nature of sustainability that must balance trade-offs between social, environmental, and economic concerns, modeling challenges abound for both the scientific and engineering disciplines. In [72] we propose a vision that synergistically combines engineering and scientific models to enable broader engagement of society for addressing sustainability concerns, informed decision-making based on more accessible scientific models and data, and automated feed-back to the engineering models to support dynamic adaptation of sustainability systems. To support this vision, we identify a number of challenges to be addressed with particular emphasis on the socio-technical benefits of modeling.

As first experiments, we presented at the Inria-Industry meeting 2015 on energy transition and EclipseCon 2015, an approach to develop smart cyber physical systems in charge of managing the production, distribution and consumption of energies (e.g., water, electricity). The main objective is to enable a broader engagement of society, while supporting a more informed decision-making, possibly automatically, on the development and run-time adaptation of sustainability systems (e.g., smart grid, home automation, smart cities). We illustrate this approach through a system that allows farmers to simulate and optimize their water consumption by combining the model of a farming system together with agronomical models (e.g., vegetable and animal lifecycle) and open data (e.g., climate series). To do so, we use Model Driven Engineering (MDE) and Domain Specific Languages (DSL) to develop such systems driven by scientific models that define the context (e.g., environment, social and economy), and model experiencing environments to engage general public and policy makers.

7.2. Results on Variability Modeling and Engineering

7.2.1. Reverse engineering variability

We have developed automated techniques and a comprehensive environment for synthesizing feature models from various kinds of artefacts (e.g. propositional formula, dependency graph, FMs or product comparison matrices). Specifically we have elaborated a support (through ranking lists, clusters, and logical heuristics) for choosing a sound and meaningful hierarchy [93]. We have performed an empirical evaluation on hundreds of feature models, coming from the SPLOT repository and Wikipedia [92]. We have showed that a hybrid approach mixing logical and ontological techniques outperforms state-of-the-art solutions (to appear in Empirical Software Engineering journal in 2015 [20]). We have also considered numerical information and feature *attributes* so that we are now capable of synthesizing attributed feature models from product descriptions [51].

Besides, we have developed techniques for reverse engineering variability in generators and configurators (e.g., video generators) [50]. We have identified new research directions for protecting variability [44] mainly due to the fact reverse engineering techniques (previously presented) are effective .

7.2.2. Product comparison matrices

Product Comparison Matrices (PCMs) constitute a rich source of data for comparing a set of related and competing products over numerous features. PCMs can be seen as a formalism for modeling a family of products, including variability information. Despite their apparent simplicity, PCMs contain heterogeneous, ambiguous, uncontrolled and partial information that hinders their efficient exploitations. We have formalized PCMs through model-based automated techniques and developed additional tooling to support the edition and re-engineering of PCMs [94]. 20 participants used our editor to evaluate our PCM metamodel and automated transformations. The empirical results over 75 PCMs from Wikipedia show that (1) a significant proportion of the formalization of PCMs can be automated: 93.11% of the 30061 cells are correctly formalized; (2) the rest of the formalization can be realized by using the editor and mapping cells to existing concepts of the metamodel. The ASE'2014 paper opens avenues for engaging a community in the mining, re-engineering, edition, and exploitation of PCMs that now abound on the Internet. We have launched an open, collaborative initiative towards this direction https://opencompare.org/

Another axis is the mining of PCMs since (1) the manual elaboration of PCMs has limitations (2) numerous sources of information can be combined and are amenable to PCMs. We have developed MatrixMiner a tool for automatically synthesizing PCMs from a set of product descriptions written in natural language [46]. MatrixMiner is capable of identifying and organizing features and values in a PCM despite the informality and absence of structure in the textual descriptions of products. More information is available online: https://matrix-miner.variability.io/

7.3. Results on Heterogeneous and dynamic software architectures

7.3.1. Resource Monitoring and Reservation in Heterogeneous and dynamic software architectures

Software systems are more pervasive than ever nowadays. Occasionally, applications run on top of resource-constrained devices where efficient resource management is required; hence, they must be capable of coping with such limitations. However, applications require support from the runtime environment to properly deal with resource limitations. This thesis addresses the problem of supporting resource-aware programming in execution environments. In particular, it aims at offering efficient support for collecting data about the consumption of computational resources (e.g., CPU, memory), as well as efficient mechanisms to reserve resources for specific applications. In existing solutions we find two important drawbacks. First, they impose performance overhead on the execution of applications. Second, creating resource management tools for these abstractions is still a daunting task. The outcomes of this work [12] are three contributions:

- An optimistic resource monitoring framework that reduces the cost of collecting resource consumption data.
- A methodology to select components' bindings at deployment time in order to perform resource reservation.
- A language to build customized memory profilers that can be used both during applications' development, and also in a production environment.

7.3.2. Dynamic Reasoning on Heterogeneous and dynamic software architectures

Multi-Objective Evolutionary Algorithms (MOEAs) have been successfully used to optimize various domains such as finance, science, engineering, logistics and software engineering. Nevertheless, MOEAs are still very complex to apply and require detailed knowledge about problem encoding and mutation operators to obtain an effective implementation. Software engineering paradigms such as domain-driven design aim to tackle this complexity by allowing domain experts to focus on domain logic over technical details. Similarly, in order to handle MOEA complexity, we propose an approach, using model-driven software engineering (MDE) techniques, to define fitness functions and mutation operators without MOEA encoding knowledge. Integrated into an open source modelling framework, our approach can significantly simplify development and maintenance of multi-objective optimizations. By leveraging modeling methods, our approach allows reusable

optimizations and seamlessly connects MOEA and MDE paradigms. We evaluate our approach on a cloud case study and show its suitability in terms of i) complexity to implement an MOO problem, ii) complexity to adapt (maintain) this implementation caused by changes in the domain model and/or optimization goals, and iii) show that the efficiency and effectiveness of our approach [56] remains comparable to ad-hoc implementations.

7.3.3. A Precise Metamodel for Open Cloud Computing Interface

Open Cloud Computing Interface (OCCI) proposes one of the first widely accepted, community-based, open standards for managing any kinds of cloud resources. But as it is specified in natural language, OCCI is imprecise, ambiguous, incomplete, and needs a precise definition of its core concepts. Indeed, the OCCI Core Model has conceptual drawbacks: an imprecise semantics of its type classification system, a nonextensible data type system for OCCI attributes, a vague and limited extension concept and the absence of a configuration concept. To tackle these issues, this work proposes a precise metamodel for OCCI. This metamodel defines rigourously the static semantics of the OCCI core concepts, of a precise type classification system, of an extensible data type system, and of both extension and configuration concepts. This metamodel is based on the Eclipse Modeling Framework (EMF), its structure is encoded with Ecore and its static semantics is rigourously defined with Object Constraint Language (OCL). As a consequence, this metamodel provides a concrete language to precisely define and exchange OCCI models. The validation of our metamodel is done on the first worldwide dataset of OCCI extensions already published in the literature, and addressing inter-cloud networking, infrastructure, platform, application, service management, cloud monitoring, and autonomic computing domains, respectively. This validation highlights simplicity, consistency, correctness, completeness, and usefulness of the proposed metamodel[38], [41].

7.3.4. Using Novelty Search Approach and models@runtime for Automatic Testing Environment Setup

In search-based structural testing, metaheuristic search techniques have been frequently used to automate the test data generation. In Genetic Algorithms (GAs) for example, test data are rewarded on the basis of an objective function that represents generally the number of statements or branches covered. However, owing to the wide diversity of possible test data values, it is hard to find the set of test data that can satisfy a specific coverage criterion. In this work, we introduce the use of Novelty Search (NS) algorithm to the test data generation problem based on statement-covered criteria. We believe that such approach to test data generation is attractive because it allows the exploration of the huge space of test data within the input domain. In this approach, we seek to explore the search space without regard to any objectives. In fact, instead of having a fitness-based selection, we select test cases based on a novelty score showing how different they are compared to all other solutions evaluated so far [47], [48]. We also create an architecture generation framework for setup testing environment for a distributed and heterogeneous service.

7.3.5. Using Models@Run.time to embed an Energetic Cloud Simulator in a MAPE-K Loop

Due to high electricity consumption in the Cloud datacenters, providers aim at maximizing energy efficiency through VM consolidation, accurate resource allocation or adjusting VM usage. More generally, the provider attempts to optimize resource utilization. However, while minimizing expenses, the Cloud operator still needs to conform to SLA constraints negotiated with customers (such as latency, downtime, affinity, placement, response time or duplication). Consequently, optimizing a Cloud configuration is a multi-objective problem. As a nontrivial multi-objective optimization problem, there does not exist a single solution that simultaneously optimizes each objective. There exists a (possibly infinite) number of Pareto optimal solutions. Evolutionary algorithms are popular approaches for generating Pareto optimal solutions to a multi-objective optimization problem. Most of these solutions use a fitness function to assess the quality of the candidates. However, regarding the energy consumption estimation, the fitness function can be approximative and lead to some imprecisions compared to the real observed data. This work presents a system that uses a genetic algorithm to optimize Cloud energy consumption and machine learning techniques to improve the fitness function regarding a real distributed cluster of server. We have carried out experiments on the OpenStack platform to validate our solution. This experimentation shows that the machine learning produces an accurate energy model, predicting precise values for the simulation [124][40].

7.4. Results on Diverse Implementations for Resilience

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

7.4.1. Software diversification

Early experiments with software diversity in the mid 1970's investigated N-version programming and recovery blocks to increase the reliability of embedded systems. Four decades later, the literature about software diversity has expanded in multiple directions: goals (fault-tolerance, security, software engineering); means (managed or automated diversity) and analytical studies (quantification of diversity and its impact). We contribute to the field of software diversity with the very first literature survey that adopts an inclusive vision of the area, with an emphasis on the most recent advances in the field. This survey includes classical work about design and data diversity for fault tolerance, as well as the cybersecurity literature that investigates randomization at different system levels. It broadens this standard scope of diversity, to include the study and exploitation of natural diversity and the management of diverse software products [17].

We also contribute to software diversity with novel techniques and methods. The interdisciplinary investigations within the DIVERSIFY project have led to the definition of novel principles for open-ended evolution in software systems. The main intuition is that software should have the ability to spontaneously and continuously evolve without waiting for specific environmental conditions. Our proposal analogizes the software consumer / provider network, which can be found in any types of distributed systems, to a bipartite ecological graph. This analogy provides the foundations for the design of an individual-based simulator used to experiment with decentralized adaptation strategies for providers and consumers. The initial model of a software network is tuned according to observations gathered from real-world software networks. The key insights about our experiments are that, 1) we can successfully model software systems as an ALife system, and 2) we succeed in emerging a global property from local decisions: when consumers and providers adapt with local decision strategies, the global robustness of the network increases. We show that these results hold with different initial situations, different scales and different topological constraints on the network [55]. In order to move towards the open-ended evolution of actual systems, we also developed a novel tool for the runtime modification of Java programs, as an extension to the JVM [60].

Our second contribution to the filed of software diversity consists in experimenting its application in different fields. First, we have proposed a novel approach to exploit software diversity at multiple granularity levels simultaneously [15]. The main idea is to reconcile two aspects of the massive software reuse in web applications: on the one hand, reuse and modularity favor much writing the next killer application; on the other hand, reuse and modularity facilitates much the next massive BOBE attack. We demonstrate the feasiblity of diversifying web applications at multiple levels, mitigating the risks of reuse.

The second application of automatic software diversification for Java programs aimed at answering the following question: which product line operators, applied to which program elements , can synthesize variants of programs that are incorrect , correct or perhaps even conforming to test suites? We implement source code transformations, based on the derivation operators of the Common Variability Language. We automatically synthesize more than 370,000 program variants from a set of 8 real large Java projects (up to 85,000 lines of code), obtaining an extensive panorama of the sanity of the operations [68].

The third application of software diversification is against browser fingerprinting. Browser fingerprint tracking relies on the following mechanisms: web browsers allow remote servers to discover sufficient information about a user's platform to create a digital fingerprint that uniquely identifies the platform. We argue that fingerprint uniqueness and stability are the key threats to browser fingerprint tracking, and we aim at breaking fingerprint stability over time, by exploiting software diversity and automatic reconfiguration. We leverage

virtualization and modular software architectures to automatically assemble and reconfigure a user's software components at multiple levels. We operate on the operating system, the browser, the lists of fonts and plugins. This work is the first application of software reconfiguration to build a moving target defense against browser fingerprint tracking. We have developed a prototype called *Blink* to experiment the effectiveness of our approach at randomizing fingerprints [33].

7.4.2. Software testing

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

Graphical user interfaces (GUIs) are integral parts of software systems that require interactions from their users. Software testers have paid special attention to GUI testing in the last decade, and have devised techniques that are effective in finding several kinds of GUI errors. However, the introduction of new types of interactions in GUIs presents new kinds of errors that are not targeted by current testing techniques. We believe that to advance GUI testing, the community needs a comprehensive and high level GUI fault model, which incorporates all types of interactions. In this work, we first propose a GUI fault model designed to identify and classify GUI faults [37]. We then studied the impact of the new types of interactions in GUIs on their testing process. We show that the current GUI model-based testing approaches have limits when applied to test such new advanced GUIs [36].

Specifying a model transformation is challenging as it must be able to give a meaningful output for any input model in a possibly infinite modeling domain. Transformation preconditions constrain the input domain by rejecting input models that are not meant to be transformed by a model transformation. In our latest work [39], we present a systematic approach to discover such preconditions when it is hard for a human developer to foresee complex graphs of objects that are not meant to be transformed. The approach is based on systematically generating a finite number of test models using our tool, PRAMANA to first cover the input domain based on input domain partitioning. Tracing a transformation's execution reveals why some preconditions are missing. Using a benchmark transformation from simplified UML class diagram models to RDBMS models we discover new preconditions that were not initially specified.

We also initiated a new line of research in order to invesitgate Novelty Search (NS) for the automatic generation of test data. This allows the exploration of the huge space of test data within the input domain. In this approach, we select test cases based on a novelty score showing how different they are compared to all other solutions evaluated so far [47].

In Model Driven Engineering (MDE), models are first-class citizens, and model transformation is MDE's "heart and soul". Since model transformations are executed for a family of (conforming) models, their validity becomes a crucial issue. In [16] we proposes to explore the question of the formal verification of model transformation properties through a tridimensional approach: the transformation involved, the properties of interest addressed, and the formal verification techniques used to establish the properties. This work is intended for a double audience. For newcomers, it provides a tutorial introduction to the field of formal verification of model transformations. For readers more familiar with formal methods and model transformations, it proposes a literature review (although not systematic) of the contributions of the field. Overall, this work allows to better understand the evolution, trends and current practice in the domain of model transformation verification. This work opens an interesting research line for building an engineering of model transformation verification guided by the notion of model transformation intent.

FOCUS Project-Team

7. New Results

7.1. Service-oriented computing

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

7.1.1. Orchestrations

Orchestration models and languages in the context of Service-Oriented Architectures (SOA) are used to describe the composition of services focusing on their interactions. Concrete web services are connected to abstract service definitions for the aim of service discovery. In [16] we study a natural notion of compliance between clients and services in terms of their bpel (abstract) descriptions. The induced preorder shows interesting connections with the must preorder and has normal form representatives that are parallel-free finite-state activities, called contracts. Moreover, in [22] we focus on advancements of the orchestration language Jolie aiming at the development of dynamically adaptable orchestrated systems.

7.1.2. Choreographies

Choreographies are high-level descriptions of distributed interacting systems featuring as basic unit a communication between two participants. A main feature of choreographies is that they ensure deadlock-freedom by construction. From a choreography one can automatically derive a description of the behaviour of each participant using a notion of projection. Choreographies can be used both at the level of types (multiparty session types) or as a programming language. In [19] we surveyed our results about verification of adaptable processes, focusing in particular on distributed adaptability, where a process can update part of a protocol (specified by a choreography) by performing dynamic distributed updates over a set of protocol participants. In [14] we illustrate our approach to develop and verify distributed, adaptive software systems. The cornerstone of our framework is the use of choreography languages, which allow us to obtain correctness by construction. Moreover, in [36] we present DIOC, a language for programming distributed applications that are free from deadlocks and races by construction. A DIOC program describes a whole distributed application as a unique entity (choreography). DIOC allows the programmer to specify which parts of the application can be updated. At runtime, these parts may be replaced by new DIOC fragments from outside the application. DIOC programs are compiled, generating code for each site, in a lower-level language called DPOC. As a consequence DPOC applications are free from communication deadlocks and races, even in presence of runtime updates.

7.2. Models for reliability

Participants: Elena Giachino, Ivan Lanese.

7.2.1. Reversibility

We have continued the study of causal-consistent reversibility started in the past years. In [42] we defined the causal-consistent reversible semantics (both controlled and uncontrolled) of muKlaim, a formal coordination language based on distributed tuple spaces, by adapting the approach developed for message passing calculi in the past years. A major novelty is that the muKlaim read primitive allows two processes to access a shared resource independently, giving rise to a causality structure which is not found in message passing calculi.

In [31] we studied the issue of compliance of a client w.r.t. a server in a reversible setting using behavioural contracts. In particular, when an agreement cannot be reached, the client and the server can synchronously rollback to the last point of choice, looking for alternatives. As a main result, we showed that compliance is decidable even for recursive contracts.

7.3. Cloud Computing and Deployment

Participants: Elena Giachino, Saverio Giallorenzo, Claudio Guidi, Cosimo Laneve, Michael Lienhardt, Jacopo Mauro, Gianluigi Zavattaro.

7.3.1. Cloud application deployment

Configuration and management of applications in the cloud is a complex task that requires advanced methodologies and tools. A foundational study of the problem has been carried out in [21] where we have identified the critical tasks to be solved, quantified their computational complexity, and proposed simplifications to the problem with the idea of limiting the computational complexity at the cost of having approximated (but acceptable, in most cases) solutions. Our attention has been dedicated to the implementation of a tool for the efficient solution of one of these tasks, namely, the automatic planning of the management actions needed to properly configure a cloud application [17]. This tool, called Metis, has been already exploited in the realization of an integrated platform for the automatic deployment of the cloud application called Blender [39] as well as in the context of the ABS modeling language [37] in order to be able to support the automatic reasoning about deployment costs already during the early phases of application design and development. We have also performed a foundational study of the problem of reconfiguring an application instead of deploying it from scratch. Our foundational study allowed us, on the one hand, to quantify the computational complexity of the problem (PSpace-Complete) and, on the other hand, to precisely identify the source of such complexity (the presence of legacy components that cannot be re-deployed from scratch).

7.3.2. Cloud resource management

One of the main challenges in the management of cloud applications is the quantification of the computing resources needed by the applications to be deployed. More precisely, it is important to quantify upper bounds to the number of needed computing resources in order to either previously acquire them or have a precise quantification of the costs for executing an application. In [40] a static analysis technique is proposed that computes upper bounds of virtual machine usages in a concurrent language with explicit acquire and release operations of virtual machines. See the section on deadlock analysis for more details.

7.4. Resource Control and Probabilities

Participants: Michele Alberti, Martin Avanzini, Flavien Breuvart, Alberto Cappai, Ugo Dal Lago, Simone Martini, Giulio Pellitta, Alessandro Rioli, Davide Sangiorgi, Marco Solieri, Valeria Vignudelli.

7.4.1. Resource Control

7.4.1.1. Time Complexity Analysis of Concurrent and Higher-Order Functional Programs

We have extensively studied the problem of automatically analysing the complexity of programs. We first of all studied the problem for concurrent object-oriented programs [41]. To determine this complexity we have used intermediate abstract descriptions that record relevant information for the time analysis, called behavioural types. Behavioural types are then translated into so-called cost equations, making parallelism explicit. Cost equations are finally fed into an automatic off-the-shelf solver for obtaining the actual time complexity. The same problem has been also analysed when the underlying program is functional [29]. We showed how the complexity of higher-order functional programs can be analysed automatically by applying program transformations to a defunctionalized version of them, and feeding the result to existing tools for the complexity analysis of first-order term rewrite systems. This is done while carefully analysing complexity preservation and reflection of the employed transformations such that the complexity of the obtained term rewrite system reflects on the complexity of the initial program. This approach turns out to work well in practice, in particular since off-the-shelf complexity tool for first-order rewrite systems matured to a state where they are both fast and powerful. However, the implementation of such tools is quite sophisticated. To ensure correctness of the obtained complexity bounds, we extended CeTA, a certified proof checker for rewrite tools, with the formalisation of various complexity techniques underlying state-of-the-art complexity tools [30]. This way, we detected conflicts in theoretical results as well as bugs in existing complexity provers.

7.4.1.2. Function Algebras and Implicit Complexity

A fundamental result about ramified recurrence, one of the earliest systems in implicit complexity, has been proved [28]. This has been obtained through a careful analysis on how the adoption of an evaluation mechanism with sharing and memoization impacts the class of functions which can be computed in polynomial time. We have first shown how a natural cost model in which lookup for an already computed result has no cost is indeed invariant. As a corollary, we have then proved that the most general notion of ramified recurrence is sound for polynomial time.

7.4.1.3. Geometry of Interaction

We see the the geometry of interaction as a foundational framework in which the efficiency of higher-order computation can be analyzed. This has produced some very interesting results, also stimulated by the bilateral Inria project CRECOGI, which started this year. We have first of all studied the geometry of interaction of the resource lambda-calculus, a model of linear and nondeterministic functional languages. In a strictly typed restriction of the resource lambda-calculus, we have studied the notion of path persistence, and defined a geometry of interaction that characterises it [18]. Furthermore, we have carried out our work on multitoken machines, started in 2014. More specifically, we have studied multitoken interaction machines in the context of a very expressive linear logical system with exponentials, fixpoints and synchronization [34]. On the one hand, we have proved that interaction is guaranteed to be deadlock-free. On the other hand, the resulting logical system has been proved to be powerful enough to embed PCF and to adequately model its behaviour, both when call-by-name and when call-by-value evaluation are considered.

7.4.2. Probabilistic Models

7.4.2.1. Applicative Bisimilarity

Notions of equivalences for probabilistic programming languages have been studied and analysed, together with their relationships with context equivalence. More specifically, we have studied how applicative bisimilarity behaves when instantiated on a call-by-value probabilistic lambda-calculus, endowed with Plotkin's parallel disjunction operator [20]. We have proved that congruence and coincidence with the corresponding context relation hold for both bisimilarity and similarity, the latter known to be impossible in sequential languages. We have also shown that applicative bisimilarity works well when the underlying language of programs takes the form of a linear lambda-calculus extended with quantum data [35]. The main results are proofs of soundness for the obtained notion of equivalence.

7.4.2.2. From Equivalences to Metrics

The presence of probabilistic (thus quantitative) notions of observation makes equivalence relations too coarse-grained as ways to compare programs. This opens the way to metrics in which, indeed, not all non-equivalent programs are at the same distance. We have studied the problem of evaluating the distance between affine lambda-terms [33]. A natural generalisation of context equivalence has been shown to be characterised by a notion of trace distance, and to be bounded from above by a coinductively defined distance based on the Kantorovich metric on distributions. A different, again fully-abstract, tuple-based notion of trace distance has been shown to be able to handle nontrivial examples. A similar thing has been done in a calculus for probabilistic polynomial time computation [32], thus paving the way towards getting effective proof methodologies for computational indistinguishability, a key notion in modern cryptography.

7.5. Verification techniques for extensional properties

Participants: Daniel Hirschkoff, Elena Giachino, Michael Lienhardt, Cosimo Laneve, Jean-Marie Madiot, Davide Sangiorgi.

Extensional properties are those properties that constrain the behavioural descriptions of a system (i.e., how a system looks like from the outside). Examples of such properties include classical functional correctness, deadlock freedom and resource usage. Related to techniques for extensional properties are the issues of decidability (the problem of establishing whether certain properties are computationally feasible).

7.5.1. Static analysis of deadlock freedom and resource usage

Deadlock detection in concurrent programs that create networks with an arbitrary number of nodes is extremely complex and solutions either give imprecise answers or do not scale. To enable the analysis of such programs, we have studied an algorithm for detecting deadlocks in a basic concurrent object-oriented language. The algorithm (i) associates behavioural types, called *lam*, to programs by means of a type inference system and (ii) uses an ad-hoc verification technique highlighting circular dependencies in lam [15]. The algorithm has been prototyped and has been extended to a full-fledged programming language, called ABS.

A technique similar to [15] has been used for computing upper bounds of resource usages in [40]. In particular, the metaphor in this paper has been virtual machines usage in a concurrent language with explicit acquire and release operations. The problematic issue in such languages is when the release is delegated to other (ad-hoc or third party) concurrent codes (by passing them as arguments of invocations) – a feature that is currently used in Amazon Elastic Cloud Computing or in the Docker FiWare. As for deadlock analysis, the technique is modular and consists of (i) a type system associating programs with behavioural types that records relevant information for resource usage (creations, releases, and concurrent operations), (ii) a translation function that takes behavioural types and returns cost equations, and (iii) an automatic off-the-shelf solver for the cost equations. A soundness proof of the type system establishes the correctness of the above technique with respect to the cost equations. The technique has also been experimentally evaluated and the experiments show that it allows us to derive bounds for programs that are better than other techniques, such as those based on amortized analysis.

Another technique for enforcing program correctness is the one used in [36], [14], where the programming of distributed applications is guaranteed to be free from communication deadlocks and races by means of *choreographies*. Choreographies are behavioural types which allow one to obtain correctness by construction (more details on this paper in Section 7.1).

7.5.2. Name mobility

The article [44] studies the behavioural theory of πP, a π-calculus featuring restriction as the only binder. In contrast with calculi such as Fusions and Chi, reduction in πP generates a preorder on names rather than an equivalence relation. Two characterisations of barbed congruence in πP are analyzed: the first is based on a compositional LTS, and the second is an axiomatisation. The results in this paper bring out basic properties of πP, mostly related to the interplay between the restriction operator and the preorder on names.

7.5.3. Coinductive techniques

Coinductive techniques, notably those based on bisimulation, are widely used in concurrency theory to reason about systems of processes. The bisimulation proof method can be enhanced by employing "bisimulations up-to" techniques. A comprehensive theory of such enhancements has been developed for first-order (i.e., CCS-like) LTSs and bisimilarity, based on the notion of compatible function for fixed-point theory.

A proof method different from bisimulation is investigated in [46], [23]. This method is based on unique solution of special forms of inequations called contractions, and inspired by Milner's theorem on unique solution of equations. The method is as powerful as the bisimulation proof method and its "up-to context" enhancements. The definition of contraction can be transferred onto other behavioural equivalences, possibly contextual and non-coinductive. This enables a coinductive reasoning style on such equivalences, either by applying the method based on unique solution of contractions, or by injecting appropriate contraction pre-orders into the bisimulation game. The technique can be applied both to first-order languages and to higher-order languages.

7.5.4. Expressiveness and decidability in actor-like systems

In [48], the limit of classical Petri nets is studied by discussing when it is necessary to move to the so-called Transfer nets, in which transitions can also move to a target place all the tokens currently present in a source place. More precisely, we consider a simple calculus of processes that interact by generating/consuming messages into/from a shared repository. For this calculus classical Petri nets can faithfully model the process

behavior. Then we present a simple extension with a primitive allowing processes to atomically rename all the data of a given kind. We show that with the addition of such primitive it is necessary to move to Transfer nets to obtain a faithful modeling.

7.6. Constraint Programming

Participants: Roberto Amadini, Maurizio Gabbrielli, Jacopo Mauro.

The Constraint Programming (CP) paradigm is a general and powerful framework that enables to express relations between different entities in form of constraints that must be satisfied. The concept of constraint is ubiquitous and not confined to the sciences: constraints appear in every aspect of daily life in the form of requirements, obligations, or prohibitions. Historically, the FOCUS group has always had an interest in CP, see e.g., [53], [54]. The possible applications of CP are in fact numerous and disparate. As an example, CP can be used for the deployment of services in the cloud [21], [39].

CP essentially consists of two layers: (i) a modeling level, in which a real-life problem is identified, examined, and formalized into a mathematical model by human experts; (ii) a solving level, aimed at resolving as efficiently and comprehensively as possible the model defined in (i) by means of software agents called constraint solvers. Over the last years we dealt with a particular aspect of CP, that is, the so called portfolio approaches [12], [27], [10]. In a nutshell, a portfolio approach in CP can be seen as the problem of predicting which is (are) the best constraint solver(s) —among a portfolio of available solvers— for solving a given CP problem. A constraint solver that relies on a portfolio of underlying, individual solvers is also dubbed a portfolio solver.

Our studies on portfolio approaches lead to development of the SUNNY-CP portfolio solver [26], [25]. SUNNY-CP relies on underlying state-of-the-art constraint solvers for solving a given CP problem encoded in the MiniZinc language, nowadays a de-facto standard for modeling CP problems. Initially developed as a sequential solver [26], SUNNY-CP has been later on enhanced by enabling the simultaneous execution of its solvers on different cores [25]. This extension allowed SUNNY-CP to win the gold medal in the open track of 2015 MiniZinc Challenge [cite], the annual competition for CP solvers.

However, we did not restrict the work on portfolio approaches to the CP field only. Indeed, we also performed some preliminary studies for evaluating SUNNY (i.e., the algorithm on which SUNNY-CP relies) in other application domains like, e.g., Boolean satisfiability (SAT), Quantified Boolean Formula (QBF), and Answer-Set Programming (ASP) [47], [24].

INDES Project-Team

6. New Results

6.1. Web programming

Participants: Yoann Couillec, Vincent Prunet, Manuel Serrano [correspondant].

6.1.1. Hop.js

Multitier programming languages unify within a single formalism and a single execution environment the programming of the different tiers of distributed applications. On the Web, this programming paradigm unifies the client tier, the server tier, and, when one is used, the database tier. This homogenization offers several advantages over traditional Web programming that rely on different languages and different environments for the two or three tiers of the Web application: programmers have only one language to learn, maintenance and evolution are simplified by the use of a single formalism, global static analyses are doable as a single semantics is involved, debugging and other runtime tools are more powerful as they access global informations about the execution.

The three first multitier platforms for the Web all appeared in 2006: GWT (a.k.a., Google Web Toolkit), Links, and Hop [6], [5]. Each relied on a different programming model and languages. GWT maps the Java programming model on the Web, as it allows, Java/Swing likes programs to be compiled and executed on the Web; Links is functional language with experimental features such as the storing of the whole execution context on the client; Hop is based on the Scheme programming language. These three pioneers have open the path for the other multitier languages such as, Ocsigen for Ocaml, UrWeb, js-scala, etc.

In spite of their interesting properties, multitier languages have not become that popular on the Web. Today, only GWT is widely used in industrial applications but arguably GWT is not a fully multitier language as developing applications with GWT requires explicit JavaScript and HTML programming. This lack of popularity of other systems is likely due to their core based languages than to the programming model itself.

JavaScript is the *defacto* standard on the Web. Since the mid 90's, it is the language of the client-side programming and more recently, with systems like Node.js, it is also a viable solution for the server-side programming. As we are convinced by the virtues of multitier programming we have started a new project consisting of enabling multitier programming JavaScript. We have created a new language called HopScript, which is a minimalist extension of JavaScript for multitier programming, and we have implemented a brand new runtime environment called Hop.js. This environment contains a builtin Web server, on-the-fly HopScript compilers, and many runtime libraries.

HopScript is a super set of JavaScript, *i.e.,* all JavaScript programs are legal HopScript programs. Hop.js is a compliant JavaScript execution environment as it succeeds at 99% of the Ecma 262 tests suite. The Hop.js environment also aims at Node.js compatibility. In its current version it supports about 70% of the Node.js runtime environment. In particular, it fully supports the Node.js modules, which lets Hop programs reuse existing Node.js modules as is.

After a full year of active development to enhance JavaScript and Node.js compatibility, to incorporate features of JavaScript 1.6, and to design new language constructs for machine-to-machine communication, we are now ready to release Hop.js. This will appear at the beginning of 2016.

6.1.2. Data source

During the past few years the volume of accumulated data has increased dramatically. New kinds of data stores have emerged as NoSQL family stores. Many modern applications now collect, analyze, and produce data from several heterogeneous sources. However implementing such applications is still difficult because of lack of appropriate tools and formalisms. We propose a solution to this problem in the context of the JavaScript

programming language by extending array comprehensions. Our extension allows programmers to query data from usual stores, such as SQL databases, NoSQL databases, Semantic Web data repositories, Web pages, or even custom user defined data structures. The extension has been implemented in the Hop.js system. It has been described in the paper [10], which has been presented at the ACM DBPL'15 conference.

6.2. Distributed programming

Participants: Gérard Boudol, Johan Grande, Manuel Serrano [correspondant].

Shared-memory concurrency is a classic concurrency model which, among other things, makes it possible to take advantage of multicore processors that are now widespread in personal computers. Concurrent programs are prone to deadlocks which are notoriously hard to predict and debug. Programs using mutexes, a very popular synchronization mechanism, are no exception.

We have studied deadlock avoidance methods with the aim of making programming with mutexes easier. We first studied a method that uses a static analysis by means of a type and effect system, then a variation on this method in a dynamically typed language.

We developed more the second method. It mixes deadlock prevention and avoidance to provide an easy-to-use and expressive deadlock-free locking function. We implemented it as a Hop library. This lead us to develop a starvation-free algorithm to simultaneously acquire an arbitrary number of mutexes, and to identify the concept of asymptotic deadlock. While doing so, we also developed an optimization of exceptions (finally blocks).

Our performance tests seem to show that using our library has negligible impact on the performance of real-life applications. Most of our work could be applied to other structured programming languages such as Java.

This work has been presented at the 17th International Symposium on Principles and Practice of Declarative Programming (PPDP'15) [13]. More details can be found in Grande's PhD thesis [8].

6.3. Types

Participants: Ilaria Castellani, Bernard Serpette.

6.3.1. Behavioural Types

The survey paper https://hal.inria.fr/hal-01213201 presents a state-of-the-art of a recent trend of research on the use of behavioural types for specifying and analysing security properties of communication-centred systems. It is essentially an outcome of the working group on security of the BETTY COST Action, and it offers a unified overview of various proposals that have been put forward in the last few years, both within the BETTY community and outside it, to combine security analysis with behavioural types.

6.3.2. Abstract Rewriting Systems

We have formalised, with the Coq system, the beginning of Paul-André Melliès's thesis concerning abstract rewriting systems. Behind the interest of studying rewriting systems, which are the roots of all small step semantics of programming languages, this particular formalisation was attractive since it gives a concrete example where we have to manage dependant types.

This was done in collaboration with Eduardo Bonelli and Pablo Barenbaum of University of Quilmes, Argentina. The specification and the proofs of this work take 2200 lines of Coq.

6.4. Security

Participants: Ilaria Castellani, Francis Doliere Some, Nataliia Bielova, Bernard Serpette, Tamara Rezk [correspondant].

6.4.1. Hybrid Typing of Secure Information Flow in a JavaScript-like Language

We propose a novel type system for securing information flow in a core of JavaScript. This core takes into account the defining features of the language, such as prototypical inheritance, extensible objects, and constructs that check the existence of object properties. We design a hybrid version of the proposed type system. This version infers a set of assertions under which a program can be securely accepted and instruments it so as to dynamically check whether these assertions hold. By deferring rejection to runtime, the hybrid version can typecheck secure programs that purely static type systems cannot accept.

This work has been published at the 10th International Symposium on Trustworthy Global Computing [11].

6.4.2. Modular Monitor Extensions for Information Flow Security in JavaScript

Client-side JavaScript programs often interact with the web page into which they are included, as well as with the browser itself, through APIs such as the DOM API, the XMLHttpRequest API, and the W3C Geolocation API. Precise reasoning about JavaScript security must therefore take API invocation into account. However, the continuous emergence of new APIs, and the heterogeneity of their forms and features, renders API behavior a moving target that is particularly hard to capture. To tackle this problem, we propose a methodology for modularly extending sound JavaScript information flow monitors with a generic API. Hence, to verify whether an extended monitor complies with the proposed noninterference property, our methodology requires only to prove that the API satisfies a predefined set of conditions. In order to illustrate the practicality of our methodology, we show how an information flow monitor-inlining compiler can take into account the invocation of arbitrary APIs, without changing the code or the proofs of the original compiler. We provide an implementation of such a compiler with an extension for handling a fragment of the DOM Core Level 1 API. Furthermore, our implementation supports the addition of monitor extensions for new APIs at runtime. This work has been published at the 10th International Symposium on Trustworthy Global Computing [12].

6.4.3. Relaxed Noninterference

We have began a study concerning the use of gradual typing for down casting or declassification for information flow. The particularity of this work is to use a finite state machine to gradually accept the down casting process.

This work is done with Éric Tanter of University of Santiago de Chile, in the context of the project Conicyt Redes CEV Challenges on Electronic Voting.

6.4.4. Hybrid Monitoring of Attacker knowledge

Enforcement of non-interference requires to prove that an attacker's knowledge about the initial state remains the same after observing a programs public output. We define a powerful hybrid monitoring mechanism which evaluates dynamically the knowledge that is contained in program variables. To get a precise estimate of the knowledge, the monitor statically analyses non-executed branches. We show that our knowledge-based approach can be combined with existing dynamic monitors for non-interference. A distinguishing feature of such a combination is that the combined monitor is provably more powerful than each mechanism taken separately. We demonstrate this by proposing a knowledge-enhanced version of a dynamic monitor based on the no-sensitive-upgrade principle. We show how to use the knowledge computed by our hybrid monitor to quantify information leakage associated to the program output. The monitor and its static analysis has been formalized and proved correct within the Coq proof assistant.

6.4.5. A Taxonomy of Information Flow Monitors

We propose a rigorous comparison of information flow monitors with respect to two dimensions: soundness and transparency.

For soundness, we notice that the standard information flow security definition called *Termination-Insensitive Non-interference (TINI)* allows the presence of termination channels, however it does not describe whether the termination channel was present in the original program, or it was added by a monitor. We propose a stronger notion of noninterference, that we call *Termination-Aware Non-interference (TANI)*, that captures this fact, and thus allows us to better evaluate the security guarantees of different monitors. We further investigate TANI, and state its formal relations to other soundness guarantees of information flow monitors. For transparency, we identify different notions from the literature that aim at comparing the behaviour of monitors. We notice that one common notion used in the literature is not adequate since it identifies as better a monitor that accepts insecure executions, and hence may augment the knowledge of the attacker. To discriminate between monitors' behaviours on secure and insecure executions, we factorized two notions that we call true and false transparency. These notions allow us to compare monitors that were deemed to be incomparable in the past.

We analyse five widely explored information flow monitors: no-sensitive- upgrade (NSU), permissive-upgrade (PU), hybrid monitoring (HM), se- cure multi-execution (SME), and multiple facets (MF).

This work has been accepted for publication in the International Conference on Principles of Security and Trust (POST 2016).

6.4.6. A Study of JavaScript constructs used in Top Alexa Sites

Several works on JavaScript analysis have shown that including remote scripts can introduce severe security implications in the behavior of the whole web application. To deal with different kinds of attacks, a number of research groups are developing automatic tools to analyze JavaScript programs. However, most of these works rely on one assumption: the scripts are written in a subset of JavaScript language meaning that only certain constructs are used (that are easier to analyse automatically) and others are omitted (for example, eval is impossible to analyze statically). The goal of the internship was to account for the use of each JavaScript construct in real world programs. To achieve that, we first did a large-scale crawl of the top 10,000 Alexa sites, collecting both inlined scripts and remote scripts. Second, we established the popularity of remote scripts. Next, we accounted for the occurrence of JavaScript constructs in the collected programs. Finally, we use the occurrence of different constructs as basis to propose a subset of JavaScript language, which covers most of JavaScript programs found in the wild. One can rely on this evidence-based subset of JavaScript in future works on that language.

PHOENIX Project-Team

7. New Results

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

Including children with Autism Spectrum Disorders (ASD) in mainstreamed environments creates a need for new interventions whose efficacy must be assessed in situ. We present a tablet-based application for activity schedules that has been designed following a participatory design approach involving mainstream teachers, special-education teachers and school aides. This applications addresses two domains of activities: classroom routines and verbal communications. We assessed the efficiency of our application with two overlapping user-studies in mainstream inclusion, sharing a group of children with ASD. The first experiment involved 10 children with ASD, where 5 children were equipped with our tabled-based application and 5 were not equipped. We show that (1) the use of the application is rapidly self-initiated (after two months for almost all the participants) and that (2) the tablet-supported routines are better performed after three months of intervention. The second experiment involved 10 children equipped with our application; it shared the data collected for the 5 children with ASD and compared them with data collected for 5 children with Intellectual Disabilities – ID. We show that (1) children with ID are not autonomous in the use of the application at the end of the intervention; (2) both groups exhibited the same benefits on classroom routines; and, (3) children with ID improve significantly less their performance on verbal communication routines. These results are discussed in relation with our design principles. Importantly, the inclusion of a group with another neurodevelopmen- tal condition provided insights about the applicability of these principles beyond the target population of children with ASD.

7.2. Age and active navigation effects on episodic memory: A virtual reality study

We investigated the navigation-related age effects on learning, proactive interference semantic clustering, recognition hits, and false recognitions in a naturalistic situation using a virtual apartment-based task. We also examined the neuropsychological correlates (executive functioning [EF] and episodic memory) of navigation-related age effects on memory. Younger and older adults either actively navigated or passively followed the computer-guided tour of an apartment. The results indicated that active navigation increased recognition hits compared with passive navigation, but it did not influence other memory measures (learning, proactive interference, and semantic clustering) to a similar extent in either age group. Furthermore, active navigation helped to reduce false recognitions in younger adults but increased those made by older adults. This differential effect of active navigation for younger and older adults was accounted for by EF score. Like for the subject-performed task effects, the effects from the navigation manipulation were well accounted for by item-specific/relational processing distinction, and they were also consistent with a source monitoring deficit in older adults.

7.3. Constraining application behaviour by generating languages

Writing a platform for reactive applications which enforces operational constraints is difficult, and has been approached in various ways. In this experience report, we detail an approach using an embedded DSL which can be used to specify the structure and permissions of a program in a given application domain. Once the developer has specified which components an application will consist of, and which permissions each one needs, the specification itself evaluates to a new, tailored, language. The final implementation of the application is then written in this specialised environment where precisely the API calls associated with the permissions which have been granted, are made available. Our prototype platform targets the domain of mobile computing ,

and is implemented using Racket. It demonstrates resource access control (e.g., camera, address book, etc.) and tries to prevent leaking of private data. Racket is shown to be an extremely effective platform for designing new programming languages and their run-time libraries. We demonstrate that this approach allows reuse of an inter-component communication layer, is convenient for the application developer because it provides high-level building blocks to structure the application, and provides increased control to the platform owner, preventing certain classes of errors by the developer.

7.4. A Unifying Notification System To Scale Up Assistive Services

Aging creates needs for assistive technology to support all activities of daily living (meal preparation, dressing, social participation, stove monitoring, *etc.*). These needs are mostly addressed by a silo-based approach that requires a new assistive service (*e.g.*, a reminder system, a pill prompter) to be acquired for every activity to be supported. In practice, these services manifest their silo-based nature in their user interactions , and more specifically, in the heterogeneity of their notification system. This heterogeneity incurs a cognitive cost that prevents scaling up assistive services and compromises adoption by older adults. We present an approach to scaling up the combination of technology-based, assistive services by proposing a unifying notification system. To do so, (1) we propose a decomposition of assistive services to expose their needs in notification; (2) we introduce a notification framework, allowing heterogeneous assistive services to homogeneously notify users; (3) we present how this notification framework is carried out in practice for an assisted living platform. We successfully applied our approach to a range of existing and new assistive services. We used our notification framework to implement an assistive platform that combines a variety of assistive services. This platform has been deployed and used 24/7 in the home of 15 older adults for up to 6 months. This study provides empirical evidence of the effectiveness and learnability of the notification system of our platform, irrespective of the cognitive and sensory resources of the user. Additional results show that our assisted living platform achieved high user acceptance and satisfaction.

7.5. Orchestrating Masses of Sensors: A Design-Driven Development Approach

We propose a design-driven development approach that is dedicated to the domain of orchestration of masses of sensors. The developer declares what an application does using a domain-specific language (DSL). Our compiler processes domain-specific declarations to generate a customized programming framework that guides and supports the programming phase.

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

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

RMOD Project-Team

7. New Results

7.1. Tools for understanding evolution

Automatic Detection of System-Specific Conventions. In Apache Ant, a convention to improve maintenance was introduced in 2004 stating a new way to close files instead of the Java generic InputStream.close(). Yet, six years after its introduction, this convention was still not generally known to the developers. Two existing solutions could help in these cases. First, one can deprecate entities, but, in our example, one can hardly deprecate Java's method. Second, one can create a system-specific rule to be automatically enforced. In a preceding publication , we showed that system-specific rules are more likely to be noticed by developers than generic ones. However, in practice, developers rarely create specific rules. We therefore propose to free the developers from the need to create rules by automatically detecting such conventions from source code repositories. This is done by mining the change history of the system to discover similar changes being applied over several revisions. The proposed approach is applied to real-world systems, and the extracted rules are validated with the help of experts. The results show that many rules are in fact relevant for the experts. [16]

DeltaImpactFinder. In software development, version control systems (VCS) provide branching and merging support tools. Such tools are popular among developers to concurrently change a code-base in separate lines and reconcile their changes automatically afterwards. However, two changes that are correct independently can introduce bugs when merged together. We call semantic merge conflicts this kind of bugs. Change impact analysis (CIA) aims at estimating the effects of a change in a codebase. We propose to detect semantic merge conflicts using CIA. On a merge, DELTAIMPACTFINDER analyzes and compares the impact of a change in its origin and destination branches. We call the difference between these two impacts the delta-impact. If the delta-impact is empty, then there is no indicator of a semantic merge conflict and the merge can continue automatically. Otherwise, the delta-impact contains what are the sources of possible conflicts. [26]

OrionPlanning. Many techniques have been proposed in the literature to support architecture definition, conformance, and analysis. However, there is a lack of adoption of such techniques by the industry. Previous work have analyzed this poor support. Specifically, former approaches lack proper analysis techniques (e.g., detection of architectural inconsistencies), and they do not provide extension and addition of new features. We present ORIONPLANNING, a prototype tool to assist refactorings at large scale. The tool provides support for model-based refactoring operations. These operations are performed in an interactive visualization. The contributions of the tool consist in: (i) providing iterative modifications in the architecture, and (ii) providing an environment for architecture inspection and definition of dependency rules. [37]

Recording and Replaying System-Specific Conventions. During its lifetime, a software system is under continuous maintenance to remain useful. Maintenance can be achieved in activities such as adding new features, fixing bugs, improving the system's structure, or adapting to new APIs. In such cases, developers sometimes perform sequences of code changes in a systematic way. These sequences consist of small code changes (e.g., create a class, then extract a method to this class), which are applied to groups of related code entities (e.g., some of the methods of a class). MacroRecorder is a proof-of-concept tool that records a sequence of code changes, then it allows the developer to generalize this sequence in order to apply it in other code locations. The evaluation is based on previous work on repetitive code changes related to rearchitecting. MacroRecorder was able to replay 92% of the examples, which consisted in up to seven code entities modified up to 66 times. The generation of a customizable, large-scale transformation operator has the potential to efficiently assist code maintenance. [39], [38]

7.2. Software Quality: Taming Software Evolution

Software metrics do not predict the health of a project. More and more companies would like to mine software data with the goal of assessing the health of their software projects. The hope is that some software

metrics could be tracked to predict failure risks or confirm good health. If a factor of success was found, projects failures could be anticipated and early actions could be taken by the organisation to help or to monitor closely the project, allowing one to act in a preventive mode rather than a curative one. We were called by a major IT company to fulfil this goal. We conducted a study to check whether software metrics can be related to project failure. The study was both theoretic with a review of literature on the subject, and practical with mining past projects data and interviews with project managers. We found that metrics used in practice are not reliable to assess project outcome. [22]

How Do Developers React to API Evolution? Software engineering research now considers that no system is an island, but it is part of an ecosystem involving other systems, developers, users, hardware,. .. When one system (e.g., a framework) evolves, its clients often need to adapt. Client developers might need to adapt to functionalities, client systems might need to be adapted to a new API, client users might need to adapt to a new User Interface. The consequences of such changes are yet unclear, what proportion of the ecosystem might be expected to react, how long might it take for a change to diffuse in the ecosystem, do all clients react in the same way? We report on an exploratory study aimed at observing API evolution and its impact on a large-scale software ecosystem, Pharo, which has about 3,600 distinct systems, more than 2,800 contributors, and six years of evolution. We analyze 118 API changes and answer research questions regarding the magnitude, duration, extension, and consistency of such changes in the ecosystem. The results of this study help to characterize the impact of API evolution in large software ecosystems, and provide the basis to better understand how such impact can be alleviated. [27]

Does JavaScript software embrace classes? JavaScript is the de facto programming language for the Web. It is used to implement mail clients, office applications, or IDEs, that can weight hundreds of thousands of lines of code. The language itself is prototype based, but to master the complexity of their application, practitioners commonly rely on some informal class abstractions. This practice has never been the target of empirical investigations in JavaScript. Yet, understanding it would be key to adequately tune programming environments and structure libraries such as they are accessible to programmers. We report a large and in-depth study to understand how class emulation is employed in JavaScript applications. We propose a strategy to statically detect class-based abstractions in the source code of JavaScript systems. We used this strategy in a dataset of 50 popular JavaScript applications available from GitHub. We found systems structured around hundreds of classes, suggesting that JavaScript developers are standing on traditional class-based abstractions to tackle the growing complexity of their systems. [28]

7.3. Software Quality: History and Changes

Mining Architectural Violations from Version History. Software architecture conformance is a key software quality control activity that aims to reveal the progressive gap normally observed between concrete and planned software architectures. However, formally specifying an architecture can be difficult, as it must be done by an expert of the system having a high level understanding of it. We present a lightweighted approach for architecture conformance based on a combination of static and historical source code analysis. The proposed approach relies on four heuristics for detecting absences (something expected was not found) and divergences (something prohibited was found) in source code based architectures. We also present an architecture conformance process based on the proposed approach. We followed this process to evaluate the architecture of two industrial-strength information systems, achieving an overall precision of 62.7% and 53.8%. We also evaluated our approach in an open-source information retrieval library, achieving an overall precision of 59.2%. We envision that an heuristic-based approach for architecture conformance can be used to rapidly raise architectural warnings, without deeply involving experts in the process. [17]

Untangling Fine-Grained Code Changes. After working for some time, developers commit their code changes to a version control system. When doing so, they often bundle unrelated changes (e.g., bug fix and refactoring) in a single commit, thus creating a so-called tangled commit. Sharing tangled commits is problematic because it makes review, reversion, and integration of these commits harder and historical analyses of the project less reliable. Researchers have worked at untangling existing commits, i.e., finding which part of a commit relates to which task. We contribute to this line of work in two ways: (1) A publicly available

dataset of untangled code changes, created with the help of two developers who accurately split their code changes into self contained tasks over a period of four months; (2) a novel approach, EpiceaUntangler, to help developers share untangled commits (aka. atomic commits) by using fine-grained code change information. EpiceaUntangler is based and tested on the publicly available dataset, and further evaluated by deploying it to 7 developers, who used it for 2 weeks. We recorded a median success rate of 91% and average one of 75%, in automatically creating clusters of untangled fine-grained code changes. [25]

Developers' Perception of Co-Change Patterns: An Empirical Study. Co-change clusters are groups of classes that frequently change together. They are proposed as an alternative modular view, which can be used to assess the traditional decomposition of systems in packages. To investigate developer's perception of co-change clusters, we report a study with experts on six systems, implemented in two languages. We mine 102 co-change clusters from the version history of such systems, which are classified in three patterns regarding their projection to the package structure: Encapsulated, Crosscutting, and Octopus. We then collect the perception of expert developers on such clusters, aiming to ask two central questions: (a) what concerns and changes are captured by the extracted clusters? (b) do the extracted clusters reveal design anomalies? We conclude that Encapsulated Clusters are often viewed as healthy designs and that Crosscutting Clusters tend to be associated to design anomalies. Octopus Clusters are normally associated to expected class distributions, which are not easy to implement in an encapsulated way, according to the interviewed developers. [40]

7.4. Dynamic Languages: Debugging

Practical domain-specific debuggers. Understanding the run-time behavior of software systems can be a challenging activity. Debuggers are an essential category of tools used for this purpose as they give developers direct access to the running systems. Nevertheless, traditional debuggers rely on generic mechanisms to introspect and interact with the running systems, while developers reason about and formulate domain-specific questions using concepts and abstractions from their application domains. This mismatch creates an abstraction gap between the debugging needs and the debugging support leading to an inefficient and error-prone debugging effort, as developers need to recover concrete domain concepts using generic mechanisms. To reduce this gap, and increase the efficiency of the debugging process, we propose a framework for developing domain-specific debuggers, called the Moldable Debugger, that enables debugging at the level of the application domain. The Moldable Debugger is adapted to a domain by creating and combining domain-specific debugging operations with domain-specific debugging views, and adapts itself to a domain by selecting, at run time, appropriate debugging operations and views. To ensure the proposed model has practical applicability (i.e., can be used in practice to build real debuggers), we discuss, from both a performance and usability point of view, three implementation strategies. We further motivate the need for domain-specific debugging, identify a set of key requirements and show how our approach improves debugging by adapting the debugger to several domains. [14]

Mercury: Properties and Design of a Remote Debugging Solution using Reflection. Remote debugging facilities are a technical necessity for devices that lack appropriate input/output interfaces (display, keyboard, mouse) for programming (e.g., smartphones, mobile robots) or are simply unreachable for local development (e.g., cloud-servers). Yet remote debugging solutions can prove awkward to use due to re-deployments. Empirical studies show us that on average 10.5 minutes per coding hour (over five 40-hour work weeks per year) are spent for redeploying applications (including re-deployments during debugging). Moreover current solutions lack facilities that would otherwise be available in a local setting because it is difficult to reproduce them remotely. Our work identifies three desirable properties that a remote debugging solution should exhibit, namely: run-time evolution, semantic instrumentation and adaptable distribution. Given these properties we propose and validate Mercury, a remote debugging model based on reflection. Mercury supports run-time evolution through a causally connected remote meta-level, semantic instrumentation through the reification of the underlying execution environment and adaptable distribution through a modular architecture of the debugging middleware. [19]

7.5. Reconciling Dynamic Languages and Isolation

Handles. Controlling object graphs and giving specific semantics to references (such as read-only, ownership, scoped sharing) have been the focus of a large body of research in the context of static type systems. Controlling references to single objects and to graphs of objects is essential to build more secure systems, but is notoriously hard to achieve in the absence of static type systems. In this article we embrace this challenge by proposing a solution to the following question: What is an underlying mechanism that can support the definition of properties (such as revocable, read-only, lent) at the reference level in the absence of a static type system? We present handles: first-class references that propagate behavioral change dynamically to the object subgraph during program execution. In this article we describe handles and show how handles support the implementation of read-only references and revocable references. Handles have been fully implemented by modifying an existing virtual machine and we report their costs. [13]

Delegation Proxies. Scoping behavioral variations to dynamic extents is useful to support non-functional concerns that otherwise result in cross-cutting code. Unfortunately, such forms of scoping are difficult to obtain with traditional reflection or aspects. We propose delegation proxies, a dynamic proxy model that supports behavioral intercession through the interception of various interpretation operations. Delegation proxies permit different behavioral variations to be easily composed together. We show how delegation proxies enable behavioral variations that can propagate to dynamic extents. We demonstrate our approach with examples of behavioral variations scoped to dynamic extents that help simplify code related to safety, reliability, and monitoring. [21]

Access Control to Reflection with Object Ownership. Reflection is a powerful programming language feature that enables language extensions, generic code, dynamic analyses , development tools, etc. However, uncontrolled reflection breaks object encapsulation and considerably increases the attack surface of programs e.g., malicious libraries can use reflection to attack their client applications. To bring reflection and object encapsulation back together, we use dynamic object ownership to design an access control policy to reflective operations. This policy grants objects full reflective power over the objects they own but limited reflective power over other objects. Code is still able to use advanced reflective operations but reflection cannot be used as an attack vector anymore. [41]

7.6. Tailoring Applications and boostrapping

Virtualization Support for Dynamic Core Library Update. Dynamically updating language runtime and core libraries such as collections and threading is challenging since the update mechanism uses such libraries at the same time that it modifies them. To tackle this challenge, we present Dynamic Core Library Update (DCU) as an extension of Dynamic Software Update (DSU) and our approach based on a virtualization architecture. Our solution supports the update of core libraries as any other normal library, avoiding the circular dependencies between the updater and the core libraries. Our benchmarks show that there is no evident performance overhead in comparison with a default execution. Finally, we show that our approach can be applied to real life scenario by introducing a critical update inside a web application with 20 simulated concurrent users. [34]

Bootstrapping Infrastructure. Bootstrapping is well known in the context of compilers, where a bootstrapped compiler can compile its own source code. Bootstrapping is a beneficial engineering practice because it raises the level of abstraction of a program making it easier to understand, optimize, evolve, etc. Bootstrapping a reflective object-oriented language is however more challenging , as we need also to initialize the runtime of the language with its initial objects and classes besides writing its compiler. We present a novel bootstrapping infrastructure for Pharo-like languages that allows us to easily extend and modify such languages. Our bootstrapping process relies on a first-class runtime. A first-class runtime is a meta-object that represents a program's runtime and provides a MOP to easily load code into it and manipulate its objects. It decou-ples the virtual machine (VM) and language concerns by introducing a clear VM-language interface. Using this process, we show how we succeeded to bootstrap a Smalltalk-based language named Candle and then extend it with traits in less than 250 lines of high-level Smalltalk code. We also show how we can bootstrap with minimal effort two other languages (Pharo and MetaTalk) with similar execution semantics but different object models. [35]

7.7. Dynamic Languages: Virtual Machines

Towards Fully Reflective Environments. Modern development environments promote live programming (LP) mechanisms because it enhances the development experience by providing instantaneous feedback and interaction with live objects. LP is typically supported with advanced reflective techniques within dynamic languages. These languages run on top of Virtual Machines (VMs) that are built in a static manner so that most of their components are bound at compile time. As a consequence, VM developers are forced to work using the traditional edit-compile-run cycle, even when they are designing LP-supporting environments. We explore the idea of bringing LP techniques to VM development to improve the observabil-ity, evolution and adaptability of VMs at run-time. We define the notion of fully reflective execution environments , systems that provide reflection not only at the application level but also at the level of the execution environment (EE). We characterize such systems, propose a design, and present Mate v1, a prototypical implementation. Based on our prototype, we analyze the feasibility and applicability of incorporating reflective capabilities into different parts of EEs. Furthermore, the evaluation demonstrates the opportunities such reflective capabilities provide for unanticipated dynamic adaptation scenarios, benefiting thus, a wider range of users. [23]

Tracing vs. Partial Evaluation. Tracing and partial evaluation have been proposed as meta-compilation techniques for interpreters to make just-in-time compilation language-independent. They promise that programs executing on simple interpreters can reach performance of the same order of magnitude as if they would be executed on state-of-the-art virtual machines with highly optimizing just-in-time compilers built for a specific language. Tracing and partial evaluation approach this meta-compilation from two ends of a spectrum, resulting in different sets of tradeoffs. This study investigates both approaches in the context of self-optimizing interpreters, a technique for building fast abstract-syntax-tree interpreters. Based on RPython for tracing and Truffle for partial evaluation, we assess the two approaches by comparing the impact of various optimizations on the performance of an interpreter for SOM, an object-oriented dynamically-typed language. The goal is to determine whether either approach yields clear performance or engineering benefits. We find that tracing and partial evaluation both reach roughly the same level of performance. SOM based on meta-tracing is on average 3x slower than Java, while SOM based on partial evaluation is on average 2.3x slower than Java. With respect to the engineering, tracing has however significant benefits, because it requires language implementers to apply fewer optimizations to reach the same level of performance. [29]

Zero-Overhead Metaprogramming. Runtime metaprogramming enables many useful applications and is often a convenient solution to solve problems in a generic way, which makes it widely used in frameworks, middleware, and domain-specific languages. However, powerful metaobject protocols are rarely supported and even common concepts such as reflective method invocation or dynamic proxies are not optimized. Solutions proposed in literature either restrict the metaprogramming capabilities or require application or library developers to apply performance improving techniques. For overhead-free runtime metaprogramming, we demonstrate that dispatch chains, a generalized form of polymorphic inline caches common to self-optimizing interpreters, are a simple optimization at the language-implementation level. Our evaluation with self-optimizing interpreters shows that unrestricted metaobject protocols can be realized for the first time without runtime overhead, and that this optimization is applicable for just-in-time compilation of interpreters based on meta-tracing as well as partial evaluation. In this context, we also demonstrate that optimizing common reflective operations can lead to significant performance improvements for existing applications [30].

A Partial Read Barrier for Efficient Support of Live Object-oriented Programming. Live programming, originally introduced by Smalltalk and Lisp, and now gaining popularity in contemporary systems such as Swift, requires on-the-fly support for object schema migration, such that the layout of objects may be changed while the program is at one and the same time being run and developed. In Smalltalk schema migration is supported by two primitives, one that answers a collection of all instances of a class, and one that exchanges the identities of pairs of objects, called the become primitive. Existing instances are collected, copies using the new schema created, state copied from old to new, and the two exchanged with become, effecting the schema migration. Historically the implementation of become has either required an extra level of indirection between an object's address and its body, slowing down slot access, or has required a sweep of all objects, a very slow operation on large heaps. Spur, a new object representation and memory manager for Smalltalk-like

languages, has neither of these deficiencies. It uses direct pointers but still provides a fast become operation in large heaps, thanks to forwarding objects that when read conceptually answer another object and a partial read barrier that avoids the cost of explicitly checking for forwarding objects on the vast majority of object accesses [31].

TACOMA Team

6. New Results

6.1. Self-describing objects and tangible data structures

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

A development in the line of the composite objects (see section 3.3) are self-describing objects. While previous works enabled integrity checking over a set of physical objects, these mechanisms were limited in two aspects: expressiveness and autonomy. More precisely, objects support the detection of special conditions (such as a missing element), but not the characterization of these conditions (such as describing the problem, identifying the missing element). Moreover, this compromises the autonomous feature of coupled objects, which would depend on external systems for analysing these special conditions. Self-describing objects are an attempt to overcome these limitations, and to broaden the application perspectives of autonomous RFID systems.

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

However, given the extreme storage limitation of RFID chips, designing such systems is difficult:

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

In the context of RFID system, the resiliency property of such data structures enables new information architecture and autonomous (offline) operation, which is very important for some RFID applications. We previously applied the self-describing objects approach to the waste management domain [1], which has shown to be a specially challenging situation for RFID. This challenge is found more generally in pervasive computing scenarios involving RFID reading in uncontrolled environments (see section 4.3).

Pervasive support for RFIDs.

We propose to apply our approach to improve the robustness of RFID inventories / batch checking: when many objects are read at once by an RFID reader, miss read are common and raise reliability and operational issues for applications. An innovative solution to this problem is to take advantage of the multiplicity of tags by leveraging them as a distributed memory shared by a logical group. In this way, it is possible to support error detection as well as information recovery. We proposed a flexible protocol to support robust EPC retrieval in adverse reading conditions. The proposed protocol uses erasure correcting techniques to enable error-free recovery of misread EPCs [2]. It is further customizable with respect to the rate of misread tags and application requirements. This work was the object of an Inria patent [0]. Fine-tuning the protocol parameters is still the object of on going experimentation in the context of the Pervasive_RFID project (see section 7.1.1).

[0] Patent filed in April 2015 - Inria 179

At the software level, RFID inventory reliability issue is usually addressed by anti-collisions mechanisms and redundancy mechanisms. Anti-collisions protocols limit the risk of data corruption when multiples tags have to reply to an inventory request. Redundancy is often implemented in RFID readers by aggregating the results of multiple inventory requests over a time frame, to give the tags multiple opportunities to reply. While useful, these strategies cannot ensure that a given inventory is valid or not (in other words, one or more tags may be missing without being noticed). In situations where we have to read large collection of objects of various types, the performance is difficult to predict but may still be adequate for a given application. For example, some application can tolerate missing some tags, provided that miss read probability could be characterized. In some cases, read reliability could be improved using mechanical approaches, such as introducing movements in objects or antenna to introduce *radio diversity* during read. Finally, distributed data structure can be used over a set of tags to be used to mitigate the impact of misread (by using data redundancy) and to help the reading protocol by integrating hints about the tag set collection being read.

We studied extensively by experimentation the behaviour of existing RFID solutions in the context of uncontrolled environment (meaning, random placement of tags on objects mixing various materials) in order to characterize their real-world performance regarding the parameters of such as tags numbers, density, frequencies, reader antenna design, dynamicity of objects (movements), etc. From these experimentations, we would like to identify the conditions that are favorable to acceptable performance, and the way where there are hopes of improvement with specific design for these difficult environments. These results should also allow improving the performance: high level integrity checks can guide low level operations by determining whether inventories are complete or not. This cross layer strategy enables faster and more efficient inventory protocols.

6.2. Interactions between connected objects in a Smart Building

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

Tacoma group is focussed on the conception and implementation of innovative services for the Smart Home/Building. The range of considered services is broad: from "optimizing the energy consumption" to "helping users to find their way in a building". One of our goals is to build a pervasive platform with constrained performance and cost [7], without disrupting existing spaces. Within this idea, we explored in 2015 the services provided by different modes of interaction in a physical space between neighboring objects, and also between an object and a nearby user.

More precisely, we conducted some experiments with LEDs. Instrumented via a short distance radio interface, a lighting device becomes an unobtrusive connected object that is easy to integrate to a mesh network. A relevant aspect of this platform is the consideration of potential conflicts in data access offered by the connected objects. One of the first scenarios we considered is to operate an LED-based light path to guide the evacuation of a building in the case of a fire alarm. When our objective is to multiply the uses of LED devices ("go beyond lighting", see section 7.1.2), the question is then the priority of access to resources offered by the platform distributed in the environment. Specifically, we addressed the following issues (similar to some of the issues presented in section 3.2):

- How to prioritize the lighting functions (classic) and occasional (but priority) uses of the LED to help in the care of a fire alarm?
- How do you prioritize access to the objects and/or resources that carry these items?

6.3. Context computing for Smart Home

Participants: Yoann Maurel, Frédéric Weis [contact].

To provide services for smart Homes, automation based on pre-set scenarios is ineffective: human behavior is hardly predictable and application should be able to adapt their behavior at runtime depending on the context. We focused on recognizing user's activities to adapt applications behaviours. Our aim is to compute small pieces of context we called *context attributes*. Those context attributes are diverse, for example a presence in a room, the number of people in a room etc.

Building efficient and accurate context information using inexpensive and non-invasive sensors was and is still a great challenge 3.1 . We proved, through the use of dedicated algorithms and a layered architecture that it is achievable when the targeted Home is known - due to the specific and non automated calibration process we used. Among all the available theories, we used the Belief Function Theory (BFT) [8] [9] as it allows to express **uncertainty** and **imprecision**.

Context is computed by a chain a tasks as illustrated in figure 5 :

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

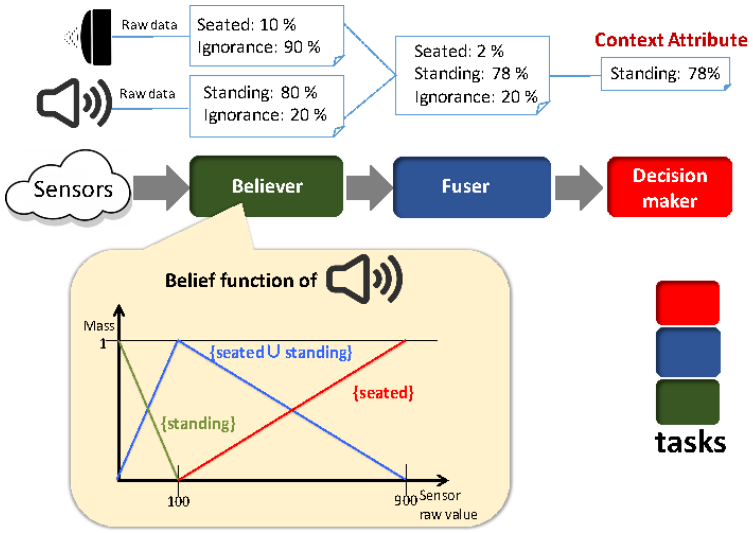

Figure 5. Context attribute computation

Alleviating the complexity of the platform configuration and maintenance is a prerequisite for the adoption of Smart-Home environments by consumers. Currently the BFT theories requires a huge calibration process. We focussed our efforts on the semi-automated building of belief functions, required by the theory, that have to be provided by each sensor.

Automated configuration of sensors.
The belief model is provided to the platform by us and a component is in charge of transforming a sensor value in a *belief function*. The fine tuning of a belief function can be a tedious task. It must be done by a specialist who understands the belief function theory and knows the behavior of the sensors. The model is often built iteratively by experimenting. This may take several hours or days. Moreover, this method is directly connected to the output of each sensor. Biased and noisy measures can cause major modifications on the resulting beliefs.

Ideally, the calibration of the model should be as automatic as possible (few interaction with the user during calibration). The person setting up the sensors should not have to understand the belief function theory. We proposed to generate our belief model from a training set of sensor data. We mainly focused on k-nearest neighbors (KNN) algorithm [6]. We used a training data set to compute the presence belief model. We acquired a set of data with someone present in the experimentation room and a second data set with nobody in the room, which gives us a labelled data set. This principle is illustrated in figure 6 .

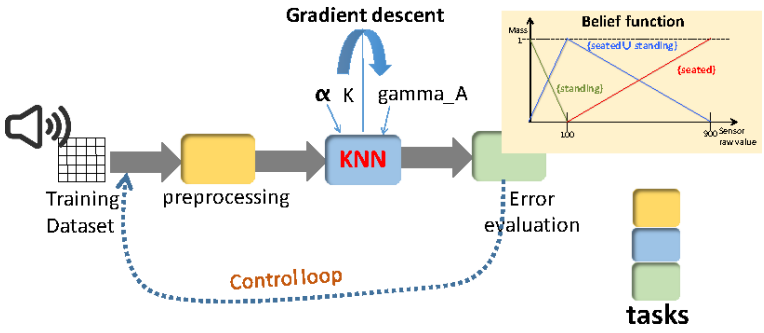

Figure 6. Sensor autocalibration

6.4. Design of a framework for distributed pervasive environments

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

Pervasive environment brings into play complex interactions between a large number of heterogeneous entities: computing units executing third-party applications delivering multiple services to users, with various (sometimes conflicting) requirements, based on the information provided by dynamically (un)available smart object or sensors. The development of pervasive application is consequently hard and must be supported by architectures and frameworks that propose solution to manage the heterogeneity, to organize the interaction of distributed entities, to support the dynamic discovery of the entities, to ensure the privacy of collected data and inferred context, to organize and structure information sharing, and to enforce access control over data and entities.

To alleviate the development of such application (see section 3.4), we worked on a distributed pervasive environment made of several processing nodes (or gateway) managing interacting Smart Spaces (*i.e* a room, a corridor etc.). A Smart Space contains one or more nodes that coordinate to provide services to users. A node is a low cost computing unit with constrained performances. Each node is responsible for the management of entities and services available in their close proximity: they dynamically discover available devices and source of information, computes contextual information and offer services to nearby users. Nodes are organized hierarchically: in each space one *supervisor node* is responsible for coordination between nodes (*e.g.* managing conflicting requirements and enforcing global policies) and communication with neighboring Smart Spaces. The whole environment (a Smart Building or a Smart Home) is controlled by a master node that distributes policy between Smart Spaces to provide global services (*e.g.* global energy management). Data are stored and processed by nodes as close as possible to the users in order to enforce data privacy (see section 2.1).

Each node supports the execution of several services and application. To help the development of these services, applications are built on top of a framework (**Matriona**) running on each gateways.

Matriona proposes a unified representation of the concepts manipulated by applications in order to hide the heterogeneity of technologies to the application. We rely on the concept of *resource*. A Matriona's resource is quite similar to a REST resource but is not identical: HTTP protocol is only used for remote call; the structure of a resource is constrained; a resource can be dynamically discovered; it can provide notification (PUSH operation); it is uniquely identified in the environment; it is any object used or shared by applications, by bridges, or by the system itself: a device, a room, a platform, a user or a contextual information; it is implemented as a standard object and has a type specified by its interfaces (annotated Java interface). Types describe the data and the operations on resources. *CRUD* (**C**reate, **R**ead, **U**pdate, **D**elete) and *PUSH* operations are used to represent the semantic of operations. Specifying the semantic of operations allows to operate some

generic processing (*e.g.* gathering all the information provided by a resource) on resources without knowing their types.

Matriona enables dynamic discovery between nodes and inter-platform communication between the applications and resources. The platforms automatically discovers other platforms using a discovery mechanism. Each platform enables the discovery and the use of its resources to other platforms' applications. Remote resources are using exactly the same API as their internal counterparts: using remote resources is completely transparent to the application. The data is serialized / deserialized automatically by the platform during the call. Calling remote resources induces a performance cost equivalent to the use of a traditional REST resource.

Matriona allows to dynamically add new properties and new behaviors to a resource using decorators. Resources are built using multiple layers in the same way as "Russian dolls". Each layer is responsible for implementing a specific behavior such as retrieving, conversion operation or adding new properties (*e.g.* localization). For instance, the implementation of a thermometer resource will consist of 1) a core layer providing standard information (id, date of creation, the groups to which it belongs); 2) a protocol layer able to communicate with the real devices; 3) a conversion layer (Kelvin to Celsius). The application interacts with the top layer that exposes all or part of information and treatments offered by lower layers. While some layers are static (part of the resource declaration) and cannot be removed, most can be added afterward by the applications. It creates a virtual resource composed of the original resource and the new layers. This virtual resource can be used and discovered as any other resources.

Matriona allows to organize the information. Resource may reference other resources, for example the localization of a "thermometer resource" refers to the resource representing the room in which it is located. The value of the property is the *id* of the referenced resource. This allows applications to easily find resources and their interactions. It is also possible to create composite resources using aggregation mechanisms provided by the framework. The virtual resource can be used directly by applications as any other resources.

Matriona provides a basic language queries. Applications use resources directly or send queries to the platform registry. The query language allows to apply filters and to aggregate data on available resources. A request is represented by a specific URL. For instance, the mean temperature of thermometers in the whole meeting rooms of a building can be obtained using the URL `/*/*/$ location/meeting_room/temperature!mean`. The query language can be extended by providing new decorators and new filters.

Matriona provides access management: each resources belongs to one or many groups. The groups are defined when the resource is created or during its lifecycle by the owner of the resource. Groups gather applications that share the same permissions on access resources. Groups are managed by a "group owner" that can limit members permissions. Permissions describe the ability of an application to read, write, update, delete, manage or lock a resource. Resource locking avoids conflicting requests to be performed by different applications. Locks are given to applications for a fixed period of time. A resource can always be unlocked by the platform itself or by "critical" applications (*e.g.* emergency fire alarm, see section 6.2).

Matriona allows applications to extend resource properties and to share these meta-information with others. Each application can add new information to a given resource. Tagged information are available only for the application and its group. Meta-information are stored by the platform and associated to the resource until the latter is destroyed. This mechanism allows application to easily share information on the resource they used. For instance, this can be used to retrieve previously used resources or to rate the quality of service provided by a given resource. For the application, meta-information are part of the resource. It is then possible for an application to only use resources that have been approved by other applications of their group. This mechanism can also be used by application to add some task-relevant information (*e.g.* a medical application can tag resources that have been used by a patient).

6.5. Towards Metamorphic Housing: the on-demand room

Participant: Michele Dominici [contact].

This research activity is supported by Fondation Rennes 1 through the chair "Smart Home and Innovation", since January 2014. This activity is centered on the concept of metamorphic housing (see section 4.2). During the first year, we had identified the goals of the research project, also taking into account the trends of future housing industry, provided by the enterprises and public authorities that support the chair. We also had identified a case study, the on-demand room, to be displayed as the main application of the research results in scientific communications and vulgarization. It consists in a space that is physically shared by a small group of apartments, but is assigned for the sole use of one or few particular ones at the time. The room is designed so as to make occupants feel they did not leave their apartment at all. They seamlessly move from their dwelling to the on-demand room and conversely, without noticing the difference, as the room adapts to their preferences. During 2015, second year of the chair, we organized our work following two main axes: *(i)* solving the research problems, illustrated in the rest of this section; *(ii)* demonstrating the results using mixed reality as combination of virtual reality and off-the-shelf domotic devices, described in section 5.2.2 .

The research problems underlying the on-demand room are numerous: we illustrated them in the research report "A Case Study of Metamorphic Housing: The on-Demand Room" [3]. We started by addressing the problems associated with the goal of "plugging" the room into different apartments. This requires to dynamically change the rights to control and customize the room's equipment, including lights, appliances, heating, ventilation and air conditioning systems (HVAC), etc. This must be done in a transparent fashion, so that off-the-shelf devices and appliances can be used.

To solve these problems, we started a collaboration with the DIVERSE team [0]. The goal is to use the Kevoree [0] software framework to dynamically reconfigure the networks and domotic system of the room and of the apartments. When the on-demand room is owned by an apartment, their computer networks are interconnected; appliances, sensors and controllers in the room and the apartment can communicate with each others; devices reflect user preferences. Kevoree will enable these reconfigurations by running on key appliances and dynamically adapting and customizing their behavior to the owner of the on-demand room.

As part of the collaboration, some research goals have already been identified. The underlying challenges will be addressed and the results will be integrated in a comprehensive mixed reality demonstrator. This will represent the final iteration of the ongoing demonstration process, illustrated in the platform section (for more details, see 5.2.2).

[0] http://diverse.irisa.fr/
[0] http://kevoree.org/

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, Alvinice Kodjo, Fatima Zahra Moataz, Joanna Moulierac, Nicolas Nisse, Stéphane Pérennes.

7.1.1. Wireless Networks

7.1.1.1. Dimensioning Microwave Wireless Networks

In [47], we aim at dimensioning fixed broadband microwave wireless networks under unreliable channel conditions. As the transport capacity of microwave links is prone to variations due to, e.g., weather conditions, such a dimensioning requires special attention. It can be formulated as the determination of the minimum cost bandwidth assignment of the links in the network for which traffic requirements can be met with high probability, while taking into account that transport link capacities vary depending on channel conditions. The proposed optimization model represents a major step forward since we consider dynamic routing. Experimental results show that the resulting solutions can save up to 45% of the bandwidth cost compared to the case where a bandwidth over-provisioning policy is uniformly applied to all links in the network planning. Comparisons with previous work also show that we can solve much larger instances in significantly shorter computing times, with a comparable level of reliability.

7.1.1.2. Data Gathering and Personalized Broadcasting in Radio Grids with Interference

In the gathering problem, a particular node in a graph, the base station, aims at receiving messages from some nodes in the graph. At each step, a node can send one message to one of its neighbors (such an action is called a call). However, a node cannot send and receive a message during the same step. Moreover, the communication is subject to interference constraints, more precisely, two calls interfere in a step, if one sender is at distance at most dI from the other receiver. Given a graph with a base station and a set of nodes having some messages, the goal of the gathering problem is to compute a schedule of calls for the base station to receive all messages as fast as possible, i.e., minimizing the number of steps (called makespan). The gathering problem is equivalent to the personalized broadcasting problem where the base station has to send messages to some nodes in the graph, with same transmission constraints. In [24], we focus on the gathering and personalized broadcasting problem in grids. Moreover, we consider the non-buffering model: when a node receives a message at some step, it must transmit it during the next step. In this setting, though the problem of determining the complexity of computing the optimal makespan in a grid is still open, we present linear (in the number of messages) algorithms that compute schedules for gathering with $dI \in \{0, 1, 2\}$. In particular, we present an algorithm that achieves the optimal makespan up to an additive constant 2 when $dI = 0$. If no messages are "close" to the axes (the base station being the origin), our algorithms achieve the optimal makespan up to an additive constant 1 when $dI = 0$, 4 when $dI = 2$, and 3 when both $dI = 1$ and the base station is in a corner. Note that, the approximation algorithms that we present also provide approximation up to a ratio 2 for the gathering with buffering. All our results are proved in terms of personalized broadcasting.

7.1.2. Elastic Optical Networks

7.1.2.1. On Spectrum Assignment in Elastic Optical Tree-Networks

To face the explosion of the Internet traffic, a new generation of optical networks is being developed; the Elastic optical Networks (EONs). The aim with EONs is to use the optical spectrum efficiently and flexibly. The benefit of the flexibility is, however, accompanied by more difficulty in the resource allocation problems. In [54], [51], [14], we study the problem of Spectrum Allocation in Elastic Optical Tree-Networks. In trees, even though the routing is fixed, the spectrum allocation is NP-hard. We survey the complexity and approximability results that have been established for the SA in trees and prove new results for stars and binary trees.

7.1.3. Fault Tolerance

7.1.3.1. Shared Risk Link Group

The notion of Shared Risk Link Groups (SRLG) captures survivability issues when a set of links of a network may fail simultaneously. The theory of survivable network design relies on basic combinatorial objects that are rather easy to compute in the classical graph models: shortest paths, minimum cuts, or pairs of disjoint paths. In the SRLG context, the optimization criterion for these objects is no longer the number of edges they use, but the number of SRLGs involved. Unfortunately, computing these combinatorial objects is NP-hard and hard to approximate with this objective in general. Nevertheless some objects can be computed in polynomial time when the SRLGs satisfy certain structural properties of locality which correspond to practical ones, namely the star property (all links affected by a given SRLG are incident to a unique node) and the span 1 property (the links affected by a given SRLG form a connected component of the network). The star property is defined in a multi-colored model where a link can be affected by several SRLGs while the span property is defined only in a mono-colored model where a link can be affected by at most one SRLG. In [59], we extend these notions to characterize new cases in which these optimization problems can be solved in polynomial time. We also investigate the computational impact of the transformation from the multi-colored model to the mono-colored one. Experimental results are presented to validate the proposed algorithms and principles.

In [22], we investigate the k-diverse routing problem which is to find a set of k pairwise SRLG-disjoint paths between a given pair of end nodes of the network. This problem has been proven NP-complete in general and some polynomial instances have been characterized. We consider more specifically the case where the SRLGs are localized and satisfy the star property. We first provide counterexamples to the polynomial time algorithm proposed by X. Luo and B. Wang (DRCN'05) for computing a pair of SRLG-disjoint paths in networks with SRLGs satisfying the star property, and then prove that this problem is in fact NP-complete. We then characterize instances that can be solved in polynomial time or are fixed parameter tractable, in particular when the number of SRLGs is constant, the maximum degree of the vertices is at most 4, and when the network is a directed acyclic graph. Finally we consider the problem of finding the maximum number of SRLG-disjoint paths in networks with SRLGs satisfying the star property. We prove that this problem is NP-hard to approximate within $O(|V|^{1-\varepsilon})$ for any $0 < \varepsilon < 1$, where V is the set of nodes in the network. Then, we provide exact and approximation algorithms for relevant subcases.

7.1.3.2. Design of Fault-tolerant On-board Networks with Variable Switch Sizes

In [29], we focus on designing networks that are capable, in the presence of faulty output ports, of rerouting input signals to operational output ports. Since the components of a satellite cannot be repaired, redundant amplifiers are added, and the interconnection network satisfies the following fault tolerance property: the network connects the set of input ports with the set of output ports, and for any set of at most k output port failures, there exists a set of edge-disjoint paths connecting the input ports to the operational output ports. Since each switching device is expensive, these interconnection networks are constructed using the fewest possible switches, or at least a number of switches close to the minimum value. The networks are controlled centrally from Earth. Each time an amplifier in use develops a fault, the controller sends messages to the switches to change their settings, so as to ensure that the inputs remain connected to functioning amplifiers.

Current switches have four ports. Obviously, the larger the number of ports, the more expensive will be the switches, but then fewer will be required. So the cost of such a network involves a trade-off between the total number of switches and their unit cost. In order to determine the minimum-cost network, we give some bounds on the minimum number $\mathcal{N}(n,k,r)$ of $2r$-port switches in interconnection networks with n inputs and $n+k$ outputs.

We first show $\mathcal{N}(n,k,r) \leq \left\lceil \frac{k+2}{2r-2} \right\rceil \left\lceil \frac{n}{2} \right\rceil$. When $r \geq k/2$, we prove a better upper bound: $\mathcal{N}(n,k,r) \leq \frac{r-2+k/2}{r^2-2r+k/2} n + O(1)$. Next, we establish some lower bounds. We show that if $k \geq r$, then $\mathcal{N}(n,k,r) \geq \frac{3n+k}{2r}$. We improve this bound when $k \geq 2r$: $\mathcal{N}(n,k,r) \geq \frac{3n+2k/3-r/2}{2r-2+\frac{3r}{\lfloor \frac{k}{r} \rfloor}}$. Finally, we determine $\mathcal{N}(n,k,r)$ up to additive constants for $k \leq 6$.

7.1.4. Reducing Networks' Energy Consumption

Due to the increasing impact of ICT (Information and Communication Technology) on power consumption and worldwide gas emissions, energy efficient ways to design and operate backbone networks are becoming a new concern for network operators. Recently, energy-aware routing (EAR) has gained an increasing popularity in the networking research community. The idea is that traffic demands are redirected over a subset of the network links, allowing other links to sleep to save energy. We studied variant of this problems.

7.1.4.1. Robust Energy-aware Routing with Redundancy Elimination

In [31], we propose GreenRE – a new EAR model with the support of data redundancy elimination (RE). This technique, enabled within routers, can virtually increase the capacity of network links. Based on real experiments on a Orange Labs platform, we show that performing RE increases the energy consumption for routers. Therefore, it is important to determine which routers should enable RE and which links to put into sleep mode so that the power consumption of the network is minimized. We model the problem as a Mixed Integer Linear Program and propose greedy heuristic algorithms for large networks. Simulations on several network topologies show that the GreenRE model can gain further 37% of energy savings compared to the classical EAR model. In [27], we introduce an extended model of the classical multi-commodity flow problem with compressible flows which is also robust with fluctuation of traffic demand and compression rate. An heuristic built on this model allows for 16-28% extra energy saving.

7.1.4.2. Optimizing IGP Link Weights for Energy-efficiency in Multi-period Traffic Matrices

To guarantee QoS while implementing EAR, all traffic demands should be routed without violating capacity constraints and the network should keep its connectivity. From the perspective of traffic engineering, we argue that stability in routing configuration also plays an important role in QoS. In details, frequent changes in network configuration (link weights, slept and activated links) to adapt with traffic fluctuation in daily time cause network oscillations. In [35], we propose a novel optimization method to adjust the link weights of Open Shortest Path First (OSPF) protocol while limiting the changes in network configurations when multi-period traffic matrices are considered.

7.1.4.3. Energy Efficient Content Distribution

Recently, there is a trend to introduce content caches as an inherent capacity of network equipment, with the objective of improving the efficiency of content distribution and reducing the network congestion. In [18], we study the impact of using in-network caches and content delivery network (CDN) cooperation on an EAR. We formulate this problem as Energy Efficient Content Distribution, we propose an integer linear program (ILP) and a heuristic algorithm to solve it. The objective of this problem is to find a feasible routing, so that the total energy consumption of the network is minimized while the constraints given by the demands and the link capacity are satisfied. We exhibit for which the range of parameters (size of caches, popularity of content, demand intensity, etc.) it is useful to use caches. Experimental results show that by placing a cache on each backbone router to store the most popular content, along with well choosing the best content provider server for each demand to a CDN, we can save about 20% of power in average of all the backbone networks considered.

7.1.5. Routing Theory and Forwarding Index

Motivated by finding the best set of links that should be on for energy efficiency, we study the problem of determining the minimum forwarding index of a graph. The (edge) forwarding index of a graph is the minimum, over all possible routings of all the demands, of the maximum load of an edge. This metric is of a great interest since it captures the notion of global congestion in a precise way: the lesser the forwarding-index, the lesser the congestion. This parameter has been studied for different graph classes in the literature. In [42], we determine, for different numbers of edges, the best spanning graphs of a square grid, namely those with a low forwarding index. In [61], [43], we study the following design question: Given a number e of edges and a number n of vertices, what is the least congested graph that we can construct? and what forwarding-index can we achieve? We answer here these questions for different families of graphs: general graphs, graphs with bounded degree, sparse graphs with a small number of edges by providing constructions, most of them

asymptotically optimal. Doing so, we partially answer the practical problem that initially motivated our work: If an operator wants to power only e links of its network, in order to reduce the energy consumption (or wiring cost) of its networks, what should be those links and what performance can be expected?

On the complexity of equal shortest path routing.

Additionally, we studied the complexity of configuring the OSPF-ECMP (for Open Shortest Path First-Equal Cost Multiple Path) protocol. In [32], we show that the problem of maximizing even a single commodity flow for the OSPF-ECMP protocol cannot be approximated within any constant factor ratio. Besides this main theorem, we derive some positive results which include polynomial-time approximations and an exponential-time exact algorithm.

7.1.6. Routing in Software Defined Networks (SDN)

Software Defined Networking (SDN) is gaining momentum with the support of major manufacturers. While it brings flexibility in the management of flows within the data center fabric, this flexibility comes at the cost of smaller routing table capacities. In [50], we investigate compression techniques to reduce the forwarding information base (FIB) of SDN switches. We validate our algorithm, called MINNIE, on a real testbed able to emulate a 20 switches fat tree architecture. We demonstrate that even with a small number of clients, the limit in terms of number of rules is reached if no compression is performed, increasing the delay of all new incoming flows. MINNIE, on the other hand, reduces drastically the number of rules that need to be stored with a limited impact on the packet loss rate. We also evaluate the actual switching and reconfiguration times and the delay introduced by the communications with the controller. In parallel, we considered the algorithmic problem of compressing bidimensional routings table with priorities on the rules. We carry out in [40] a study of the problem complexity, providing results of NP-completeness, of Fixed-Parameter Tractability and approximation algorithms. In [44], we then propose green routing schemes performing simultaneously the selection of the routes, the compression of the routing tables, and decide to put in sleep mode unused links. These algorithms are tested on networks from the SNDLib library.

7.1.7. Video Streaming

7.1.7.1. Study of Repair Protocols for Live Video Streaming Distributed Systems

In [41], we study distributed systems for live video streaming. These systems can be of two types: structured and un-structured. In an unstructured system, the diffusion is done opportunistically. The advantage is that it handles churn, that is the arrival and departure of users, which is very high in live streaming systems, in a smooth way. On the opposite, in a structured system, the diffusion of the video is done using explicit diffusion trees. The advantage is that the diffusion is very efficient, but the structure is broken by the churn. In this paper, we propose simple distributed repair protocols to maintain, under churn, the diffusion tree of a structured streaming system. We study these protocols using formal analysis and simulation. In particular, we provide an estimation of the system metrics, bandwidth usage, delay, or number of interruptions of the streaming. Our work shows that structured streaming systems can be efficient and resistant to churn.

7.2. Graph Algorithms

Participants: Nathann Cohen, David Coudert, Frédéric Giroire, Fatima Zahra Moataz, Benjamin Momège, Nicolas Nisse, Stéphane Pérennes.

COATI is also interested in the algorithmic aspects of Graph Theory. In general we try to find the most efficient algorithms to solve various problems of Graph Theory and telecommunication networks. We use graph theory to model various network problems. We study their complexity and then we investigate the structural properties of graphs that make these problems hard or easy. In particular, we try to find the most efficient algorithms to solve the problems, sometimes focusing on specific graph classes from which the problems are polynomial-time solvable. Many results introduced here are presented in detail in the PhD thesis of F. Z. Moataz [14].

7.2.1. Graph Hyperbolicity

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

7.2.1.1. Exact Algorithms for Computing the Gromov Hyperbolicity

The best known theoretical algorithm computing this parameter runs in $O(n^{3.69})$ time, which is prohibitive for large-scale graphs. In [26], we propose an algorithm for determining the hyperbolicity of graphs with tens of thousands of nodes. Its running time depends on the distribution of distances and on the actual value of the hyperbolicity. Although its worst case runtime is $O(n^4)$, it is in practice much faster than previous proposals as observed in our experimentations on benchmark instances. We also propose a heuristic algorithm that can be used on graphs with millions of nodes.

In [37], we provide a more efficient algorithm: although its worst-case complexity remains in $O(n^4)$, in practice it is much faster, allowing, for the first time, the computation of the hyperbolicity of graphs with up to $200,000$ nodes. We experimentally show that our new algorithm drastically outperforms the best previously available algorithms, by analyzing a big dataset of real-world networks. We have also used the new algorithm to compute the hyperbolicity of random graphs generated with the Erdös-Renyi model, the Chung-Lu model, and the Configuration Model.

7.2.1.2. Hyperbolicity of Particular Graph Classes

Topologies for data center networks have been proposed in the literature through various graph classes and operations. A common trait to most existing designs is that they enhance the symmetric properties of the underlying graphs. Indeed, symmetry is a desirable property for interconnection networks because it minimizes congestion problems and it allows each entity to run the same routing protocol. However, despite sharing similarities these topologies all come with their own routing protocol. Recently, generic routing schemes have been introduced which can be implemented for any interconnection networks. The performances of such universal routing schemes are intimately related to the hyperbolicity of the topology. Motivated by the good performances in practice of these new routing schemes, we propose in [56] the first general study of the hyperbolicity of data center interconnection networks. Our findings are disappointingly negative: we prove that the hyperbolicity of most data center topologies scales linearly with their diameter, that it the worst-case possible for hyperbolicity. To obtain these results, we introduce original connection between hyperbolicity and the properties of the endomorphism monoid of a graph. In particular, our results extend to all vertex and edge-transitive graphs. Additional results are obtained for de Bruijn and Kautz graphs, grid-like graphs and networks from the so-called Cayley model.

In [57], we investigate more specifically on the hyperbolicity of bipartite graphs. More precisely, given a bipartite graph $B = (V_0 \cup V_1, E)$ we prove it is enough to consider any one side V_i of the bipartition of B to obtain a close approximate of its hyperbolicity $\delta(B)$ — up to an additive constant 2. We obtain from this result the sharp bounds $\delta(G) - 1 \leq \delta(L(G)) \leq \delta(G) + 1$ and $\delta(G) - 1 \leq \delta(K(G)) \leq \delta(G) + 1$ for every graph G, with $L(G)$ and $K(G)$ being respectively the line graph and the clique graph of G. Finally, promising extensions of our techniques to a broader class of intersection graphs are discussed and illustrated with the case of the biclique graph $BK(G)$, for which we prove $(\delta(G) - 3)/2 \leq \delta(BK(G)) \leq (\delta(G) + 3)/2$.

7.2.2. Tree-decompositions

We study the computational complexity of different variants of tree-decompositions. We also study their relationship with various pursuit-evasion games.

7.2.2.1. Diameter of Minimal Separators in Graphs (structure vs metric in graphs)

In [39], we establish general relationships between the topological properties of graphs and their metric properties. For this purpose, we upper-bound the diameter of the *minimal separators* in any graph by a function of their sizes. More precisely, we prove that, in any graph G, the diameter of any minimal separator S in G is

at most $\lfloor \frac{\ell(G)}{2} \rfloor \cdot (|S|-1)$ where $\ell(G)$ is the maximum length of an isometric cycle in G. We refine this bound in the case of graphs admitting a *distance preserving ordering* for which we prove that any minimal separator S has diameter at most $2(|S|-1)$. Our proofs are mainly based on the property that the minimal separators in a graph G are connected in some power of G. Our result easily implies that the *treelengthtl*(G) of any graph G is at most $\lfloor \frac{\ell(G)}{2} \rfloor$ times its *treewidthtw*(G). In addition, we prove that, for any graph G that excludes an *apex graph*H as a minor, $tw(G) \leq c_H \cdot tl(G)$ for some constant c_H only depending on H. We refine this constant when G has bounded genus. As a consequence, we obtain a very simple $O(\ell(G))$-approximation algorithm for computing the treewidth of n-node m-edge graphs that exclude an apex graph as a minor in $O(nm)$-time.

7.2.2.2. Minimum Size Tree-decompositions

Tree-decompositions are the cornerstone of many dynamic programming algorithms for solving graph problems. Since the complexity of such algorithms generally depends exponentially on the width (size of the bags) of the decomposition, much work has been devoted to compute tree-decompositions with small width. However, practical algorithms computing tree-decompositions only exist for graphs with treewidth less than 4. In such graphs, the time-complexity of dynamic programming algorithms is dominated by the size (number of bags) of the tree-decompositions. It is then interesting to minimize the size of the tree-decompositions. In [48], [14], we consider the problem of computing a tree-decomposition of a graph with width at most k and minimum size. We prove that the problem is NP-complete for any fixed $k \geq 4$ and polynomial for $k \leq 2$; for $k = 3$, we show that it is polynomial in the class of trees and 2-connected outerplanar graphs.

7.2.2.3. Non-deterministic Graph Searching in Trees

Non-deterministic graph searching was introduced by Fomin et al. to provide a unified approach for pathwidth, treewidth, and their interpretations in terms of graph searching games. Given $q \geq 0$, the q-limited search number, $s_q(G)$, of a graph G is the smallest number of searchers required to capture an invisible fugitive in G, when the searchers are allowed to know the position of the fugitive at most q times. The search parameter $s_0(G)$ corresponds to the pathwidth of a graph G, and $s_\infty(G)$ to its treewidth. Determining $s_q(G)$ is NP-complete for any fixed $q \geq 0$ in general graphs and $s_0(T)$ can be computed in linear time in trees, however the complexity of the problem on trees has been unknown for any $q > 0$. We introduce in [16] a new variant of graph searching called restricted non-deterministic. The corresponding parameter is denoted by rs_q and is shown to be equal to the non-deterministic graph searching parameter s_q for $q = 0, 1$, and at most twice s_q for any $q \geq 2$ (for any graph G). Our main result is a polynomial time algorithm that computes $rs_q(T)$ for any tree T and any $q \geq 0$. This provides a 2-approximation of $s_q(T)$ for any tree T , and shows that the decision problem associated to s_1 is polynomial in the class of trees. Our proofs are based on a new decomposition technique for trees which might be of independent interest.

7.2.2.4. k-Chordal Graphs: from Cops and Robber to Compact Routing via Treewidth

Cops and robber games, introduced by Winkler and Nowakowski and independently defined by Quilliot, concern a team of cops that must capture a robber moving in a graph. We consider in [34] the class of k-chordal graphs, i.e., graphs with no induced (chordless) cycle of length greater than k, $k \geq 3$. We prove that $k-1$ cops are always sufficient to capture a robber in k-chordal graphs. This leads us to our main result, a new structural decomposition for a graph class including k-chordal graphs. We present a polynomial-time algorithm that, given a graph G and $k \geq 3$, either returns an induced cycle larger than k in G, or computes a tree-decomposition of G, each bag of which contains a dominating path with at most $k-1$ vertices. This allows us to prove that any k-chordal graph with maximum degree Δ has treewidth at most $(k-1)(\Delta-1)+2$, improving the $O(\Delta(\Delta-1)k-3)$ bound of Bodlaender and Thilikos (1997). Moreover, any graph admitting such a tree-decomposition has small hyperbolicity. As an application, for any n-vertex graph admitting such a tree-decomposition, we propose a compact routing scheme using routing tables, addresses and headers of size $O(k \log \Delta + \log n)$ bits and achieving an additive stretch of $O(k \log \Delta)$. As far as we know, this is the first routing scheme with $O(k \log \Delta + \log n)$-routing tables and small additive stretch for k-chordal graphs.

7.2.2.5. Connected Surveillance Game

The surveillance game [68] models the problem of web-page prefetching as a pursuit evasion game played on a graph. This two-player game is played turn-by-turn. The first player, called the observer, can mark a fixed

amount of vertices at each turn. The second one controls a surfer that stands at vertices of the graph and can slide along edges. The surfer starts at some initially marked vertex of the graph, its objective is to reach an unmarked node before all nodes of the graph are marked. The surveillance number $sn(G)$ of a graph G is the minimum amount of nodes that the observer has to mark at each turn ensuring it wins against any surfer in G. Fomin et al. also defined the connected surveillance game where the observer must ensure that marked nodes always induce a connected subgraph. They ask what is the cost of connectivity, i.e., is there a constant $c > 0$ such that the ratio between the connected surveillance number $csn(G)$ and $sn(G)$ is at most c for any graph G. It is straightforward to show that $csn(G) \leq \Delta sn(G)$ for any graph G with maximum degree Δ. Moreover, it has been shown that there are graphs G for which $csn(G) = sn(G) + 1$. In [30], we investigate the question of the cost of the connectivity. We first provide new non-trivial upper and lower bounds for the cost of connectivity in the surveillance game. More precisely, we present a family of graphs G such that $csn(G) > sn(G) + 1$. Moreover, we prove that $csn(G) \leq \sqrt{sn(G)n}$ for any n-node graph G. While the gap between these bounds remains huge, it seems difficult to reduce it. We then define the online surveillance game where the observer has no a priori knowledge of the graph topology and discovers it little-by-little. This variant, which fits better the prefetching motivation, is a restriction of the connected variant. Unfortunately, we show that no algorithm for solving the online surveillance game has competitive ratio better than $\Omega(\Delta)$. That is, while interesting, this variant does not help to obtain better upper bounds for the connected variant. We finally answer an open question [68] by proving that deciding if the surveillance number of a digraph with maximum degree 6 is at most 2 is NP-hard.

7.2.3. Distributed Algorithms

7.2.3.1. Allowing each Node to Communicate only once in a Distributed System: Shared Whiteboard Models

In [21] we study distributed algorithms on massive graphs where links represent a particular relationship between nodes (for instance, nodes may represent phone numbers and links may indicate telephone calls). Since such graphs are massive they need to be processed in a distributed way. When computing graph-theoretic properties, nodes become natural units for distributed computation. Links do not necessarily represent communication channels between the computing units and therefore do not restrict the communication flow. Our goal is to model and analyze the computational power of such distributed systems where one computing unit is assigned to each node. Communication takes place on a whiteboard where each node is allowed to write at most one message. Every node can read the contents of the whiteboard and, when activated, can write one small message based on its local knowledge. When the protocol terminates its output is computed from the final contents of the whiteboard. We describe four synchronization models for accessing the whiteboard. We show that message size and synchronization power constitute two orthogonal hierarchies for these systems.We exhibit problems that separate these models, i.e., that can be solved in one model but not in a weaker one, even with increased message size. These problems are related to maximal independent set and connectivity. We also exhibit problems that require a given message size independently of the synchronization model.

7.2.3.2. Computing on Rings by Oblivious Robots: a Unified Approach for Different Tasks

A set of autonomous robots have to collaborate in order to accomplish a common task in a ring-topology where neither nodes nor edges are labeled (that is, the ring is anonymous). In [36], we present a unified approach to solve three important problems: the exclusive perpetual exploration, the exclusive perpetual clearing, and the gathering problems. In the first problem, each robot aims at visiting each node infinitely often while avoiding that two robots occupy a same node (exclusivity property); in exclusive perpetual clearing (also known as searching), the team of robots aims at clearing the whole ring infinitely often (an edge is cleared if it is traversed by a robot or if both its endpoints are occupied); and in the gathering problem, all robots must eventually occupy the same node. We investigate these tasks in the Look-Compute-Move model where the robots cannot communicate but can perceive the positions of other robots. Each robot is equipped with visibility sensors and motion actuators, and it operates in asynchronous cycles. In each cycle, a robot takes a snapshot of the current global configuration (Look), then, based on the perceived configuration, takes a decision to stay idle or to move to one of its adjacent nodes (Compute), and in the latter case it eventually moves to this neighbor (Move). Moreover, robots are endowed with very weak capabilities. Namely, they are anonymous, asynchronous, oblivious, uniform (execute the same algorithm) and have no common sense of

orientation. In this setting, we devise algorithms that, starting from an exclusive and rigid (i.e. aperiodic and asymmetric) configuration, solve the three above problems in anonymous ring-topologies.

7.2.4. Miscellaneous

7.2.4.1. Finding Paths in Grids with Forbidden Transitions

A transition in a graph is a pair of adjacent edges. Given a graph $G = (V, E)$, a set of forbidden transitions $F \subseteq E \times E$ and two vertices $s, t \in V$, we study in [64], [45], [46], [14] the problem of finding a path from s to t which uses none of the forbidden transitions of F. This means that it is forbidden for the path to consecutively use two edges forming a pair in F. The study of this problem is motivated by routing in road networks in which forbidden transitions are associated to prohibited turns as well as routing in optical networks with asymmetric nodes, which are nodes where a signal on an ingress port can only reach a subset of egress ports. If the path is not required to be elementary, the problem can be solved in polynomial time. On the other side, if the path has to be elementary, the problem is known to be NP-complete in general graphs [69]. In [45], we study the problem of finding an elementary path avoiding forbidden transitions in planar graphs. We prove that the problem is NP-complete in planar graphs and particularly in grids. In addition, we show that the problem can be solved in polynomial time in graphs with bounded treewidth. More precisely, we show that there is an algorithm which solves the problem in time $O((3\Delta(k+1))2k + 4n))$ in n-node graphs with treewidth at most k and maximum degree Δ.

7.3. Graph theory

Participants: Nathann Cohen, Frédéric Havet.

7.3.1. Graph Colouring

7.3.1.1. Steinberg-like Theorems for Backbone Colouring

Motivated by some channel assignment problem, we study the following variation of graph colouring problem. A function $f : V(G) \to \{1, \cdots, k\}$ 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 c 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 $CBC_q(G, H)$, is the minimum k such that (G, H) admits a circular q-backbone k-colouring. In [19], 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 $CBC_2(G, H) \leq 7$. Then, we prove that if $H \subseteq G$ is a forest whose connected components are paths, then $CBC_2(G, H) \leq 6$.

7.3.1.2. Complexity of Greedy Edge-colouring

The Grundy index of a graph $G = (V, E)$ is the greatest number of colours that the greedy edge-colouring algorithm can use on G. In [33], we prove that the problem of determining the Grundy index of a graph $G = (V, E)$ is NP-hard for general graphs. We also show that this problem is polynomial-time solvable for caterpillars. More specifically, we prove that the Grundy index of a caterpillar is $\Delta(G)$ or $\Delta(G) + 1$ and present a polynomial-time algorithm to determine it exactly.

7.3.1.3. Proper Orientation Number

An *orientation* of a graph G is a digraph D obtained from G by replacing each edge by exactly one of the two possible arcs with the same endvertices. For each $v \in V(G)$, the *indegree* of v in D, denoted by $d_D^-(v)$, is the number of arcs with head v in D. An orientation D of G is *proper* if $d_D^-(u) \neq d_D^-(v)$, for all $uv \in E(G)$. The *proper orientation number* of a graph G, denoted by $\overrightarrow{\chi}(G)$, is the minimum of the maximum indegree over all its proper orientations. It is well-known that $\chi(G) \leq \overrightarrow{\chi}(G) + 1 \leq \Delta(G) + 1$, for every graph G, where $\chi(G)$ and $\Delta(G)$ denotes the chromatic number and the maximum degree of G. In other words, the proper orientation number (plus one) is an upper bound on the chromatic number which is tighter than the maximum degree.

In [17], we ask whether the proper orientation number is really a more accurate bound than the maximum degree in the following sense : does there exists a positive ϵ and such that $\overrightarrow{\chi}(G) \leq \epsilon \cdot \chi(G) + (1-\epsilon)\Delta(G)$.

As an evidence to this, we prove that if G is bipartite (i.e. $\chi(G) \leq 2$) then $\overrightarrow{\chi}(G) \leq \left(\Delta(G) + \sqrt{\Delta(G)}\right)/2 + 1$.

However, the proper orientation number has the drawback to be difficult to compute. We prove in [17] that deciding whether $\overrightarrow{\chi}(G) \leq \Delta(G) - 1$ is already an NP-complete problem on graphs with $\Delta(G) = k$, for every $k \geq 3$. We also show that it is NP-complete to decide whether $\overrightarrow{\chi}(G) \leq 2$, for planar *subcubic* graphs G. Moreover, we prove that it is NP-complete to decide whether $\overrightarrow{\chi}(G) \leq 3$, for planar bipartite graphs G with maximum degree 5.

Nevertheless, it might be interesting to bound the proper orientation number on some graph families. In particular, if we prove that for a graph with treewidth at most t, the proper orientation number is bounded by a function of t, this would imply that finding the proper orientation number of a graph with bounded treewidth is polynomial-time solvable. In [17] we prove $\overrightarrow{\chi}(G) \leq 4$ if G is a tree (or equivalently a graph with treewidth at most 1). In [53], we study the cacti which is a special class of graphs with treewidth at most 2. We prove that $\overrightarrow{\chi}(G) \leq 7$ for every cactus. We also prove that the bound 7 is tight by showing a cactus having no proper orientation with maximum indegree less than 7. We also prove that any planar claw-free graph has a proper orientation with maximum indegree at most 6 and that this bound can also be attained.

7.3.2. Subdivisions of Digraphs

An important result in the Roberston and Seymour minor theory is the polynomial-time algorithm to solve the so-called Linkage Problem. This implies in particular, that for any fixed graph H, deciding whether a graph G contains a subdivision of H as a subgraph can be solved in polynomial time.

We consider the directed analogue F-subdivision problem, which is an analogue for directed graphs (i.e. digraphs). Given a directed graph D, does it contain a subdivision of a prescribed digraph F? In [20], we give a number of examples of polynomial instances, several NP-completeness proofs as well as a number of conjectures and open problems. In [62], we give further support to several open conjectures and speculations about algorithmic complexity of finding F-subdivisions. In particular, 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.

7.4. Applications to Other Domains

Participants: Christelle Caillouet, David Coudert, Nicolas Nisse.

7.4.1. Unveiling Contacts within Macro-molecular assemblies by solving Minimum Weight Connectivity Inference Problems

Consider a set of oligomers listing the subunits involved in sub-complexes of a macro-molecular assembly, obtained e.g. using native mass spectrometry or affinity purification. Given these oligomers, connectivity inference (CI) consists in finding the most plausible contacts between these subunits, and minimum connectivity inference (MCI) is the variant consisting in finding a set of contacts of smallest cardinality. MCI problems avoid speculating on the total number of contacts, but yield a subset of all contacts and do not allow exploiting a priori information on the likelihood of individual contacts. In this context, we present in [15] two novel algorithms, MILP-W and MILP-WB. The former solves the minimum weight connectivity inference (MWCI), an optimization problem whose criterion mixes the number of contacts and their likelihood. The latter uses the former in a bootstrap fashion, to improve the sensitivity and the specificity of solution sets. Experiments on three systems (yeast exosome, yeast proteasome lid, human eiF3), for which reference contacts are known (crystal structure, cryo electron microscopy, cross-linking), show that our algorithms predict contacts with high specificity and sensitivity, yielding a very significant improvement over previous work, typically a twofold increase in sensitivity. The software accompanying this paper is made available, and should prove of ubiquitous interest whenever connectivity inference from oligomers is faced.

7.4.2. Recovery of Disrupted Airline Operations using k-Maximum Matching in Graphs

When an aircraft is approaching an airport, it gets a short time interval (called *slot*) that it can use to land. If the landing of the aircraft is delayed (because of bad weather, or if it arrives late, or if other aircrafts have to land first), it loses its slot and Air traffic controllers have to assign it a new slot. However, slots for landing are a scarce resource of the airports and, to avoid that an aircraft waits too much time, Air traffic controllers have to regularly modify the assignment of the slots of the aircrafts. Unfortunately, for legal and economical reasons, Air traffic controllers can modify the slot-assignment only using two kind of operations: either assign to aircraft A a slot that was free, or give to A the slot of another aircraft B and assign to B a free slot. The problem is then the following. Let $k \geq 1$ be an odd integer and let G be a graph and M be a matching (set of pairwise disjoint edges) of G. What is the maximum size of a matching that can be obtained from M by using only augmenting paths of length at most k? Moreover, how to compute such a maximum matching? This problem has already been studied in the context of wireless networks, mainly because it provides a simple approximation for the classical matching problem. We prove in [65], [49] that this problem can be solved in polynomial-time when $k \leq 3$. Then, we show that, for any odd integer $k \geq 5$, the problem is NP-complete in planar bipartite graphs with maximum degree at most 3.

7.4.3. Inference of Curvilinear Structure based on Learning a Ranking Function and Graph Theory

To detect curvilinear structures in natural images, we propose in [63] a novel ranking learning system and an abstract curvilinear shape inference algorithm based on graph theory. We analyze the curvilinear structures as a set of small line segments. In this work, the rankings of the line segments are exploited to systematize the topological feature of the curvilinear structures. Structured Support Vector Machine is employed to learn the ranking function that predicts the correspondence of the given line segments and the latent curvilinear structures. We first extract curvilinear features using morphological profiles and steerable filtering responses. Also, we propose an orientation-aware feature descriptor and a feature grouping operator to improve the structural integrity during the learning process. To infer the curvilinear structure, we build a graph based on the output rankings of the line segments. We progressively reconstruct the curvilinear structure by looking for paths between remote vertices in the graph. Experimental results show that the proposed algorithm faithfully detects the curvilinear structures within various datasets.

7.4.4. Web Transparency for Complex Targeting: Algorithms, Limits, and Tradeoffs

Big Data promises important societal progress but exacerbates the need for due process and accountability. Companies and institutions can now discriminate between users at an individual level using collected data or past behavior. Worse, today they can do so in near perfect opacity. The nascent field of web transparency aims to develop the tools and methods necessary to reveal how information is used, however today it lacks robust tools that let users and investigators identify targeting using multiple inputs. In [67], [38], we formalize for the first time the problem of detecting and identifying targeting on combinations of inputs and provide the first algorithm that is asymptotically exact. This algorithm is designed to serve as a theoretical foundational block to build future scalable and robust web transparency tools. It offers three key properties. First, our algorithm is service agnostic and applies to a variety of settings under a broad set of assumptions. Second, our algorithm's analysis delineates a theoretical detection limit that characterizes which forms of targeting can be distinguished from noise and which cannot. Third, our algorithm establishes fundamental tradeoffs that lead the way to new metrics for the science of web transparency. Understanding the tradeoff between effective targeting and targeting concealment lets us determine under which conditions predatory targeting can be made unprofitable by transparency tools.

DANTE Project-Team

7. New Results

7.1. Graph & Signal Processing

Participants: Paulo Gonçalves Andrade, Éric Fleury, Benjamin Girault, Sarra Ben Alaya.

Isometric Graph shift operator. In [14], [40], we proposed a new shift operator for graph signals, enforcing that our operator is isometric. Doing so, we ensure that as many properties of the time shift as possible get carried over. Finally, we show that our operator behaves reasonably for graph signals.

Stationary graph signals. We extended the concept of stationary temporal signals to stationary graph signals [24]. We introduced the concept of strict sense stationary and wide sense stationary graph signals as a statistical invariance through an isometric graph translation. Using these definitions, we proposed a spectral characterisation of WSS graph signals allowing to study stationarity using only the spectral components of a graph signal. Finally, we applied this characterisation to a synthetic graph in order to study a few important stochastic graph signals. Also, using geographic data, we analysed data from a graph set of weather stations and showed evidence of stationarity in the temperature signal [36].

Community mining with graph filters for correlation matrices. Communities are an important type of structure in networks. Graph filters, such as wavelet filter-banks, have been used to detect such communities as groups of nodes more densely connected together than with the outsiders. When dealing with times series, it is possible to build a relational network based on the correlation matrix. However, in such a network, weights assigned to each edge have different properties than those of usual adjacency matrices. As a result, classical community detection methods based on modularity optimisation are not consistent and the modularity needs to be redefined to take into account the structure of the correlation from random matrix theory. In our contribution [34], we addressed how to detect communities from correlation matrices, by filtering global modes and random parts using properties that are specific to the distribution of correlation eigenvalues. Based on a Louvain approach, an algorithm to detect multiscale communities was also developed, which yields a weighted hierarchy of communities. The implementation of the method using graph filters was also discussed.

A strong Tauberian theorem for characteristic functions. In [20], we showed that a characteristic function which can be approximated at 0 by any polynomial of order n is actually n-times differentiable at 0. This fact is exploited to strengthen a tauberian-type result by Lukacs and provides the theoretical basis for a wavelet based non-parametric estimator of the tail index of a distribution. This work is a technical improvement of our previous contribution [53].

Fractal Analysis of Fetal Heart Rate Variability. The fetal heart rate (FHR) is commonly monitored during labor to detect early fetal acidosis. FHR variability is traditionally investigated using Fourier transform, often with adult predefined frequency band powers and the corresponding LF/HF ratio. However, fetal conditions differ from adults and modify spectrum repartition along frequencies. The study we reported in [12] questioned the arbitrariness definition and relevance of the frequency band splitting procedure, and thus of the calculation of the underlying LF/HF ratio, as efficient tools for characterising intrapartum FHR variability. Then, we showed that the intrapartum FHR is characterised by fractal temporal dynamics and promotes the Hurst parameter as a potential marker of fetal acidosis. This parameter preserves the intuition of a power frequency balance, while avoiding the frequency band splitting procedure and thus the arbitrary choice of a frequency separating bands. The study also shows that extending the frequency range covered by the adult-based bands to higher and lower frequencies permits the Hurst parameter to achieve better performance for identifying fetal acidosis.

7.2. Performance analysis and networks protocols

Participants: Paulo Gonçalves Andrade, Thomas Begin, Anthony Busson, Isabelle Guérin Lassous, Laurent Reynaud, Thiago Wanderley Matos de Abreu.

Global computing-network-visualisation. The PetaFlow application aims to contribute to the use of high performance computational resources for the benefit of society. To this goal the emergence of adequate information and communication technologies with respect to high performance computing-networking-visualisation and their mutual awareness is required. In the work published in [5], we present the developed technology and the algorithms that we applied to a real global peta-scale data intensive scientific problem with social and medical importance, i.e. human upper airflow modeling.

Performance analysis of multi-hop flows in IEEE 802.11 networks Multi-hop wireless networks are often regarded as a promising means to extend the limited coverage area offered by WLANs. However, they are usually associated with poor and uncertain performance in terms of available bandwidth and packet losses, which clearly stands as a limitation to their use. In [7], we consider the performance evaluation of a multi-hop path (also called chain), based on the IEEE 802.11 DCF. The proposed modeling framework is constructive and versatile, so that it can handle various types of multi-hop wireless paths, including scenarios with two flows in opposite directions, and topologies where nodes are exposed to the well-known hidden node problem. The models derived from our framework are conceptually simple, easy to implement and produce generally accurate results for the attained goodput of flows, as well as the datagram loss probability. Typical relative errors for these two quantities are below a few percent. Also, fundamental phenomena occurring in multi-hop wireless networks such as performance collapse and starvation, are well captured by the models.

Passive Measurement-based Estimator for the Standard Deviation of the End-to-End Delay. Emerging architectures for computer networks such as SDN aim at offering a better handling of flows with stringent requirements of QoS. On the one hand, operators would benefit from a detailed description of common network performance (e.g., end-to-end delay and end-to end loss ratio) including their first two moments, namely mean and standard deviation. Indeed, for many applications, the variability in the end-to-end delay (e.g., jitter) deeply affects the actual QoS experienced by a flow. On the other hand, the cost and nuisance associated with the instrumentation, the measurements, and the computations must be kept as low as possible. This typically prevents the availability of end-to-end measurements. In [30], we propose an algorithm to estimate the second moment of the end-to-end delay experienced by the packets of a flow based only on delay measurements locally collected by the network nodes. Our solution estimates the standard deviation of the end-to-end delay in an easy and computationally efficient way. Based on thousands of simulations using a real-life trace, our solution is found to be accurate, typically differing by only a few percent from the actual value of the standard deviation of the end-to-end delay.

Design of a force-based controlled mobility on aerial vehicles for pest management. Vespa velutina, also known as the Asian hornet, is considered as an invasive species out of its native zone. In particular, since it preys on honey bees, its recent progression in Europe could soon pose a significant risk to the local apiculture activity. European beekeepers are therefore investigating adapted control strategies, including V. velutina nest destruction. Unfortunately, nest location pinpointing generally follows a manual process which can prove tedious, time-consuming and inaccurate. In [31], we propose the use of a network of micro aerial vehicles featuring autonomous and cooperative flight capabilities. We describe an adapted controlled mobility strategy and detail the design of our Virtual Force Protocol (VFP) which allows a swarm of vehicles to track and follow hornets to their nests, while maintaining connectivity through a wireless multi-hop communication route with a remote ground station used to store applicative data such as hornet trajectory and vehicle telemetry. In order to achieve the mission objectives with a minimum of vehicles, we identify through simulations appropriate value for the key parameters of VFP and discuss the obtained network performance.

Channel assignment in IEEE 802.11-based substitution networks. A substitution network is a rapidly deployable wireless network that provides a backup solution to quickly react to failures

on an existing network. We assume that the substitution network uses Wi-Fi technology and that wireless routers are equipped with several Wi-Fi cards. The problem, addressed in this work, deals with the channel assignment to these wireless interfaces. In this particular context, there is only one source-destination pair for which paths are known in advance. It is then possible to derive an objective function, function of the channel assignment, that very precisely reflects the overall throughput that can be achieved in this network. This problemis formulated through a linear optimization problem for which we propose different heuristics. Simulation results, performed with ns-3, consider several scenarios, and compare our heuristics to the optimum. Simulations show that, with only a few wireless cards, the throughput is signifficantly increased. Also, we show that the objective function fits to the throughput measured with ns-3.

Performance evaluation and message dissemination in vehicular networks. Vehicular Ad-Hoc Network (VANET) is becoming a promising technology for improving the efficiency and the safety of Intelligent Transportation Systems (ITS). Smart vehicles are expected to continuously exchange a huge amount of data either through safety or non-safety messages dedicated for road safety or infotainment and passenger comfort applications, respectively. In this context we proposed two contributions: the estimation of the capacity offered by the wireless network [13] in order to dimension the applications, and the proposal of an efficient message dissemination protocol [25].

Performance Evaluation of Cloud Computing Centers with General Arrivals and Service. Cloud providers need to size their systems to determine the right amount of resources to allocate as a function of customer's needs so as to meet their SLAs (Service Level Agreement), while at the same time minimizing their costs and energy use. Queueing theory based tools are a natural choice when dealing with performance aspects of the QoS (Quality of Service) part of the SLA and forecasting resource utilization. The characteristics of a cloud center lead to a queueing system with multiple servers (nodes) in which there is potentially a very large number of servers and both the arrival and service process can exhibit high variability. We propose to use a G/G/c-like model to represent a cloud system and assess expected performance indices. Given the potentially high number of servers in a cloud system, we present an efficient, fast and easy-to-implement approximate solution. We have extensively validated our approximation against discrete-event simulation for several QoS performance metrics such as task response time and blocking probability with excellent results. We apply our approach to examples of system sizing and our examples clearly demonstrate the importance of taking into account the variability of the tasks arrivals and thus expose the risk of under- or over-provisioning if one relies on a model with Poisson assumptions [8].

Prediction of the System Performance from components models. In this paper we consider the problem of combining calibrated performance models of system components in order to predict overall system performance. We focus on open workload system models, in which, under certain conditions, obtaining and validating the overall system performance measures can be a simple application of Little's law. We discuss the conditions of applicability of such a simple validation methodology, including examples of successful application, as well as examples where this approach fails. Additionally, we propose to analyze the deviations between the model predictions and system measurements, so as to decide if they correspond to "measurement noise" or if an important system component has not been correctly represented. This approach can be used as an aid in the design of validated system performance models [26].

7.3. Modeling of Dynamics of Complex Networks

Participants: Christophe Crespelle, Éric Fleury, Márton Karsai, Yannick Leo, Matteo Morini.

Non-Altering Time Scales for Aggregation of Dynamic Networks into Series of Graphs [29] Many dynamic networks coming from real-world contexts are *link streams*, i.e. a finite collection of triplets (u, v, t) where u and v are two nodes having a link between them at time t. A great number of studies on these objects start by aggregating the data on disjoint time windows of length Δ in order to obtain a series of graphs on which are made all subsequent analyses. Here we are concerned

with the impact of the chosen Δ on the obtained graph series. We address the fundamental question of knowing whether a series of graphs formed using a given Δ faithfully describes the original link stream. We answer the question by showing that such dynamic networks exhibit a threshold for Δ, which we call the *saturation scale*, beyond which the properties of propagation of the link stream are altered, while they are mostly preserved before. We design an automatic method to determine the saturation scale of any link stream, which we apply and validate on several real-world datasets.

Termination of the Iterated Strong-Factor Operator on Multipartite Graphs [10] The clean-factor operator is a multipartite graph operator that has been introduced in the context of complex network modelling. Here, we consider a less constrained variation of the clean-factor operator, named strong-factor operator, and we prove that, as for the clean-factor operator, the iteration of the strong-factor operator always terminates, independently of the graph given as input. Obtaining termination for all graphs using minimal constraints on the definition of the operator is crucial for the modelling purposes for which the clean-factor operator has been introduced. Moreover we show that the relaxation of constraints we operate not only preserves termination but also preserves the termination time, in the sense that the strong-factor series always terminates before the clean-factor series.

On the Termination of Some Biclique Operators on Multipartite Graphs [9] We define a new graph operator, called the *weak-factor graph*, which comes from the context of complex network modelling. The weak-factor operator is close to the well-known clique-graph operator but it rather operates in terms of bicliques in a multipartite graph. We address the problem of the termination of the series of graphs obtained by iteratively applying the weak-factor operator starting from a given input graph. As for the clique-graph operator, it turns out that some graphs give rise to series that do not terminate. Therefore, we design a slight variation of the weak-factor operator, called *clean-factor*, and prove that its associated series terminates for all input graphs. In addition, we show that the multipartite graph on which the series terminates has a very nice combinatorial structure: we exhibit a bijection between its vertices and the chains of the inclusion order on the intersections of the maximal cliques of the input graph.

Directed Cartesian-Product Decomposition [11]. In this paper, we design an algorithm that, given a directed graph G and the Cartesian-product decomposition of its underlying undirected graph $\widetilde{G}$, produces the directed Cartesian-product decomposition of G in linear time. This is the first time that the linear complexity is achieved for this problem, which has two major consequences. Firstly, it shows that the directed and undirected versions of the Cartesian-product decomposition of graphs are linear-time equivalent problems. And secondly, as there already exists a linear-time algorithm for solving the undirected version of the problem, combined together, it provides the first linear-time algorithm for computing the directed Cartesian-product decomposition of a directed graph.

An $O(n^2)$ time Algorithm for the Minimal Permutation Completion Problem [28] We provide an $O(n^2)$ time algorithm computing a minimal permutation completion of an arbitrary graph $G = (V, E)$, i.e., a permutation graph $H = (V, F)$ on the same vertex set, such that $E \subseteq F$ and F is inclusion-minimal among all possibilities.

Linearity is Strictly More Powerful than Contiguity for Encoding Graphs [27] Linearity and contiguity are two parameters devoted to graph encoding. Linearity is a generalisation of contiguity in the sense that every encoding achieving contiguity k induces an encoding achieving linearity k, both encoding having size $\Theta(k.n)$, where n is the number of vertices of G. In this paper, we prove that linearity is a strictly more powerful encoding than contiguity, i.e. there exists some graph family such that the linearity is asymptotically negligible in front of the contiguity. We prove this by answering an open question asking for the worst case linearity of a cograph on n vertices: we provide an $O(\log n / \log \log n)$ upper bound which matches the previously known lower bound.

Socioeconomic correlations in communication networks [37], [38] In this work we study the socioeconomic structure of a communication network by combining mobile communication records and bank credit informations of a large number of individuals living in Mexico. We provide empirical evidences about present economic unbalances suggesting not only the distribution of wealth but also

the distribution of debts to follow the Pareto principle. Further we study the internal and interconnected structure of socioeconomic groups. Through a weighted core analysis we signal assortative correlations between people regarding their economic capacities, and show the existence of "rich-clubs" indicating present social stratification in the social structure. This project is ongoing with final results expected in 2016.

Detecting global bridges in networks [15] The identification of nodes occupying important positions in a network structure is crucial for the understanding of the associated real-world system. Usually, betweenness centrality is used to evaluate a node capacity to connect different graph regions. However, we argue here that this measure is not adapted for that task, as it gives equal weight to "local" centers (i.e. nodes of high degree central to a single region) and to "global" bridges, which connect different communities. This distinction is important as the roles of such nodes are different in terms of the local and global organisation of the network structure. In this paper we propose a decomposition of betweenness centrality into two terms, one highlighting the local contributions and the other the global ones. We call the latter bridgeness centrality and show that it is capable to specifically spot out global bridges. In addition, we introduce an effective algorithmic implementation of this measure and demonstrate its capability to identify global bridges in air transportation and scientific collaboration networks.

Collective attention in the age of (mis)information [17] We study, on a sample of 2.3 million individuals, how Facebook users consumed different information at the edge of political discussion and news during the last Italian electoral competition. Pages are categorized, according to their topics and the communities of interests they pertain to, in a) alternative information sources (diffusing topics that are neglected by science and main stream media); b) online political activism; and c) main stream media. We show that attention patterns are similar despite the different qualitative nature of the information, meaning that unsubstantiated claims (mainly conspiracy theories) reverberate for as long as other information. Finally, we categorize users according to their interaction patterns among the different topics and measure how a sample of this social ecosystem (1279 users) responded to the injection of 2788 false information posts. Our analysis reveals that users which are prominently interacting with alternative information sources (i.e. more exposed to unsubstantiated claims) are more prone to interact with false claims.

The Scaling of Human Contacts in Reaction-Diffusion Processes [22] We present new empirical evidence, based on millions of interactions on Twitter, confirming that human contacts scale with population sizes. We integrate such observations into a reaction-diffusion metapopulation framework providing an analytical expression for the global invasion threshold of a contagion process. Remarkably, the scaling of human contacts is found to facilitate the spreading dynamics. Our results show that the scaling properties of human interactions can significantly affect dynamical processes mediated by human contacts such as the spread of diseases, and ideas.

From calls to communities: a model for time varying social networks [16] Social interactions vary in time and appear to be driven by intrinsic mechanisms, which in turn shape the emerging structure of the social network. Large-scale empirical observations of social interaction structure have become possible only recently, and modelling their dynamics is an actual challenge. Here we propose a temporal network model which builds on the framework of activity-driven time-varying networks with memory. The model also integrates key mechanisms that drive the formation of social ties - social reinforcement, focal closure and cyclic closure, which have been shown to give rise to community structure and the global connectedness of the network. We compare the proposed model with a real-world time-varying network of mobile phone communication and show that they share several characteristics from heterogeneous degrees and weights to rich community structure. Further, the strong and weak ties that emerge from the model follow similar weight-topology correlations as real-world social networks, including the role of weak ties.

Kinetics of Social Contagion [21] Diffusion of information, behavioural patterns or innovations follows diverse pathways depending on a number of conditions, including the structure of the underlying social network, the sensitivity to peer pressure and the influence of media. Here we study

analytically and by simulations a general model that incorporates threshold mechanism capturing sensitivity to peer pressure, the effect of 'immune' nodes who never adopt, and a perpetual flow of external information. While any constant, non-zero rate of dynamically-introduced innovators leads to global spreading, the kinetics by which the asymptotic state is approached show rich behaviour. In particular we find that, as a function of the density of immune nodes, there is a transition from fast to slow spreading governed by entirely different mechanisms. This transition happens below the percolation threshold of fragmentation of the network, and has its origin in the competition between cascading behaviour induced by innovators and blocking of adoption due to immune nodes. This change is accompanied by a percolation transition of the induced clusters.

DIANA Project-Team

6. New Results

6.1. Service Transparency

6.1.1. From Network-level Measurements to Expected QoE: the Skype Use Case

Participants: Thierry Spetebroot, Nicolas Aguilera, Damien Saucez and Chadi Barakat.
Modern Internet applications rely on rich multimedia contents making the quality of experience (QoE) of end users sensitive to network conditions. Several models were developed in the literature to express QoE as a function of measurements carried out on the traffic of the applications themselves. In this contribution, we propose a new methodology based on machine learning able to link expected QoE to network and device level measurements outside the applications' traffic. This direct linking to network and device level measurements is important for the prediction of QoE. We prove the feasibility of the approach in the context of Skype. In particular, we derive and validate a model to predict the Skype QoE as a function of easily measurable network performance metrics. One can see our methodology as a new way of performing measurements in the Internet, where instead of expressing the expected performance in terms of network and device level measurements that only specialists can understand, we express performance in clear terms related to expected quality of experience for different applications. More details on this approach and on our application ACQUA can be found in section 5.1 , in the paper summarizing the results [16] and on the application web page http://team.inria.fr/diana/acqua/.

6.1.2. Towards a General Solution for Detecting Traffic Differentiation at the Internet Access

Participants: Ricardo Ravaioli and Chadi Barakat.
In recent years network neutrality has been widely debated from both technical and economic points of view. Various cases of traffic differentiation at the Internet access have been reported throughout the last decade, in particular aimed at bandwidth consuming traffic flows. In this contribution we present a novel application-agnostic method for the detection of traffic differentiation, through which we are able to correctly identify where a shaper is located with respect to the user and evaluate whether it affected delays, packet losses or both. The tool we propose, ChkDiff, replays the user's own traffic in order to target routers at the first few hops from the user. By comparing the resulting flow delays and losses to the same router against one other, and analyzing the behaviour on the immediate router topology spawning from the user end-point, ChkDiff manages to detect instances of traffic shaping. This contribution is published in [15] where we provide a detailed description of the design of the tool for the case of upstream traffic, the technical issues it overcomes and a validation in controlled scenarios. It is the result of collaboration with the SIGNET group at I3S in the context of a PhD thesis funded by the UCN@SOPHIA Labex.

6.1.3. A Diagnostic Tool for Content-Centric Networks

Participant: Thierry Turletti
In collaboration with our colleagues at NICT, Japan, we have proposed the Contrace tool for Measuring and Tracing Content-Centric Networks (CCNs). CCNs are fundamental evolutionary technologies that promise to form the cornerstone of the future Internet. The information flow in these networks is based on named data requesting, in-network caching, and forwarding – which are unique and can be independent of IP routing. As a result, common IP-based network tools such as ping and traceroute can neither trace a forwarding path in CCNs nor feasibly evaluate CCN performance. We designed "contrace," a network tool for CCNs (particularly, CCNx implementation running on top of IP) that can be used to investigate 1) the Round-Trip Time (RTT) between content forwarder and consumer, 2) the states of in-network cache per name prefix, and 3) the forwarding path information per name prefix. We report a series of experiments conducted using contrace on a CCN topology created on a local testbed and the GEANT network topology emulated by the Mini-CCNx emulator.

The results confirm that contrace is not only a useful tool for monitoring and operating a network, but also a helpful analysis tool for enhancing the design of CCNs. Further, contrace can report the number of received interests per cache or per chunk on the forwarding routers. This enables us to estimate the content popularity and design more effective cache control mechanisms in experimental networks (see our publication in the IEEE Communication Magazine [9]).

6.1.4. An efficient packet extraction tool for large experimentation traces

Participants: Thierry Turletti and Walid Dabbous

Network packet tracing has been used for many different purposes during the last few decades, such as network software debugging, networking performance analysis, forensic investigation, and so on. Meanwhile, the size of packet traces becomes larger, as the speed of network rapidly increases. Thus, to handle huge amounts of traces, we need not only more hardware resources, but also efficient software tools. However, traditional tools are inefficient at dealing with such big packet traces. In this work, we propose pcapWT, an efficient packet extraction tool for large traces. PcapWT provides fast packet lookup by indexing an original trace using a Wavelet Tree structure. In addition, it supports multi-threading for avoiding synchronous I/O and blocking system calls used for file processing, and it is particularly efficient on machines with SSD disks. PcapWT shows remarkable performance enhancements in comparison with traditional tools such as tcpdump and most recent tools such as pcapIndex in terms of index data size and packet extraction time. Our benchmark using large and complex traces shows that pcapWT reduces the index data size down below 1% of the volume of the original traces. Moreover, packet extraction performance is 20% better than with pcapIndex. Furthermore, when a small amount of packets are retrieved, pcapWT is hundreds of times faster than tcpdump. This work has been done in collaboration with our colleagues at Universidad Diego Portales (UDP) and Universidad de Chile and has been published in the Computer Networks journal [10].

6.1.5. Social Clicks: What and Who Gets Read on Twitter?

Participants: Maksym Gabielkov and Arnaud Legout

Online news domains increasingly rely on social media to drive traffic to their website. Yet we know surprisingly little about how social media conversation mentioning an online article actually generates a click to it. Posting behaviors, in contrast, have been fully or partially available and scrutinized over the years. While this has led to multiple assumptions on the diffusion of information, each were designed or validated while ignoring this important step. We made a large scale, validated and reproducible study of social clicks, that is also the first data of its kind, gathering a month of web visits to online resources that are located in 5 leading news domains and that are mentioned in the third largest social media by web referral (Twitter). Our dataset amounts to 2.8 million posts, together responsible for 75 billion potential views on this social media, and 9.6 million actual clicks to 59,088 unique resources. We design a reproducible methodology, carefully corrected its biases, enabling data sharing, future collection and validation. As we prove, properties of clicks and social media Click-Through-Rates (CTR) impact multiple aspects of information diffusion, all previously unknown. Secondary resources, that are not promoted through headlines and are responsible for the long tail of content popularity, generate more clicks both in absolute and relative terms. Social media attention is actually long-lived, in contrast with temporal evolution estimated from posts or impressions. The actual influence of an intermediary or a resource is poorly predicted by their posting behavior, but we show how that prediction can be made more precise. The results are reported in an article under submission, no report available yet.

6.1.6. ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic

Participant: Arnaud Legout

It is well known that apps running on mobile devices extensively track and leak users' personally identifiable information (PII); however, these users have little visibility into PII leaked through the network traffic generated by their devices, and have poor control over how, when and where that traffic is sent and handled by third parties. In this work, we present the design, implementation, and evaluation of ReCon: a cross-platform system that reveals PII leaks and gives users control over them without requiring any special privileges or custom OSes. ReCon leverages machine learning to reveal potential PII leaks by inspecting network traffic, and provides a visualization tool to empower users with the ability to control these leaks via blocking or

substitution of PII. We evaluate ReCon's effectiveness with measurements from controlled experiments using leaks from the 100 most popular iOS, Android, and Windows Phone apps, and via an user study with 92 participants. In this study, that was approved by the Inria Ethical Board (COERELE), we show that ReCon is accurate, efficient, and identifies a wider range of PII than previous approaches. The results are reported in an article under submission, no report available yet.

6.2. Open Network Architecture

6.2.1. Storage on Wheels: Offloading Popular Contents Through a Vehicular Cloud

Participants: Luigi Vigneri and Chadi Barakat.
The increasing demand for mobile data is overloading the cellular infrastructure. Small cells and edge caching is being explored as an alternative, but installation and maintenance costs for sufficient coverage are significant. In this work, we perform a preliminary study of an alternative architecture based on two main ideas: (i) using vehicles as mobile caches that can be accessed by user devices; compared to small cells, vehicles are more widespread and require lower costs; (ii) combining the mobility of vehicles with delayed content access to increase the number of cache hits (and reduce the load on the infrastructure). Contrary to standard DTN-type approaches, in our system max delays are guaranteed to be kept to a few minutes (beyond this deadline, the content is fetched from the infrastructure). We first propose an analytical framework to compute the optimal number of content replicas that one should cache, in order to minimize the infrastructure load. We then investigate how to optimally refresh these caches to introduce new contents, as well as to react to the temporal variability in content popularity. Simulations suggest that our vehicular cloud considerably reduces the infrastructure load in urban settings, assuming modest penetration rates and tolerable content access delays. This work is currently under submission; it is the result of collaboration with the Mobile Communications Department at Eurecom in the context of a PhD thesis funded by the UCN@SOPHIA Labex.

6.2.2. Geographically Fair In-Network Caching for Mobile Data Offloading

Participant: Chadi Barakat
Data offloading from the cellular network to low-cost WiFi has been the subject of several research works in the last years. In-network caching has also been studied as an efficient means to further reduce cellular network traffic. In this contribution, done jointly with the Maestro project-team, we consider a scenario where mobile users can download popular contents (e.g., maps of a city, shopping information, social media, etc.) from WiFi-enabled caches deployed in an urban area. We study the optimal distribution of contents among the caches (i.e., what contents to put in each cache) to minimize users' access cost in the whole network. We argue that this optimal distribution does not necessarily provide geographic fairness, i.e., users at different locations can experience highly variable performance. In order to mitigate this problem, we propose two different cache coordination algorithms based on gossiping. These algorithms achieve geographic fairness while preserving the minimum access cost for end users. More details on this contribution can be found in [12].

6.2.3. Virtual Service Providers (vSP)

Participant: Damien Saucez
The ability of SOHO networks to connect to the Internet through several Internet service providers, gives high potential to enable rich cloud-based network services for enterprises. Nevertheless, it remains a huge challenge for SOHOs to leverage such multi-homing and cloud networking capabilities. For such a reason, we introduced the vSP concept (virtual Service Provider). The idea of vSP is to hide the technical complexity inherent to multi-homing and allow SOHOs to seamlessly use their cloud resources. The role of the vSP is to orchestrate traffic between the different Internet Services Providers (ISPs) in order to maximize the cloud service performance without requiring any intervention of the SOHO network administrator. This ongoing work is done in collaboration with Telecom ParisTech, Ericsson, LISPERS.net, and Cisco Systems and is presented in two papers [19], [20] and detailed in one IETF Internet-draft [19].

6.2.4. Rules Placement Problem in OpenFlow Networks

Participants: Xuan Nam Nguyen, Damien Saucez, Chadi Barakat and Thierry Turletti

Software-Defined Networking (SDN) abstracts low-level network functionalities to simplify network management and reduce costs. The OpenFlow protocol implements the SDN concept by abstracting network communications as flows to be processed by network elements. In OpenFlow, the high-level policies are translated into network primitives called rules that are distributed over the network. While the abstraction offered by OpenFlow allows to potentially implement any policy, it raises the new question of how to define the rules and where to place them in the network while respecting all technical and administrative requirements. We proposed a comprehensive study of the so-called OpenFlow rules placement problem with a survey of the various proposals intending to solve it [11] and developed an offline optimization framework for this problem with a polynomial time approximation in [13].

6.3. Experimental Evaluation

6.3.1. Automating ns-3 Experimentation in Multi-Host Scenarios

Participants: Alina Ludmila Quereilhac, Damien Saucez, Thierry Turletti and Walid Dabbous
ns-3 is a flexible simulator whose capabilities go beyond running purely synthetic simulations in a local desktop. Due to its ability to run unmodified Linux applications, to execute in real time mode, and to exchange traffic with live networks, ns-3 can be combined with live hosts to run distributed simulations or to transparently integrate live and simulated networks. Nevertheless, setting up ns-3 multi-host experiment scenarios might require considerable manual work and advanced system administration skills. The NEPI experiment management framework is capable of automating deployment, execution, and result collection of experiment scenarios that combine ns-3 with multiple hosts in various ways, reducing the burden of manual scenario set up. We proved that this approach can be used to seamlessly running parallel simulations on a cluster of hosts, running distributed simulation spanning multiple hosts, and integrating live and simulated networks. This work has been published in [18] and has been awarded as the best paper of the workshop.

6.3.2. DiG: Emulating Data Centers and Cloud Architectures in a Grid Network

Participants: Hardik Soni, Thierry Turletti, Damien Saucez
We are witnessing a considerable amount of research work related to data-center and cloud infrastructures but evaluations are often limited to small scale scenarios as very few researchers have access to a real infrastructure to confront their ideas to reality. We have designed an experiment automation tool, called DiG (Data-centers in the Grid), which explicitly allocates physical resources in grids to emulate data-center and cloud networks. DiG allows one to utilize grid infrastructures to evaluate research ideas pertaining to data-centers and cloud environments at massive scale and with real traffic workload. We have automated the procedure of building target network topologies while respecting effective performance capacity of available physical resources in the grid against the demand of links and hosts in the experiment. We demonstrate a showcase where DiG automatically builds a large data-center topology composed of hundreds of servers executing various Hadoop intensive workloads (see our demo abstract at IEEE NFV/SDN 2015 in [24]).

DIONYSOS Project-Team

6. New Results

6.1. Quality of Experience

Participants: Yassine Hadjadj-Aoul, Gerardo Rubino.

QoE in mobile networks. We consider in [43] an important Quality of Experience (QoE) indicator in cellular networks that is reneging of users due to impatience. We specifically consider a cell under heavy load conditions, modeled as a multiclass Processor Sharing system, and compute the reneging probability by using a fluid limit analysis. In order to enhance the user QoE, we propose a radio resource allocation control scheme that minimizes the global reneging rates. This control scheme is based on the α-fair scheduling framework and adapts the scheduler parameter depending on the traffic load. While the proposed scheme is simple, our results show that it achieves important performance gains. This work is extended in [42]. By solving the fixed point equation, we obtain a new QoE perturbation metric quantifying the impact of reneging on the performance of the system. This metric is then used to devise a new pricing scheme accounting of reneging. We specifically propose several flavors of this scheme around the idea of having a flat rate for accessing the network and an elastic price related to the level of QoE perturbation induced by communications.

In order to offer a high media quality and a good user satisfaction, the media streaming service requires that transport protocols can be adapted continuously to the network parameters. However, the diversity of terminals (e.g., tablets, smart phones, laptops) and their corresponding capabilities, mean that users' agnostic solutions are inefficient to cope with such diverse contexts. Indeed, the intrinsic characteristics and parameters of the terminals (i.e., devices) need to be taken into account on the video streaming adaptation process. In [17], we propose an adaptive video streaming solution to improve the user satisfaction factor by adapting the TCP parameters according to the user's parameters on mobile networks. The user satisfaction factor is calculated according to some metrics driven from the user's quality of experience (QoE). The work is validated through our proposal based on a new mobile agent developed on a Linux script platform and tested on different kinds of devices with different scenarios.

Learning tools. Our QoE measuring techniques (see 3.2) are based on statistical learning methods, and we have been using Random Neural Networks as our main learning tool. These are actually open queueing networks where customers have a "sign" and behave analogously as neural spiking signals. They have been proposed by Gelenbe in the 80s, and have been used in many areas since then. In [26], we published a survey about the tool, where we develop in some detail their use in supervised learning, not only for the case of interest in PSQA, our QoE measuring technology. We also discuss the use of powerful optimization methodology, first and second order techniques, that have proved to be very effective in the standard Neural Network area.

Recently, we started to explore new learning techniques. The first reason is not the search for more accurate tools, because ours are, we claim, as accurate as they can be, it is to improve robustness. The second reason is to extend our QoE measuring tools to richer contexts, mainly when we take into account time, that is, time series data. This comes from the observation that in many cases, the way people perceive quality has some "inertia" and depends on the quality perceived some minutes ago. In [66] we explored the capabilities of a recently proposed method called "Reservoir Computing (RC) with Random Static Projections" which combines two ideas, the now classic Reservoir Computing approach and Extreme Learning Machines (ELMs). In our paper, we replaced the ELMs by Radial Basis Functions (RBF) projections. We illustrated the good behavior of this variation of the original technique basically using known benchmarks.

In [67], we perform a detailed analysis of one of the main instances of the Reservoir Computing idea, called Echo State Network (ESN). This type of model has several parameters to adjust, that have an impact on the performances of the learning procedure. For instance, it has been shown that the spectral radius of the reservoir matrix (the recurrent network structure that doesn't learn during the process) is related to the accuracy and the memory capabilities of the technology. The size of the reservoir is also a parameter to adjust when configuring

an ESN for performing some specific task. One of the results of our work is the fact that the periodic or pseudo-period nature of data is also an important factor to be taken into account when designing an ESN, since it has an influence on the impact of parameters such as the previously mentioned spectral radius.

QoE and emergency management. As a by-product of our activities around QoE, we started to work on an application where, instead of evaluating the QoE of, say, a video or voice application, we wanted to evaluate the way users perceive a service not necessarily based on audio or video content. This was related to our participation to the European project QuEEN (see 8.2.2). We finished by building a platform where we test different ideas for managing an emergency situation. In our system, we include an automatic evaluator of the perceived quality of the related voice and video communications, since in the case of some catastrophes, the communications can be seriously damaged and it is critical to automatically detect the issue in order to report the problem and to take appropriate countermeasures, when possible. In [55], we describe some of the aspects of our system and of the implemented mechanisms, and we present some design problems and their solutions, together with illustrations of the capabilities of the tool.

6.2. Analytic models

Participants: Gerardo Rubino, Bruno Sericola.

Sojourn times in Markovian models. In [74], we discuss different issues related to the time a Markov chain spends in a part of its state space. This is relevant in many application areas including those interesting Dionysos, namely, performance and dependability analysis of complex systems. For instance, in dependability, the reliability of a system subject to failures and repairs of its components, is, in terms of a discrete-space model of it, the probability that it remains in the subset of operational or up states during the whole time interval $[0, t]$. In performance, the occupancy factor of some server is the probability that, in steady state, the model belongs to the subset of states where the server is busy. This book chapter reviews some past work done by the authors on this topic, and add some new insights on the properties of these sojourn times.

Queuing systems in equilibrium. In the late 70s, Leonard Kleinrock proposed a metric able to capture the tradeoff between the work done by a system and its cost, or, in terms of queueing systems, between throughput and mean response time. The new metric was called *power* and among its properties, it satisfies a nice one informally called "keep the pipe full", specifying that the operation point of some queues (mainly the $M/M/1$ one) giving the maximal possible value to the power is when the mean backlog is 1. In [56], we took back this idea to explore what happens when we consider Jackson queuing networks. After showing that the same property holds for them and exploring other ones, we show that the power metric has some drawbacks when considering multiserver queues and networks of queues. We then propose a new metric that we called *effectiveness*, identical to power when there is a single queue with a single server, but different otherwise, that avoids these drawbacks. We analyze it and, in particular, we show that the same "keep the pipe full" holds for it.

Transient analysis of queuing systems. In a well-known book [86], today out of press, a concept of dual of a birth-and-death process is proposed, based on stochastic monotonicity. In past work [88] we showed that this concept coupled with the classical randomization or uniformization of continuous time Markov chains and lattice path combinatorics, allowed to derive analytical expressions of the transient distribution of several Markovian queueing systems. Recently, we discovered two new things: first, that this dual concept can be generalized to arbitrary systems of ordinary differential equations (ODEs) and still keep its main properties; second, that we can define a similar transformation than uniformization, that can be applied to arbitrary systems of ODEs and again, holding similar properties than the former. We respectively called pseudo-dual and pseudo-randomization the two concepts and associated methods. In [69], we presented these ideas and first results about them. We illustrated their use, and how they allow to obtain analytical expressions of transient queues' distributions in cases where Anderson's dual doesn't exist (see [87].

In [68], we present results concerning some aspects of the behavior of a queueing system observed during a fixed time period of the form $[0, t]$. The two aspects we looked at in this work are the loss process of a finite capacity model during the considered $[0, t]$, and the maximal backlog reached at a queue over the interval.

Following the classical procedure mentioned below, consisting in using uniformization to go to discrete time and then, combinatorial techniques, we develop numerical schemes to analyze both aspects of some basic queueing systems.

Network reliability. In [28], we consider the classical network design "Capacitated m-Ring Star Problem" (CmRSP), where we look for m rings connecting two nodes in a network at minimum cost. We add to this model the fact that links can fail, and propose a new paradigm that we call "Capacitated m-Ring Star Problem with Diameter Constrained Reliability " (in short, CmRSP-DCR), where we look again for a minimal cost spanning graph of the set of nodes in the network that connects the selected source and terminal, *while satisfying a Diameter Constrained Reliability (DCR) condition*. The DCR is the probability that the two nodes can communicate by means of paths having lengths bounded by some fixed value d. We prove that this problem is NP-hard, and we propose a GRASP-based approach to solve it.

Fluid models. In [19] we study congestion periods in a finite fluid buffer when the net input rate depends upon a recurrent Markov process; congestion occurs when the buffer content is equal to the buffer capacity. We consider the duration of congestion periods as well as the associated volume of lost information. We derive their distributions in a typical stationary busy period of the buffer. Our goal is to compute the exact expression of the loss probability in the system, which is usually approximated by the probability that the occupancy of the infinite buffer is greater than the buffer capacity under consideration. Moreover, by using general results of the theory of Markovian arrival processes, we show that the duration of congestion and the volume of lost information have phase-type distributions.

6.3. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Yann Busnel, Pierre L'Ecuyer.

Detection of distributed deny of service attacks. A Deny of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Deny of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term "iceberg" has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs [36]. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modeled as items) prior to sending them to the server, and throw away all the items that appear with a small probability. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the time needed to collect c distinct items among n frequent ones. A thorough analysis of the time needed to collect c distinct items appears in [12], [11].

Stream Processing Systems. Stream processing systems are today gaining momentum as tools to perform analytics on continuous data streams. Their ability to produce analysis results with sub-second latencies, coupled with their scalability, makes them the preferred choice for many big data companies.

A stream processing application is commonly modeled as a direct acyclic graph where data operators, represented by nodes, are interconnected by streams of tuples containing data to be analyzed, the directed edges (the arcs). Scalability is usually attained at the deployment phase where each data operator can be parallelized using multiple instances, each of which will handle a subset of the tuples conveyed by the operators' ingoing stream. Balancing the load among the instances of a parallel operator is important as it yields to better resource

utilization and thus larger throughputs and reduced tuple processing latencies. We have proposed a new key grouping technique targeted toward applications working on input streams characterized by a skewed value distribution [53]. Our solution is based on the observation that when the values used to perform the grouping have skewed frequencies, the few most frequent values (the *heavy hitters*) drive the load distribution, while the remaining largest fraction of the values (the *sparse items*) appear so rarely in the stream that the relative impact of each of them on the global load balance is negligible. We have shown, through a theoretical analysis, that our solution provides on average near-optimal mappings using sub-linear spaces in the number of tuples read from the input stream in the learning phase and the support (value domain) of the tuples. In particular this analysis presents new results regarding the expected error made on the estimation of the frequency of heavy hitters.

Randomized Message-Passing Test-and-Set. In [37], we have presented a solution to the well-known Test&Set operation in an asynchronous system prone to process crashes. Test&Set is a synchronization operation that, when invoked by a set of processes, returns yes to a unique process and returns no to all the others. Recently, many advances in implementing Test&Set objects have been achieved. However, all of them target the shared memory model. In this paper we propose an implementation of a Test&Set object in the message passing model. This implementation can be invoked by any number $p \leq n$ of processes where n is the total number of processes in the system. It has an expected individual step complexity in $O(\log p)$ against an oblivious adversary, and an expected individual message complexity in $O(n)$. The proposed Test&Set object is built atop a new basic building block, called selector, that allows to select a winning group among two groups of processes. We propose a message-passing implementation of the selector whose step complexity is constant. We are not aware of any other implementation of the Test&Set operation in the message passing model.

Population Protocol Model. The population protocol model, introduced by Angluin and his colleagues in 2006, provides theoretical foundations for analyzing global properties emerging from pairwise interactions among a large number of anonymous agents. In the population protocol model, agents are modeled as identical and finite state machines, i.e each agent can be in a finite number of states while waiting to execute a transition. When two agents interact, they communicate their local state, and can move from one state to another according to a transition function. The ultimate goal of population protocols is for all the agents to converge to the same value. Examples of systems whose behavior can be modeled by population protocols range from molecule interactions of a chemical process to sensor networks in which agents, which are small devices embedded for instance in animals, interact each time two animals are in the same radio range.

In this work, we focus on a quite important related question. Namely, is there a population protocol that exactly counts the difference κ between the number of agents that initially set their state to A and the one that initially set it to B, and can it be solved in an efficient way, that is with the guarantee that each agent should converge to the exact value of κ after having triggered a sub-linear number of interactions in the size of the system [49]?. We answer this question by the affirmative by presenting a $O(n^{3/2})$-state population protocol that allows each agent to converge to the exact solution by interacting no more than $O(\log n)$ times. The proposed protocol is very simple (as is true for most known population protocols), but is general enough to be used to solve different types of tasks.

Call centers. We develop research activities around the analysis and design of call centers, from a performance perspective. The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty. Among important sources of uncertainty are call arrival rates which are typically time-varying, stochastic, dependent across time periods and across call types, and often affected by external events. Accurately modeling and forecasting future call arrival volumes is a complicated issue which is critical for making important operational decisions, such as staffing and scheduling, in the call center. In [20] we review the existing literature on modeling and forecasting call arrivals. We also develop in [58] customer delay predictors for multi-skill call centers that take as inputs the queueing state upon arrival and the waiting time of the last customer served. Barely any predictor currently exists for the multi-skill case. We introduce two new predictors that use cubic regression splines and artificial neural networks, respectively, and whose parameters are optimized (or learned) from observation data obtained by simulation.

6.4. Wireless Networks

Participants: Osama Arouk, Btissam Er-Rahmadi, Adlen Ksentini, Meriem Bouzouita, Pantelis Frangoudis, Yassine Hadjadj-Aoul, Gerardo Rubino.

We are continuing our activities around wireless and mobile networks, by focusing more on leveraging the current mobile and wireless architecture toward building the 5G systems.

LTE improvements. One of the 5G objectives is to support a high number of devices. This not only concerns User Equipment (UE) devices, but also other devices such as sensors and actuators (known also as Internet of Things (IoT)). Sensor and actuator devices communicate generally with a remote server in an automatic way, without any human intervention. This type of communication is known as Machine to Machine (M2M) communication, or Machine Type Communication (MTC). The corresponding traffic is known by its intensity and impact on increasing congestion in both main parts of 4G networks, the Radio Access Network (RAN) and the Core Network. To improve the current LTE system to support MTC, we did several contributions. We proposed in [51] an important enhancement to the Group Paging (GP) mechanism, which is responsible for relaying requests to sensors, in order to gather data. After modeling analytically the GP procedure, we proposed a mechanism that, instead of paging all MTC devices in the same period, calculates the appropriate number of MTCs that reduces the collision probability as well as increases the success probability. In [52], we modeled the Radio Access Channel (RACH) procedure when the MTC devices are activated in a highly synchronized manner during a certain period (synchronized traffic), which is represented by a Beta distribution. The proposed model estimates for each period the exact number of MTC devices that may win the contention.

To control the Random Access Network (RAN) overload and alleviate the access network congestion, 3GPP developed the Access Class Barring (ACB) procedure that depends on an access probability called the ACB factor, without proposing a procedure for calculating such probability. In [72], we have proposed a fluid-based random access model for M2M communications, which was used to determine dynamically the value of the ACB factor that avoids system overload and the radio resources' underutilization at the same time. We proposed in [60] a novel implementation of the ACB mechanism in the context of multiple M2M traffic classes. Based on a scheduling algorithm, we have applied a PID controller to adjust dynamically multiple ACB factors related to each class category, guaranteeing a number of devices around an optimal value that maximizes the Random Access (RA) success probability. In [61], we first present a simple fluid model of MTC devices' random access. This model is then used to derive a novel adaptive regulator of the ACB factor, somehow in contrast with previous existing contributions which generally rely on heuristics. The main advantages of the proposed approach are twofold. First, the proposal is fully compliant with the standard while it reduces significantly the computation and the signaling overheads. Second, it provides an efficient mean to regulate adaptively the ACB factor as it guarantees having an optimal number of MTC devices accessing concurrently to the RAN. The obtained results based on simulations show clearly the robustness of the proposed approach, and its superiority compared to existing proposals.

Another important objective of 5G mobile networks is to accommodate a diverse and ever-increasing number of user equipments (UEs). Coping with the massive signaling overhead expected from UEs is an important hurdle to tackle so as to achieve this objective. In [38], we devised an efficient tracking area list management framework that aims for finding optimal distributions of tracking areas (TAs) in the form of TA lists (TALs) and assigning them to UEs. The objective is to minimize two conflicting metrics: paging overhead and tracking area update (TAU) overhead. We used bargaining games to find the Pareto optimal solution that satisfies both objectives.

WiFi networks improvements. It is well established that WiFi is complementing LTE connections to ensure, wirelessly, high data rate. One idea to improve WiFi towards high data rates is to multiple users' transmissions on both directions, i.e. on the Down Link (DL) and the Up Link (UL). In [50] we devised a novel solution to enhance the TXOP Sharing mechanism, introduced in the 802.11ac amendment, to achieve efficient Down-Link Multi-User Multiple-Input Multiple-Output (DL-MU-MIMO) transmission. First, we give new definitions about both events of successful and failed DL-MU-MIMO transmission. Then, we devise a revised

Backoff procedure for the primary Access Category (AC). In [40] we proposed a novel 802.11ax MAC protocol aiming at reducing the elapsed time in managing the establishment of an UL-MU communication, thus enhancing considerably the system's performance.

On the other hand, the volume of mobile multimedia traffic is fast-growing, challenging the radio and backhaul network infrastructure and calling for alternative content dissemination schemes. To improve user experience and reduce infrastructure load, we exploit implicit social relationships among users and take into account content popularity, proposing push-based prefetching mechanisms which take advantage of the caching and mobile ad hoc networking capabilities of user devices. We use, in [65], bloom filters as summaries of user caches, and design mechanisms to estimate the social distance between users and the popularity of content items, which drive our algorithms. Our simulation-based evaluation shows that our scheme brings caching performance improvements in an order of 10% in terms of absolute cache hit ratio in most of the cases studied, and from 3% to 82% in terms of normalized cache hit ratio gain.

Network selection. With the explosion of mobile data traffic, the Fixed and Mobile Converged (FMC) network are being heavily required. Mobile devices have the capability of connecting to different access networks in the FMC architecture simultaneously. Access network selection becomes an issue when mobile devices are under coverage of different access networks, since a bad selection may lead to network congestion and degrade the QoE of users. In order to address this problem, we model and analyze, in [62] and [63], the interface selection procedure using control theory in the FMC architecture. Based on our model, we designed a controller which can send to mobile devices a network selection command calculated instantly for the access network selection. In [29], we investigated network decentralization in conjunction with the selective IP traffic offload approaches to handle the increased data traffic. We first devised different approaches based on a per-destination-domain-name basis, which offer operators a fine-grained control to determine whether a new IP connection should be offloaded or accommodated via the core network.

Energy efficiency. Due to the ever-growing gap between battery lifetime and hardware/software complexity in addition to application's computing power needs, the energy saving issue becomes crucial. In this context, we proposed, in [13], an end-to-end study of video decoding on different architectures. The study was achieved thanks to a two steps methodology: (1) a comprehensive characterization and evaluation of the performance and the energy consumption of video decoding, (2) an accurate high level energy model based on the characterization step. In [24], we proposed to apply data fragmentation, in slotted CSMA/CA, in a way to allow improving the bandwidth occupation while reducing the latency. We proposed to introduce a network allocation vector (NAV) in the fragmentation mechanism to reduce energy consumption in IEEE 802.15.4. A Markov chain-based analytical model of the fragmentation mechanism was given as well as an analytical model of the energy consumption using a NAV. The analytical results show that the fragmentation technique improves at the same time the throughput, the access delay and the bandwidth occupation. They also show that the NAV mechanism reduces energy consumption when applying the fragmentation technique in slotted CSMA-CA for IEEE 802.15.4.

6.5. Future networks and architectures

Participants: Adlen Ksentini, Yassine Hadjadj-Aoul, Jean-Michel Sanner.

SDN. We started an activity on Software Defined Networking (SDN), a recent idea proposed to handle network management problems. SDN are becoming an important issue with the ever-increasing network complexity. They are proposed as an alternative to the current architecture of the Internet, which cannot meet the supported services requirements such as Quality of Service/Experience (Qos/QoE), security and energy consumption. We particularly address the scalability issue by proposing in [70] an automated hierarchical controller-based architecture handling the whole control chain.

Mobile cloud. One of the 5G-architecture visions considers the usage of cloud to ease mobile networks evolution towards more flexibility and elasticity for handling resources; building the concept of carrier cloud. Software Defined Networking (SDN) and Network Function Virtualization (NFV) represent the key enabler of carrier cloud. In [57], we addressed the problem of Virtual Network Function (VNF) placement in the carrier

cloud. Indeed, we proposed a placement solution that has two main design goals: i) minimizing path between users and their respective data anchor gateways and ii) optimizing their sessions? mobility. The two design goals effectively represent two conflicting objectives that we deal with considering the mobility features and service usage behavioral patterns of mobile users, in addition to the mobile operators? cost in terms of the total number of instantiated VNFs to build a Virtual Network Infrastructure (VNI). We modeled this problem using an optimization formulation having these conflicting objectives, and then used Bargaining Game to find the Pareto optimal solution. We are continuing our improvement to the Follow Me Cloud (FMC), which was devised by our team conjointly with NEC labs. In [33], we proposed a FMC architecture that relies on PMIPv6 to handle mobility, and SDN to update the flow table of the anchor routers when a service has moved from one Data Center to another. In [10] and [32], we addressed the challenge of flow table scalability problem, which may arise in FMC to high number of mobile users. To this aim, we proposed a two-level hierarchical SDN controllers architecture in order to distribute the SDN/OpenFlow control plane. Another objective of 5G is to reduce network latency to 1ms, which will ease computation offloading. Thus, it will be possible to run applications on UE device, even if the latter has low computation capability, by offloading part of the code to a remote server. In [44], we were interested on studying the opportunities to offload part of one of the well known game engine in the literature, i.e. Unity 3D. We built a data set representing the CPU-GPU use of several games; allowing us to understand which modules might be offloaded to a remote server in the Mobile Cloud.

6.6. Network Economics

Participant: Bruno Tuffin.

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 “Telecommunication Network Economics From Theory to Applications” (P. Maillé and B. Tuffin, Cambridge U. Press, 2014.

Network neutrality. Among the topics we have particularly focused on, the network neutrality debate was a major concern in 2015. In [23], [80], [83] we recall the debate and highlight the fact that neutrality principles can be bypassed in many ways without violating the rules currently evoked in the debate. For example via Content Delivery Networks (CDNs), which deliver content on behalf of content providers for a fee, or via search engines, which can hinder competition and innovation by affecting the visibility and accessibility of content. In [23], we challenge the definition of net neutrality as it is generally discussed. Our goal there is to initiate a relevant debate for net neutrality in an increasingly complex Internet ecosystem, and to provide examples of possible neutrality rules for different levels of the delivery chain, this level separation being inspired by the OSI layer model.

As particular ways to bypass the current neutrality principles, we have particularly focused on CDNs. We for example investigate in [47] the impact of decisions made by a CDN willing to maximize its revenue through the management of cache servers. Based on a model with two network providers, we highlight that revenue-oriented management policies can affect the user-perceived quality of experience, impacting the competition among network access providers in favor of the largest one. Since this contradicts the principle underpinning network neutrality?although not with the technical net neutrality rules?we discuss the necessity to regulate CDN activity. Also, one of the main argument toward neutrality being that it favors innovation, we study in [46] the impact of CDNs" activity on other actors of the supply chain. Our findings indicate that vertically integrating a CDN helps Internet Service Providers (ISPs) collect fees from Content Providers (CPs), hence circumventing the interdiction of side payments coming from net neutrality rules. However, this outcome is socially much better in terms of user quality and innovation fostering than having separate actors providing the access and CDN services: in the latter case double marginalization (both ISP and CDN trying to get some value from the supply chain) leads to suboptimal investments in CDN storage capacities and higher prices for CPs, resulting in reduced innovation.

Another model we have developed is for understanding the behavior of some big providers actually paying side payment to ISPs while still officially in favor of neutrality. To better understand this strategical behavior, we have presented a simple model in [59] providing some insight on whether or not paying side payments for an incumbent provider is a way to create barriers to entry for competitors. It also investigates the economic consequences on all actors: incumbent and new entrant content providers, users, and the Internet Service Provider.It then describes how the side payment can be determined as a Nash bargaining solution.

Pricing access networks. Access networks in a competitive context has been a topic of research for a while. In the Internet, the data charging scheme has usually been flat rate. But more recently, especially for mobile data traffic, we have seen more diversity in the pricing offers, such as volume-based ones or cap-based ones. We study in [48] the behavior of heterogeneous users facing two offers: a volume-based one and a flat-rate one. On top of that selection, we investigate 1) the relevance for an ISP to propose the two types of offers, and optimize the corresponding prices, and 2) the existence of a solution to the pricing game when the offers come from competing providers.

Sponsored auctions. Advertisement in dedicated webpage spaces or in search engines sponsored slots is usually sold using auctions, with a payment rule that is either per view or per click. But advertisers can be both sensitive to being viewed (brand awareness effect) and being clicked (conversion into sales). In [84], we generalize the auction mechanism by including both pricing components: the displayed advertisers are charged when their ad is displayed, and pay an additional price if the ad is clicked. Applying the results for Vickrey-Clarke-Groves (VCG) auctions, we show how to compute payments to ensure incentive compatibility from advertisers as well as maximize the total value of the advertisement slot(s). We provide tight upper bounds for the loss of efficiency due to applying only pay- per-click (or pay-per-view) pricing instead of our scheme. Those bounds depend on the joint distribution of advertisement visibility and population likelihood to click on ads, and can help identify situations where our mechanism yields significant improvements. We also describe how the commonly used generalized second price (GSP) auction can be extended to this context.

6.7. Monte Carlo

Participants: Pierre L'Ecuyer, Gerardo Rubino, Bruno Tuffin.

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. However, when the events of interest are rare, simulation requires a special attention, to accelerate the occurrence of the event and get unbiased estimators of the event of interest with a sufficiently small relative variance. This is the main problem in the area. Dionysos' work focuses then on dealing with the rare event situation. For example, [39] presents an exponential tilting method for exact simulation from the truncated multivariate student-t distribution in high dimensions as an alternative to approximate Markov Chain Monte Carlo sampling.

A non-negligible part of our activity on the application of rare event simulation was about the evaluation of static network reliability models. Our paper [16] focuses on a technique known as Recursive Variance Reduction (RVR) which approaches the unreliability by recursively reducing the graph from the random choice of the first working link on selected cuts. This previously known method is shown to not verify the bounded relative error (BRE) property as reliability of individual links goes to one, i.e., the estimator is not robust in general to high reliability of links. We then propose to use the decomposition ideas of the RVR estimator in conjunction with the IS technique. Two new estimators are presented in the paper: the first one, called Balanced Recursive Decomposition estimator, chooses the first working link on cuts uniformly, while the second, called Zero-Variance Approximation Recursive Decomposition estimator, combines RVR and our zero-variance IS approximation. We show that in both cases BRE property is verified and, moreover, that a vanishing relative error (VRE) property can be obtained for the Zero-Variance Approximation RVR under specific sufficient conditions. A numerical illustration of the power of the methods is provided on several benchmark networks. In [54], we explore the use of the same powerful RVR idea, but applied in a very general

context, where the system is model by a monotone structure function. In the paper, we illustrate the approach with a very widely used model, a series of k-out-of-m modules.

In a static network reliability model one typically assumes that the failures of the components of the network are independent. This simplifying assumption makes it possible to estimate the network reliability efficiently via specialized Monte Carlo algorithms. Hence, a natural question to consider is whether this independence assumption can be relaxed, while still attaining an elegant and tractable model that permits an efficient Monte Carlo algorithm for unreliability estimation. In [14] we provide one possible answer by considering a static network reliability model with dependent link failures, based on a Marshall-Olkin copula, which models the dependence via shocks that take down subsets of components at exponential times, and propose a collection of adapted versions of permutation Monte Carlo (PMC, a conditional Monte Carlo method), its refinement called the turnip method, and generalized splitting (GS) methods, to estimate very small unreliabilities accurately under this model. The PMC and turnip estimators have bounded relative error when the network topology is fixed while the link failure probabilities converge to 0, whereas GS does not have this property. But when the size of the network (or the number of shocks) increases, PMC and turnip eventually fail, whereas GS works nicely (empirically) for very large networks, with over 5000 shocks in our examples. In [41] we focus on a method proposed by Fishman making use of bounds on the structure function describing in terms of configurations of (independent) link states if the considered nodes are connected. The bounds are based on the computation of (independent) mincuts disconnecting the set of nodes and (independent) minpaths ensuring that they are connected. We analyze here the robustness of the method when the unreliability of links goes to zero. We show that the conditions provided by Fishman are based on a bound and are therefore only sufficient, and provide more insight and examples on the behavior of the method.

PMC is an effective way of estimating the unreliability of a static network when this unreliability is very small and the network is not too large. We generalize the method in [31] to cover a wider range of applications, in which an estimation problem can be reframed in terms of the hitting time of a given set of states by a continuous-time Markov chain. The estimator is then defined as a function of the sample path of the underlying discrete time chain only, via Conditional Monte Carlo.We prove that the method gives bounded relative error for rare event probability estimation in certain settings. We show how it can be used to estimate the cumulative distribution function, or the density, or some moment of the hitting time. We provide examples for which the method can be applied and we give numerical illustrations.

Another family of models of interest in the group are the highly reliable Markovian systems, where a Markov chain models the evolution of a multicomponent system with failures and repairs of its components. In [27] we explore a new approach in the context of these models, and in the rare event case, called Conditional Monte Carlo with Intermediate Estimations (CMIE). The target are models with complex structures, where it is hard to design a good *importance function* dealing to good Importance Sampling schemes. The paper shows that the method belongs to the variance reduction family, and some examples illustrate its performances. It can be seen as a generalization of the class of splitting simulation procedures.

Finally, in Quasi-Monte Carlo (QMC), we reviewed in [64] the recent development on array-RQMC, a randomized quasi-Monte Carlo method for we had developed estimating the state distribution at each step of a Markov chain with totally ordered (discrete or continuous) state space. It can be used in particular to obtain a low-variance unbiased estimator of the expected total cost up to some random stopping time, when state-dependent costs are paid at each step. In [21], a combination of sequential MC with RQMC to accelerate convergence proposed by Gerber and Chopin is compared with our array-RQMC.

But simulation requires the use of pseudo-random generators. In [45] we provide a review of the state of the art on the design and implementation of random number generators (RNGs) for simulation, on both sequential and parallel computing environments. A general review of pseudo-random and quasi-random number generation is also provided in [73]. A tool for the generation of rank-1 lattice rules is described in [22].

DYOGENE Project-Team

7. New Results

7.1. Evaluation and optimization of the quality of service perceived by mobile users for new services in cellular networks

The goal of this thesis[1] defended in 2015 is to develop tools and methods for the evaluation of the QoS (Quality of Service) perceived by users, as a function of the traffic demand, in modern wireless cellular networks. This complex problem, directly related to network dimensioning, involves modeling dynamic processes at several time-scales, which due to their randomness are amenable to probabilistic formalization. Firstly, on the ground of information theory, we capture the performance of a single link between a base station and a user in the context of a cellular network with orthogonal channels and MIMO technology. We prove and use some lower bounds of the information-theoretic ergodic capacity of such a link, which account also for the fast channel variability caused by multi-path propagation. These bounds give robust basis for further user QoS evaluation. Next, one considers several (possibly mobile) users, arriving in the network and requesting some service from it. We consider variable (elastic) bit-rate services, in which transmissions of some amounts of data are realized in a best-effort manner, or constant bit-rate services, in which a certain transmission rate needs to be maintained during requested times. On the ground of queuing theory, one captures this traffic demand and service process using appropriate (multi-class) processor sharing (PS) or loss models. In this thesis, we adapt existing PS models and develop a new loss model for wireless streaming traffic, in which the aforementioned information-theoretic capacities of single links describe the instantaneous user service rates. The multi-class models are used to capture the spatial heterogeneity of user channels, which depends on the user geographic locations and propagation shadowing phenomenon. Finally, on top of the queueing-theoretic processes, one needs to consider a multi-cellular network, whose base stations are not necessarily regularly placed, and whose geometry is further perturbed by the shadowing phenomenon. We address this randomness aspect by using some models from stochastic geometry, notably Poisson point processes and Palm formalism applied to the typical cell of the network. Applying the above three-fold approach, supposed to represent all crucial mechanisms and engineering parameters of cellular networks (such as LTE), we establish some macroscopic relations between the traffic demand and the user QoS metrics for some elastic and constant bit-rate services. These relations are mostly obtained in a semi-analytic way, i.e., they only involve static simulations of a Poisson point process (modeling the locations of base stations) in order to evaluate its characteristics which are not amenable to analytic expressions. More precisely, regarding the data traffic (the elastic bit-rate service), we capture the inter-cell interference, making the PS queue models of individual cells dependent, via some system of cell-load equations. These equations allow one to determine the mean user throughput, the mean number of users and the mean cell load in a large network, as a function of the traffic demand. The spatial distribution of these QoS metrics in the network is also studied. We validate our approach by comparing the obtained results with those measured from live-network traces. We observe a remarkably good agreement between the model predictions and the statistical data collected in several deployment scenarios. Regarding constant bit-rate services, we propose a new stochastic model to evaluate the frequency and the number of interruptions during real-time streaming calls in function of user radio conditions. Despite some fundamental similarities with the classical Erlang loss model, a more adequate model was required for in this case, where the denial of service is not definitive for a given call: it takes the form of, hopefully short, interruptions or outage periods. Our model allows one to take into account realistic implementations of the considered streaming service. We use it to study the quality of service metrics in function of user radio conditions in LTE networks. All established results contribute to the development of network dimensioning methods and are currently used in Orange internal tools for network capacity calculations.

7.2. Interference and SINR coverage in spatial non-slotted Aloha networks

In [8] we propose two analytically tractable stochastic-geometric models of interference in ad-hoc networks using pure (non-slotted) Aloha as the medium access. In contrast the slotted model, the interference in pure Aloha may vary during the transmission of a tagged packet. We develop closed form expressions for the Laplace transform of the empirical average of the interference experienced during the transmission of a typical packet. Both models assume a power-law path-loss function with arbitrarily distributed fading and feature configurations of transmitters randomly located in the Euclidean plane according to a Poisson point process. Depending on the model, these configurations vary over time or are static. We apply our analysis of the interference to study the Signal-to-Interference-and-Noise Ratio (SINR) outage probability for a typical transmission in pure Aloha. The results are used to compare the performance of non-slotted Aloha to the slotted one, which has almost exclusively been previously studied in the same context of mobile ad-hoc networks.

7.3. Random linear multihop relaying in a general field of interferers using spatial Aloha

In [9] we study, as a basic model, a stationary Poisson pattern of nodes on a line embedded in an independent planar Poisson field of interfering nodes. Assuming slotted Aloha and the signal-to-interference-and-noise ratio capture condition, with the usual power-law path loss model and Rayleigh fading, we explicitly evaluate several local and end-to-end performance characteristics related to the nearest-neighbor packet relaying on this line, and study their dependence on the model parameters (the density of relaying and interfering nodes, Aloha tuning and the external noise power). Our model can be applied in two cases: the first use is for vehicular ad-hoc networks, where vehicles are randomly located on a straight road. The second use is to study a typical route traced in a (general) planar ad-hoc network by some routing mechanism. The approach we have chosen allows us to quantify the non-efficiency of long-distance routing in pure ad-hoc networks and evaluate a possible remedy for it in the form of additional fixed relaying nodes, called road-side units in a vehicular network. It also allows us to consider a more general field of interfering nodes and study the impact of the clustering of its nodes the routing performance. As a special case of a field with more clustering than the Poison field, we consider a Poisson-line field of interfering nodes, in which all the nodes are randomly located on random straight lines. The comparison to our basic model reveals a paradox: clustering of interfering nodes decreases the outage probability of a single (typical) transmission on the route, but increases the mean end-to-end delay.

7.4. Studying the SINR process of the typical user in Poisson networks by using its factorial moment measures

Based on a stationary Poisson point process, a wireless network model with random propagation effects (shadowing and/or fading) is considered in [7] in order to examine the process formed by the signal-to-interference-plus-noise ratio (SINR) values experienced by a typical user with respect to all base stations in the down-link channel. This SINR process is completely characterized by deriving its factorial moment measures, which involve numerically tractable, explicit integral expressions. This novel framework naturally leads to expressions for the k-coverage probability, including the case of random SINR threshold values considered in multi-tier network models. While the k-coverage probabilities correspond to the marginal distributions of the order statistics of the SINR process, a more general relation is presented connecting the factorial moment measures of the SINR process to the joint densities of these order statistics. This gives a way for calculating exact values of the coverage probabilities arising in a general scenario of signal combination and interference cancellation between base stations. The presented framework consisting of mathematical representations of SINR characteristics with respect to the factorial moment measures holds for the whole domain of SINR and is amenable to considerable model extension.

7.5. Performance laws of large heterogeneous cellular networks

In [24] we propose a model for heterogeneous cellular networks assuming a space-time Poisson process of call arrivals, independently marked by data volumes, and served by different types of base stations (having different

transmission powers) represented by the superposition of independent Poisson processes on the plane. Each station applies a processor sharing policy to serve users arriving in its vicinity, modeled by the Voronoi cell perturbed by some random signal propagation effects (shadowing). Users' peak service rates depend on their signal-to-interference-and-noise ratios (SINR) with respect to the serving station. The mutual-dependence of the cells (due to the extra-cell interference) is captured via some system of cell-load equations impacting the spatial distribution of the SINR. We use this model to study in a semi-analytic way (involving only static simulations, with the temporal evolution handled by the queuing theoretic results) network performance metrics (cell loads, mean number of users) and the quality of service perceived by the users (mean throughput) served by different types of base stations. Our goal is to identify macroscopic laws regarding these performance metrics, involving averaging both over time and the network geometry. The reveled laws are validated against real field measurement in an operational network.

7.6. Wireless networks appear Poissonian due to strong shadowing

Geographic locations of cellular base stations sometimes can be well fitted with spatial homogeneous Poisson point processes. In [6] we make a complementary observation: In the presence of the log-normal shadowing of sufficiently high variance, the statistics of the propagation loss of a single user with respect to different network stations are invariant with respect to their geographic positioning, whether regular or not, for a wide class of empirically homogeneous networks. Even in perfectly hexagonal case they appear as though they were realized in a Poisson network model, i.e., form an inhomogeneous Poisson point process on the positive half-line with a power-law density characterized by the path-loss exponent. At the same time, the conditional distances to the corresponding base stations, given their observed propagation losses, become independent and log-normally distributed, which can be seen as a decoupling between the real and model geometry. The result applies also to Suzuki (Rayleigh-log-normal) propagation model. We use Kolmogorov-Smirnov test to empirically study the quality of the Poisson approximation and use it to build a linear-regression method for the statistical estimation of the value of the path-loss exponent.

7.7. What frequency bandwidth to run cellular network in a given country? - a downlink dimensioning problem

In [25] we propose an analytic approach to the frequency bandwidth dimensioning problem, faced by cellular network operators who deploy/upgrade their networks in various geographical regions (countries) with an inhomogeneous urbanization. We present a model allowing one to capture fundamental relations between users' quality of service parameters (mean downlink throughput), traffic demand, the density of base station deployment, and the available frequency bandwidth. These relations depend on the applied cellular technology (3G or 4G impacting user peak bit-rate) and on the path-loss characteristics observed in different (urban, sub-urban and rural) areas. We observe that if the distance between base stations is kept inversely proportional to the distance coefficient of the path-loss function, then the performance of the typical cells of these different areas is similar when serving the same (per-cell) traffic demand. In this case, the frequency bandwidth dimensioning problem can be solved uniformly across the country applying the mean cell approach proposed in [Blaszczyszyn et al. WiOpt2014]. We validate our approach by comparing the analytical results to measurements in operational networks in various geographical zones of different countries.

7.8. Optimal Geographic Caching In Cellular Networks

In [23] we consider the problem of an optimal geographic placement of content in wireless cellular networks modelled by Poisson point processes. Specifically, for the typical user requesting some particular content and whose popularity follows a given law (e.g. Zipf), we calculate the probability of finding the content cached in one of the base stations. Wireless coverage follows the usual signal-to-interference-and noise ratio (SINR) model, or some variants of it. We formulate and solve the problem of an optimal randomized content placement policy, to maximize the user's hit probability. The result dictates that it is not always optimal to follow the standard policy "cache the most popular content, everywhere". In fact, our numerical results regarding three different coverage scenarios, show that the optimal policy significantly increases the chances of hit under high-coverage regime, i.e., when the probabilities of coverage by more than just one station are high enough.

7.9. Spatial distribution of the SINR in Poisson cellular networks with sector antennas

In [5] we consider a model of cellular networks where the base station locations constitute a Poisson point process and each base station is equipped with three sectorial antennas is proposed. This model permits to study the spatial distribution of the SINR in the downlink. In particular, this distribution is shown to be insensitive to the distribution of antenna azimuths. Moreover, the effect of horizontal sectorisation is shown to be equivalent to that of shadowing. Assuming ideal vertical antenna pattern, an explicit expression of the Laplace transform of the inverse of SINR is given. The model is validated by comparing its results to measurements in an operational network. It is observed numerically that, in the case of dense urban regions where interference is preponderant, one may neglect the effect of the vertical sectorization when calculating the distribution of the SINR, which provides considerable tractability. Combined with queuing theory results, the SINR's distribution permits to express the user's quality of service as function of the traffic demand. This permits in particular to operators to predict the required investments to face the continual increase of traffic demand.

7.10. Theoretical expression of link performance in OFDM cellular networks with MIMO compared to simulation and measurements

The objective of [18] is to establish a theoretical expression of the link performance in the downlink of a multiple input multiple output (MIMO) cellular network and compare it to the real Long-Term Evolution (LTE) performance. In order to account for the interference, we prove that the worst additive noise process in the MIMO context is the white Gaussian one. Based on this theoretical result, we build an analytic expression of the link performance in LTE cellular networks with MIMO. We study also the minimum mean square error (MMSE) scheme currently implemented in the field, as well as its improvement MMSE-SIC (successive interference cancellation) known to achieve the MIMO capacity. Comparison to simulation results as well as to measurements in the field shows that the theoretical expression predicts well practical link performance of LTE cellular networks. This theoretical expression of link performance is the basis of a global analytic approach to the evaluation of the quality of service perceived by the users in the long run of their arrivals and departures.

7.11. Information Theory: Boolean model in the Shannon Regime

In a paper accepted for publication in the Journal of Applied Probability, F. Baccelli and V. Anantharam consider a family of Boolean models, indexed by integers $n \geq 1$. The n-th model features a Poisson point process in $\mathbb{R}^n$ of intensity $e^{n\rho_n}$ and balls of independent and identically distributed radii distributed like $\overline{X}_n\sqrt{n}$. Assume that $\rho_n \to \rho$ as $n \to \infty$, and that $\overline{X}_n$ satisfies a large deviations principle. It is shown that there then exist three deterministic thresholds: τ_d the degree threshold; τ_p the percolation probability threshold; and τ_v the volume fraction threshold, such that asymptotically as n tends to infinity, we have the following features. (i) For $\rho < \tau_d$, almost every point is isolated, namely its ball intersects no other ball; (ii) for $\tau_d < \rho < \tau_p$, the mean number of balls intersected by a typical ball converges to infinity and nevertheless there is no percolation; (iii) for $\tau_p < \rho < \tau_v$, the volume fraction is 0 and nevertheless percolation occurs; (iv) for $\tau_d < \rho < \tau_v$, the mean number of balls intersected by a typical ball converges to infinity and nevertheless the volume fraction is 0; (v) for $\rho > \tau_v$, the whole space covered. The analysis of this asymptotic regime is motivated by problems in information theory, but it could be of independent interest in stochastic geometry. The relations between these three thresholds and the Shannon–Poltyrev threshold are discussed.

7.12. Stochastic Geometry: Wireless Modeling

In an Infocom'15 paper, F. Baccelli and X. Zhang (Qualcomm) have introduced an analytically tractable stochastic geometry model for urban wireless networks, where the locations of the nodes and the shadowing are highly correlated and different path loss functions can be applied to line-of-sight (LOS) and non-line-of-sight (NLOS) links.

Using a distance-based LOS path loss model and a blockage (shadowing)-based NLOS path loss model, one can derive the distribution of the interference observed at a typical location and the joint distribution at different locations of the network. When applied to cellular networks, this model leads to tractable coverage probabilities (SINR distribution) expressions. This model captures important features of urban wireless networks, which were difficult to analyze using existing models.

This model was lately extended in a joint work by the same authors and Robert Heath (UT Austin) in a paper presented at IEEE Globecom'15 where it received the best paper award.

7.13. Information Theory: SIMO

In a paper to be published in IEEE Transactions of Information Theory, F. Baccelli, N. Lee and Robert Heath consider large random wireless networks where transmit-and-receive node pairs communicate within a certain range while sharing a common spectrum. By modeling the spatial locations of nodes as Poisson point processes, analytical expressions for the ergodic spectral efficiency of a typical node pair are derived as a function of the channel state information available at a receiver (CSIR) in terms of relevant system parameters: the density of communication links, the number of receive antennas, the path loss exponent, and the operating signal-to-noise ratio. One key finding is that when the receiver only exploits CSIR for the direct link, the sum spectral efficiency increases linearly with the density, provided the number of receive antennas increases as a certain super-linear function of the density. When each receiver exploits CSIR for a set of dominant interfering links in addition to that of the direct link, the sum spectral efficiency in creases linearly with both the density and the path loss exponent if the number of antennas is a linear function of the density. This observation demonstrates that having CSIR for dominant interfering links provides an order gain in the scaling law. It is also shown that this linear scaling holds for direct CSIR when incorporating the effect of the receive antenna correlation, provided that the rank of the spatial correlation matrix scales super-linearly with the density. These scaling laws are derived from integral representations of the distribution of the Signal to Interference and Noise Ratio, which are of independent interest and which in turn derived from stochastic geometry and more precisely from the theory of Shot Noise fields.

7.14. Theory of point processes

In a joint work with Mir-Omid Haji-Mirsadeghi, Sharif University, Department of Mathematics, F. Baccelli studied a class of non-measure preserving dynamical systems on counting measures called point-maps. This research introduced two objects associated with a point map f acting on a stationary point process Φ:

- The f-probabilities of Φ, which can be interpreted as the stationary regimes of the action of f on Φ. These probabilities are defined from the compactification of the action of the semigroup of point-map translations on the space of Palm probabilities. The f-probabilities of Φ are not always Palm distributions.
- The f-foliation of Φ, a partition of the support of Φ which is the discrete analogue of the stable manifold of f, i.e., the leaves of the foliation are the points of Φ with the same asymptotic fate for f. These leaves are not always stationary point processes. There always exists a point-map allowing one to navigate the leaves in a measure-preserving way.

Two papers on the matter available. The first one is under revision for Annals of Probability.

7.15. Cross-Technology Interference Mitigation in Body Area Networks: An Optimization Approach

In recent years, wearable devices and wireless body area networks have gained momentum as a means to monitor people's behavior and simplify their interaction with the surrounding environment, thus representing a key element of the body-to-body networking (BBN) paradigm. Within this paradigm, several transmission technologies, such as 802.11 and 802.15.4, that share the same unlicensed band (namely, the industrial, scientific, and medical band) coexist, dramatically increasing the level of interference and, in turn, negatively

affecting network performance. In this paper, we analyze the cross-technology interference (CTI) caused by the utilization of different transmission technologies that share the same radio spectrum. We formulate an optimization model that considers internal interference, as well as CTI to mitigate the overall level of interference within the system, explicitly taking into account node mobility. We further develop three heuristic approaches to efficiently solve the interference mitigation problem in large-scale network scenarios. Finally, we propose a protocol to compute the solution that minimizes CTI in a distributed fashion. Numerical results show that the proposed heuristics represent efficient and practical alternatives to the optimal solution for solving the CTI mitigation (CTIM) problem in large-scale BBN scenarios.

7.16. Body-to-Body Area Networks

The ongoing evolution of wireless technologies has fostered the development of innovative network paradigms like the Internet of Things (IoT). Wireless Body Area Networks, and more specifically Body-to-Body Area Networks (BBNs), are emerging solutions for the monitoring of people's behavior and their interaction with the surrounding environment. These networks represent a key building block of the IoT paradigm. In BBNs several transmission technologies like 802.11 and 802.15.4 that share the same unlicensed band (namely the industrial, scientific and medical (ISM) radio band) coexist, increasing dramatically the level of interference and, in turn, negatively affecting network's performance. In [14], we investigate the Cross-Technology Interference Mitigation (CTIM) problem caused by the utilization of different transmission technologies that share the same radio spectrum, from a centralized and distributed point of view, respectively.

7.17. Exact Worst-Case Delay in FIFO-Multiplexing Feed-Forward Networks

In [11], we compute the actual worst-case end-to-end delay for a flow in a feed-forward network of FIFO-multiplexing service curve nodes, where flows are shaped by piecewise-affine concave arrival curves, and service curves are piecewise affine and convex. We show that the worst-case delay problem can be formulated as a mixed integer-linear programming problem, whose size grows exponentially with the number of nodes involved. Furthermore, we present approximate solution schemes to find upper and lower delay bounds on the worst-case delay. Both only require to solve just one linear programming problem, and yield bounds which are generally more accurate than those found in the previous work, which are computed under more restrictive assumptions.

7.18. Fast symbolic computation of the worst-case delay in tandem networks and applications

Computing deterministic performance guarantees is a defining issue for systems with hard real-time constraints, like reactive embedded systems. In [10], we use burst-rate constrained arrivals and rate-latency servers to deduce tight worst-case delay bounds in tandem networks under arbitrary multiplexing. We present a constructive method for computing the exact worst-case delay, which we prove to be a linear function of the burstiness and latencies; our bounds are hence symbolic in these parameters. Our algorithm runs in quadratic time in the number of servers. We also present an application of our algorithm to the case of stochastic arrivals and server capacities. For a generalization of the exponentially bounded burstiness (EBB) model, we deduce a polynomial-time algorithm for stochastic delay bounds that strictly improve the state-of-the-art separated flow analysis (SFA) type bounds.

7.19. Ancillary Service to the Grid Using Intelligent Deferrable Loads

Renewable energy sources such as wind and solar power have a high degree of unpredictability and time-variation, which makes balancing demand and supply challenging. One possible way to address this challenge is to harness the inherent flexibility in demand of many types of loads. Introduced in [19] is a technique for decentralized control for automated demand response that can be used by grid operators as ancillary service for maintaining demand-supply balance. A randomized control architecture is proposed, motivated by the need for decentralized decision making, and the need to avoid synchronization that can lead to large and detrimental

spikes in demand. An aggregate model for a large number of loads is then developed by examining the mean field limit. A key innovation is a linear time-invariant (LTI) system approximation of the aggregate nonlinear model, with a scalar signal as the input and a measure of the aggregate demand as the output. This makes the approximation particularly convenient for control design at the grid level.

7.20. Spectral Decomposition of Demand-Side Flexibility for Reliable Ancillary Services in a Smart Grid

[22] describes a new way of thinking about demand-side resources to provide ancillary services to control the grid. It is shown that loads can be classified based on the frequency bandwidth of ancillary service that they can offer. If demand response from loads respects these frequency limitations, it is possible to obtain highly reliable ancillary service to the grid, while maintaining strict bounds on the quality of service (QoS) delivered by each load. It is argued that automated demand response is required for reliable control. Moreover, some intelligence is needed at demand response loads so that the aggregate will be reliable and controllable.

7.21. State Estimation for the Individual and the Population in Mean Field Control with Application to Demand Dispatch

[29] concerns state estimation problems in a mean field control setting. In a finite population model, the goal is to estimate the joint distribution of the population state and the state of a typical individual. The observation equations are a noisy measurement of the population. The general results are applied to demand dispatch for regulation of the power grid, based on randomized local control algorithms. In prior work by the authors it has been shown that local control can be carefully designed so that the aggregate of loads behaves as a controllable resource with accuracy matching or exceeding traditional sources of frequency regulation. The operational cost is nearly zero in many cases. The information exchange between grid and load is minimal, but it is assumed in the overall control architecture that the aggregate power consumption of loads is available to the grid operator. It is shown that the Kalman filter can be constructed to reduce these communication requirements, and to provide the grid operator with accurate estimates of the mean and variance of quality of service (QoS) for an individual load.

7.22. Perfect sampling of Jackson queueing networks

In [12], we consider open Jackson networks with losses with mixed finite and infinite queues and analyze the efficiency of sampling from their exact stationary distribution. We show that perfect sampling is possible, although the underlying Markov chain may have an infinite state space. The main idea is to use a Jackson network with infinite buffers (that has a product form stationary distribution) to bound the number of initial conditions to be considered in the coupling from the past scheme. We also provide bounds on the sampling time of this new perfect sampling algorithm for acyclic or hyper-stable networks. These bounds show that the new algorithm is considerably more efficient than existing perfect samplers even in the case where all queues are finite. We illustrate this efficiency through numerical experiments. We also extend our approach to variable service times and non-monotone networks such as queueing networks with negative customers.

7.23. Speeding up Glauber Dynamics for Random Generation of Independent Sets

The maximum independent set (MIS) problem is a well-studied combinatorial optimization problem that naturally arises in many applications, such as wireless communication, information theory and statistical mechanics. MIS problem is NP-hard, thus many results in the literature focus on fast generation of maximal independent sets of high cardinality. One possibility is to combine Gibbs sampling with coupling from the past arguments to detect convergence to the stationary regime. This results in a sampling procedure with time complexity that depends on the mixing time of the Glauber dynamics Markov chain. We propose in [37] an adaptive method for random event generation in the Glauber dynamics that considers only the events that are effective in the coupling from the past scheme, accelerating the convergence time of the Gibbs sampling algorithm.

7.24. Approximate optimality with bounded regret in dynamic matching models

In [28], we consider a dynamic matching model with random arrivals. In prior work, authors have proposed policies that are stabilizing, and also policies that are approximately finite-horizon optimal. This paper considers the infinite-horizon average-cost optimal control problem. A relaxation of the stochastic control problem is proposed, which is found to be a special case of an inventory model, as treated in the classical theory of Clark and Scarf. The optimal policy for the relaxation admits a closed-form expression. Based on the policy for this relaxation, a new matching policy is proposed. For a parameterized family of models in which the network load approaches capacity, this policy is shown to be approximately optimal, with bounded regret, even though the average cost grows without bound.

7.25. Perfect sampling for multiclass closed queueing networks

In [27] we present an exact sampling method for multiclass closed queuing networks. We consider networks for which stationary distribution does not necessarily have a product form. The proposed method uses a compact representation of sets of states, that is used to derive a bounding chain with significantly lower complexity of one-step transition in the coupling from the past scheme. The coupling time of this bounding chain can be larger than the coupling time of the exact chain, but it is finite in expectation. Numerical experiments show that coupling time is close to that of the exact chain. Moreover, the running time of the proposed algorithm outperforms the classical algorithm.

7.26. Fast and Memory Optimal Low-Rank Matrix Approximation

In this paper, we revisit the problem of constructing a near-optimal rank k approximation of a matrix $M \in [0,1]^{m\times n}$ under the streaming data model where the columns of M are revealed sequentially. We present SLA (Streaming Low-rank Approximation), an algorithm that is asymptotically accurate, when $ks_{k+1}(M) = o(\sqrt{mn})$ where $s_{k+1}(M)$ is the $(k+1)$-th largest singular value of M. This means that its average mean-square error converges to 0 as m and n grow large (i.e., $\|\widehat{M}^{(k)} - M^{(k)}\|_F^2 = o(mn)$ with high probability, where $\widehat{M}^{(k)}$ and $M^{(k)}$ denote the output of SLA and the optimal rank k approximation of M, respectively). Our algorithm makes one pass on the data if the columns of M are revealed in a random order, and two passes if the columns of M arrive in an arbitrary order. To reduce its memory footprint and complexity, SLA uses random sparsification, and samples each entry of M with a small probability δ. In turn, SLA is memory optimal as its required memory space scales as $k(m+n)$, the dimension of its output. Furthermore, SLA is computationally efficient as it runs in $O(\delta kmn)$ time (a constant number of operations is made for each observed entry of M), which can be as small as $O(k\log(m)^4 n)$ for an appropriate choice of δ and if $n \geq m$.

7.27. Combinatorial Bandits Revisited

[42] investigates stochastic and adversarial combinatorial multi-armed bandit problems. In the stochastic setting under semi-bandit feedback, we derive a problem-specific regret lower bound, and discuss its scaling with the dimension of the decision space. We propose ESCB, an algorithm that efficiently exploits the structure of the problem and provide a finite-time analysis of its regret. ESCB has better performance guarantees than existing algorithms, and significantly outperforms these algorithms in practice. In the adversarial setting under bandit feedback, we propose COMBEXP, an algorithm with the same regret scaling as state-of-the-art algorithms, but with lower computational complexity for some combinatorial problems.

7.28. Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs

A non-backtracking walk on a graph is a directed path such that no edge is the inverse of its preceding edge. The non-backtracking matrix of a graph is indexed by its directed edges and can be used to count non-backtracking walks of a given length. It has been used recently in the context of community detection and has appeared previously in connection with the Ihara zeta function and in some generalizations of Ramanujan graphs. In [26], we study the largest eigenvalues of the non-backtracking matrix of the Erdos-Renyi random graph and of the Stochastic Block Model in the regime where the number of edges is proportional to the number of vertices. Our results confirm the "spectral redemption" conjecture that community detection can be made on the basis of the leading eigenvectors above the feasibility threshold.

7.29. Designing Adaptive Replication Schemes in Distributed Content Delivery Networks

In [32], we address the problem of content replication in large distributed content delivery networks, composed of a data center assisted by many small servers with limited capabilities and located at the edge of the network. The objective is to optimize the placement of contents on the servers to offload as much as possible the data center. We model the system constituted by the small servers as a loss network, each loss corresponding to a request to the data center. Based on large system / storage behavior, we obtain an asymptotic formula for the optimal replication of contents and propose adaptive schemes related to those encountered in cache networks but reacting here to loss events, and faster algorithms generating virtual events at higher rate while keeping the same target replication. We show through simulations that our adaptive schemes outperform significantly standard replication strategies both in terms of loss rates and adaptation speed.

7.30. Spectral Detection in the Censored Block Model

In [36], we consider the problem of partially recovering hidden binary variables from the observation of (few) censored edge weights, a problem with applications in community detection, correlation clustering and synchronization. We describe two spectral algorithms for this task based on the non-backtracking and the Bethe Hessian operators. These algorithms are shown to be asymptotically optimal for the partial recovery problem, in that they detect the hidden assignment as soon as it is information theoretically possible to do so.

7.31. A spectral method for community detection in moderately-sparse degree-corrected stochastic block models

In the ordinary stochastic block model, all degrees in a cluster have the same expected degree. The Degree-Corrected Stochastic Block Models (DC-SBM) is a generalization of the former where the expected degrees of individual nodes follow a prescribed degree-sequence. We consider community detection in the DC-SBM in a paper currently in preparation [43]. We perform spectral clustering on a suitably normalized adjacency matrix. This leads to consistent recovery of the block-membership of all but a vanishing fraction of nodes, in the regime where the lowest degree is of order $\log(n)$ or higher. The main contributions of this paper are (i) the fact that recovery succeeds for very heterogeneous degree-distributions and (ii) a clean analysis for the DC-SBM, which is a messy model.

7.32. An Impossibility Result for Reconstruction in a Degree-Corrected Planted-Partition Model

In a paper currently in preparation [44], we consider a degree-corrected planted-partition model: a random graph on n nodes with two equal-sized clusters. The model parameters are two constants $a, b > 0$ and an i.i.d. sequence $(\phi_i)_{i=1}^n$, with finite second moment Φ^2. Vertices i and j are joined by an edge with probability $\frac{\phi_i \phi_j}{n} a$ whenever they are in the same class and with probability $\frac{\phi_i \phi_j}{n} b$ otherwise. We prove that the underlying community structure cannot be accurately recovered from observations of the graph when $(a-b)^2 \Phi^2 \leq 2(a+b)$.

7.33. Universality in polytope phase transitions and message passing algorithms

In [], we consider a class of nonlinear mappings $\mathsf{F}_{A,N}$ in $\mathbb{R}^N$ indexed by symmetric random matrices $A \in \mathbb{R}^{N\times N}$ with independent entries. Within spin glass theory, special cases of these mappings correspond to iterating the TAP equations and were studied by Bolthausen [Comm. Math. Phys. 325 (2014) 333-366]. Within information theory, they are known as "approximate message passing" algorithms. We study the high-dimensional (large N) behavior of the iterates of F for polynomial functions F, and prove that it is universal; that is, it depends only on the first two moments of the entries of A, under a sub-Gaussian tail condition. As an application, we prove the universality of a certain phase transition arising in polytope geometry and compressed sensing. This solves, for a broad class of random projections, a conjecture by David Donoho and Jared Tanner.

7.34. Contagions in Random Networks with Overlapping Communities

In [13], we consider a threshold epidemic model on a clustered random graph with overlapping communities. In other words, our epidemic model is such that an individual becomes infected as soon as the proportion of her infected neighbors exceeds the threshold q of the epidemic. In our random graph model, each individual can belong to several communities. The distributions for the community sizes and the number of communities an individual belongs to are arbitrary. We consider the case where the epidemic starts from a single individual, and we prove a phase transition (when the parameter q of the model varies) for the appearance of a cascade, i.e. when the epidemic can be propagated to an infinite part of the population. More precisely, we show that our epidemic is entirely described by a multi-type (and alternating) branching process, and then we apply Sevastyanov's theorem about the phase transition of multi-type Galton-Watson branching processes. In addition, we compute the entries of the matrix whose largest eigenvalue gives the phase transition.

7.35. The Diameter of Weighted Random Graphs.

In [3], we study the impact of random exponential edge weights on the distances in a random graph and, in particular, on its diameter. Our main result consists of a precise asymptotic expression for the maximal weight of the shortest weight paths between all vertices (the weighted diameter) of sparse random graphs, when the edge weights are i.i.d. exponential random variables.

EVA Team

7. New Results

7.1. Wireless Sensor Networks

7.1.1. Time slot and channel assignment in multichannel Wireless Sensor Networks

Participants: Pascale Minet, Ridha Soua, Erwan Livolant.

Wireless sensor networks (WSNs) play a major role in industrial environments for data gathering (convergecast). Among the industrial requirements, we can name a few like 1) determinism and bounded convergecast latencies, 2) throughput and 3) robustness against interferences. The classical IEEE 802.15.4 that has been designed for low power lossy networks (LLNs) partially meets these requirements. That is why the IEEE 802.15.4e MAC amendment has been proposed recently. This amendment combines a slotted medium access with a channel hopping (i.e. Time Slotted Channel Hopping TSCH). The MAC layer orchestrates the medium accesses of nodes according to a given schedule. Nevertheless, this amendment does not specify how this schedule is computed. We propose a distributed joint time slot and channel assignment, called $Wave$ for data gathering in LLNs. This schedule targets minimized data convergecast delays by reducing the number of slots assigned to nodes. Moreover, $Wave$ ensures the absence of conflicting transmissions in the schedule provided. In such a schedule, a node is awake only during its slots and the slots of its children in the convergecast routing graph. Thus, energy efficiency is ensured. We describe in details the functioning of $Wave$, highlighting its features (e.g. support of heterogeneous traffic, support of a sink equipped with multiple interfaces) and properties in terms of worst case delays and buffer size. We discuss its features with regard to a centralized scheduling algorithm like $TMCP$ and a distributed one like $DeTAS$. Simulation results show the good performance of $Wave$ compared to $TMCP$. Since in an industrial environment, several routing graphs can coexist, we study how $Wave$ supports this coexistence.

7.1.2. Centralized Scheduling in TSCH-based Wireless Sensor Networks

Participants: Erwan Livolant, Pascale Minet, Thomas Watteyne.

Scheduling in an IEEE802.15.4e TSCH(Time Slotted Channel Hopping 6TiSCH) low-power wireless network can be done in a centralized or distributed way. When using centralized scheduling, a scheduler installs a communication schedule into the network. This can be done in a standards-based way using CoAP. In this study, we compute the number of packets and the latency this takes, on real-world examples. The result is that the cost is very high using today's standards, much higher than when using an ad-hoc solution such as OCARI. We conclude by making recommendations to drastically reduce the number of messages and improve the efficiency of the standardized approach.

7.1.3. Distributed and Optimized Deployment of WSNs

Participants: Ines Khoufi, Pascale Minet.

This is a joint work with Telecom SudParis: Anis Laouiti.
We are witnessing the deployment of many wireless sensor networks in various application domains such as pollution detection in the environment, intruder detection at home, preventive maintenance in industrial process, monitoring of temporary industrial worksites, damage assessment after a disaster.... Wireless sensor networks are deployed to monitor physical phenomena. The accuracy of the information collected depends on the position of sensor nodes. These positions must meet the application requirements in terms of coverage and connectivity, which therefore requires the use of deployment algorithms. We distinguish two cases: firstly when the nodes are autonomous, and secondly when they are static and the deployment is assisted by mobile robots. In both cases, this deployment must not only meet the application's coverage and connectivity requirements, but must also minimize the number of sensors needed while satisfying various constraints (e.g. obstacles, energy, fault-tolerant connectivity). We distinguished two cases: autonomous and mobile wireless sensor nodes on the one hand, and static wireless sensor nodes on the other hand.

We propose a distributed and optimized deployment of mobile and autonomous sensor nodes to ensure full coverage of the 2D-area considered, as well as network connectivity. With the full coverage of the area, any event occurring in this area is detected by at least one sensor node. In addition, the connectivity ensures that this event is reported to the sink in charge of analyzing the data gathered from the sensors and acting according to these data. This distributed algorithm, called OA-DVFA, can run in an unknown area with obstacles dicovered dynamically. We distinguish two types of obstacles: the transparent ones like ponds in outdoor environment, or tables in an indoor site that only prevent the location of sensor nodes inside them; whereas the opaque obstacles like walls or trees prevent the sensing by causing the existence of hidden zones behind them: such zones may remain uncovered. Opaque obstacles are much more complex to handle than transparent ones and require the deployment of additional sensors to eliminate coverage holes. OA-DVFA is based on virtual forces to obtain a fast spreading of sensor nodes and uses a virtual grid to stop node oscillations and save energy by making sleep redundant nodes. It automatically detects when the maximum area coverage is reached.
We also considered 3D volumes and proposed an algorithm, called 3D-DVFA, also based on virtual forces, to ensure full coverage of 3D volumes and ensure network connectivity. This is a joint work with Nadya Boufares from ENSI, Tunisia. Since applications of such 3D deployments may be limited, we focus on 3D surface covering, where the objective is to deploy wireless sensor nodes on a 3D-surface (e.g. a mountain) to ensure full area coverage and network connectivity. To reach this goal we propose 3D-DVFA-SC, a distributed deployment algorithm based on virtual forces strategy to move sensor nodes.

7.1.4. WSN deployment assisted by mobile robots

Participants: Ines Khoufi, Pascale Minet.

This is a joint work with Telecom SudParis: Anis Laouiti.
Autonomous deployment may be expensive when the number of mobile sensor nodes is very high. In this case, an assisted deployment may be necessary: the nodes' positions being pre-computed and given to mobile robots that place a static sensor at each position. In order to reduce both the energy consumed by the robots, their exposure time to a hostile environment, as well as the time at which the wireless network becomes operational, the optimal tour of robots is this minimizing the delay. This delay must take into account not only the time needed by the robot to travel the tour distance but also the time spent in the rotations performed by the robot each time it changes its direction. This problem is called the Multiple Robot Deploying Sensor nodes problem, in short MRDS. We first show how this problem differs from the well-known traveling salesman problem. We adopt two approaches to optimize the deployment duration. The first one is based on game theory to optimize the length of the tours of two robots (TRDS), and the second is based on a multi-objective optimization, for multiple robots (MRDS). The objectives to be met are: optimizing the duration of the longest tour, balancing the durations of the robot tours and minimizing the number of robots used, while bypassing obstacles.

The TRDS problem is modeled as a non-cooperative game with two players representing the mobile robots, these robots compete for the selection of the sensor nodes to deploy. Each robots tends to maximize its utility function.

We then propose an integer linear program formulation of the MRDS problem. We propose various algorithms relevant to iterative improvement by exchanging tour edges, genetic approach and hybridization. The solutions provided by these algorithms are compared and their closeness to the optimal is evaluated in various configurations.

7.1.5. Sinks Deployment and Packet Scheduling for Wireless Sensor Networks

Participants: Nadjib Achir, Paul Muhlethaler.

The objective of this work is to propose an optimal deployment and distributed packet scheduling of multi-sink Wireless Sensors networks (WNSs). We start by computing the optimal deployment of sinks for a given maximum number of hops between nodes and sinks. We also propose an optimal distributed packet scheduling in order to estimate the minimum energy consumption. We consider the energy consumed due to reporting, forwarding and overhearing. In contrast to reporting and forwarding, the energy used in overhearing is difficult to estimate because it is dependent on the packet scheduling. In this case, we determine the lower-bound of overhearing, based on an optimal distributed packet scheduling formulation. We also propose another estimation of the lower-bound in order to simulate non interfering parallel transmissions which is more tractable in large networks. We note that overhearing largely predominates in energy consumption. A large part of the optimizations and computations carried out in this work are obtained using ILP (Integer Linear Programming) formalization.

7.1.6. Security in wireless sensor networks

Participants: Selma Boumerdassi, Paul Muhlethaler.

Sensor networks are often used to collect data from the environment where they are located. These data can then be transmitted regularly to a special node called a *sink*, which can be fixed or mobile. For critical data (like military or medical data), it is important that sinks and simple sensors can mutually authenticate so as to avoid data to be collected and/or accessed by fake nodes. For some applications, the collection frequency can be very high. As a result, the authentication mechanism used between a node and a sink must be fast and efficient both in terms of calculation time and energy consumption. This is especially important for nodes which computing capabilities and battery lifetime are very low. Moreover, an extra effort has been done to develop alternative solutions to secure, authenticate, and ensure the confidentiality of sensors, and the distribution of keys in the sensor network. Specific researches have also been conducted for large-scale sensors. At present, we work on an exchange protocol between sensors and sinks based on low-cost shifts and xor operations.

7.1.7. Massive MIMO Cooperative Communications for Wireless Sensor Networks

Participants: Nadjib Achir, Paul Muhlethaler.

This work is a collaboration with Mérouane Debbah (Supelec, France).

The objective of this work is to propose a framework for massive MIMO cooperative communications for Wireless Sensor Networks. Our main objective is to analyze the performances of the deployment of a large number of sensors. This deployment should cope with a high demand for real time monitoring and should also take into account energy consumption. We have assumed a communication protocol with two phases: an initial training period followed by a second transmit period. The first period allows the sensors to estimate the channel state and the objective of the second period is to transmit the data sensed. We start analyzing the impact of the time devoted to each period. We study the throughput obtained with respect to the number of sensors when there is one sink. We also compute the optimal number of sinks with respect to the energy spent for different values of sensors. This work is a first step to establish a complete framework to study energy efficient Wireless Sensor Networks where the sensors collaborate to send information to a sink. Currently, we are exploring the multi-hop case.

7.2. Cognitive Radio Networks

7.2.1. Multichannel time slot assignment in Cognitive Radio Sensor Networks

Participants: Ons Mabrouk, Pascale Minet.

This is a joint work with Hanen Idoudi and Leila Saidane from ENSI, Tunisia.
The unlicensed spectrum bands become overcrowded causing an increased level of interference for current wireless sensor nodes. Cognitive Radio Sensor Networks (CRSNs) overcome this problem by allowing sensor nodes to access opportunistically the underutilized licensed spectrum bands. The sink assigns the spectrum holes to the secondary users (SUs). Therefore, it must rely on reliable information about the spectrum holes to protect the primary users (PUs). We focused on the MultiChannel Time Slot Assignment problem in CRSN and tackled this problem: first at the level of a cluster (i.e. Intra-cluster multichannel scheduling), second at the level of several clusters coexisting in the same CRSN (i.e. inter-cluster multichannel scheduling).

In 2013, we proposed an Opportunistic centralized TIme slot assignment in COgnitive Radio sensor networks (OTICOR) for the Intra-cluster multichannel scheduling. OTICOR differs from the existing schemes in its ability to allow non-interfering cognitive sensors to access the same channel and time slot pair. OTICOR takes advantages of spatial reuse, multichannel communication and multiple radio interfaces of the sink. We proved through simulations that a smaller schedule length improves the throughput. Applying OTICOR, we showed that, even in the presence of several PUs, the average throughput granted to SUs remains important. We also showed how to get the best performances of OTICOR when the channel occupancy by PUs is known.

In 2014, we extended this Intra-cluster multichannel scheduling algorithm by proposing two ways for the sink to determine the available channels and alert the SUs if an unexpected activity of PU occurs. Our objective is to design an algorithm able to detect the unexpected presence of PUs in the multi-hop network while maximizing the throughput. If the estimation of PU presence is accurate, a channel sensing at the beginning of the slotframe is sufficient. This algorithm, called Frame-ICMS (Frame Intra-Cluster Multichannel Scheduling), takes advantage of the slots dedicated to the control period by allowing noninterfering cognitive sensors to access the control/data channel and time slot pair. We showed through simulations that using the control period also for data transmission minimizes the schedule length and maximizes the throughput. However, if the estimation of PU presence is not accurate, channel sensing should be done before each slot. We proposed the Slot-ICMS algorithm.

In 2015, we focused on inter-cluster multichannel scheduling algorithm. We considered the coexistence of different clusters in a same CSRN, each cluster having an intra-cluster multichannel scheduling algorithm. Our goal is to obtain a better scalability without loosing the properties provided by OTICOR:

- collision-free schedule,
- minimized data gathering delays,
- sleeping periods per node to save node's energy.

However, the co-existence of several clusters in the same environment may lead to conflicts in the allocation of time slots and channels between these clusters. To avoid inter-cluster collisions, we donot require that different clusters use different channels, we assign distinct channels only to nodes having one-hop neighbors out of their cluster. Once the problem of inter-cluster collision is avoided, each cluster head schedules the transmissions of its members independently. This whole solution exhibits good performances as shown by the simulation results.

7.3. Learning for an efficient and dynamic management of network resources and services

7.3.1. Learning in networks

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet, Selma Boumerdassi.

This work is a joint work with Dominique Barth (University of Versailles-Saint-Quentin). To guarantee an efficient and dynamic management of network resources and services we intend to use a powerful mathematical tool: prediction and learning from prediction. Prediction will be concerned with guessing the evolution of network or network components state, based on knowledge about the past elements and/or other available information. Basically, the prediction problem could be formulated as follows: a forecaster observes the values of one or several metrics giving indications about the network state (generally speaking the network represents the environment). At each time t, before the environment reveals the new metric values, the forecaster predicts the new values based on previous observations. Contrary to classical methods where the environment evolution is characterized by stochastic process, we suppose that the environment evolution follows an unspecified mechanism, which could be deterministic, stochastic, or even adaptive to a given behavior. The prediction process should adapt to unpredictable network state changes due to its non-stationary nature. To properly address the adaptivity challenge, a special type of forecasters is used: the experts. These experts analyse the previous environment values, apply their own computation and make their own prediction. The experts predictions are given to the forecaster before the next environment values are revealed.

The forecaster can then make its own prediction depending on the experts' "advice". The risk of a prediction may be defined as the value of a loss function measuring the discrepancy between the predicted value and the real environment value. The principal notion to optimize the behavior of the forecasters is the regret, seen as a difference between the forecaster's accumulated loss and that of each expert. To optimize the prediction process means to construct a forecasting strategy that guarantees a small loss with respect to defined experts. Adaptability of the forecaster is reflected in the manner in which it is able to follow the better expert according to the context.

Our purpose is to apply on-line learning strategies to:

- Wireless Sensor Networks (WSNs) to predict the quality of a wireless link in a WSN, based on the LQI metric for instance and take advantage of wireless links with the best possible quality to improve the packet delivery rate. We model this problem as a forecaster prediction game based on the advice of several experts. The forecaster learns on-line how to adjust its prediction to better fit the environment metric values. A forecaster estimates the LQI value using the advice of experts.
- Content Delivery Networks (CDNs) to predict the number of solicitations of video contents to cache the contents with the highest popularity.
- Data centers require a huge amount of energy. As an example, in 2014, the electric consumption of all date centers will be larger than 42 TWh, and after 2020 the CO2 production will be larger then 1.27 GTons, ie. more than the aeronautic industry (GeSI SMARTer 2020 report). These "frightening" figures led the research community to work on the management of energy consumption. Several tracks have been explored, among which the optimization of computation and load balancing of servers. At present, we work on tools dedicated to traffic prediction, thus allowing a better management of servers. Our work consists in modeling the traffic specific to data centers and apply different statistical prediction methods.

7.3.2. Tools for learning and prediction

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet.

In 2015, Nesrine Ben Hassine developped an extraction tool to provide real traces from YouTube. these real traces are used as a learning sample by the different prediction algorithms used.

Nesrine Ben Hassine and Dana Marinca extended their simulation tool developed in Python to integrate:

- various prediction strategies SES (Single Explonential Smoothing), DES (Double Exponential Smoothing), Basic and enhanced basic, strategies based on averages (e.g. Average on a Moving Window), regressions (e.g. polynomial or Savitzky Golay), as well as prediction strategies adapting dynamically their parameters according to the loss obtained.
- various loss functions (e.g. absolute value, square). The prediction accuracy is evaluated by a loss function as the discrepancy between the prediction value and the real number obtained.
- different forecaster strategies: Best expert, Exponential Weighted Average, K Best-Experts, etc.

With these tools, we can now tune parameters of prediction strategies and evaluate them.

7.3.3. Popularity prediction in CDNs

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet.

To predict the popularity of video contents, expressed as the number of solicitations, we compared three prediction strategies: Single Exponential Smooting (SES), Double Exponential Smoothing (DES) and Basic. The best tuning of each strategy is determined, depending on the considered phase of the solicitation curve. For DES, values of the smoothing factor close to 1 probide the best results. We study the behavior of each strategy within a phase and around a phase change, where a phase is defined as an interval of time during which a measured metric remains relatively stable.

Basic expert makes large errors at the phase change, but it quickly corrects its prediction and it is the expert having the closest prediction to the real value within a phase. DES expert provides also good quality predictions within a phase. Since DES and Basic experts outperform the SES expert, we recommend the use of on the one hand, the best DES expert per phase within a phase and on the other hand, the Basic expert to automatically detect phase changes, because of its better reactivity. This self-learning and prediction method can be applied to optimize resources allocation in service oriented architectures and self-adaptive networks, more precisely for cache management in CDNs.

7.3.4. Automatic phase detection in popularity evolution of video contents

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet.

In Content Delivery Networks (CDNs) where experts predict the number of solicitations of video contents, simulations based on real YouTube traces show that the accuracy of prediction is improved by splitting the video content profile in contiguous phases. A phase is an interval of time during which a measured metric remains relatively stable. The best expert per phase outperforms the best expert on the whole video content profile. Different prediction methods are compared and also different phase change-points detection methods are evaluated:

- the R tool using Bayesian inference,
- the Basic expert (an important loss may indicate a phase change),
- a fixed time interval (e.g. each week).

The goal is to identify the method (or method parameters) minimizing the cumulated discrepancy compared to real solicitations of video contents. The use of this machine learning method allows the Content Delivery Network to self-adapt to users solicitations by caching the most popular contents near the end users. More generally, such method can be applied to decide which contents should be replicated to improve the performance of audio and video applications and maximize the satisfaction degree of users.

7.4. VANETs

7.4.1. Protocols for VANETs

Participants: Nadjib Achir, Younes Bouchaala, Mohamed Elhadad Or Hadded, Paul Muhlethaler, Oyunchimeg Shagdar.

7.4.1.1. Synthetic study of TDMA protocols for VANETs

Recently several Time Division Multiple Access (TDMA)-based medium access control protocols have been proposed for VANETs in an attempt to ensure that all the vehicles have enough time to send safety messages without collisions and to reduce the end-to-end delay and the packet loss ratio. In this paper, we identify the reasons for using the collision-free medium access control paradigm in VANETs. We then present a novel topology-based classification and we provide an overview of TDMA-based MAC protocols that have been proposed for VANETs. We focus on the characteristics of these protocols, as well as on their benefits and limitations. Finally, we give a qualitative comparison, and we discuss some open issues that need to be tackled in future studies in order to improve the performance of TDMA-based MAC protocols for vehicle to vehicle (V2V) communications.

7.4.1.2. A stable clustering protocol for VANETs

VANETs have a highly dynamic and portioned network topology due to the constant and rapid movement of vehicles. Currently, clustering algorithms are widely used as the control schemes to make VANET topology less dynamic for Medium Access Control (MAC), routing and security protocols. An efficient clustering algorithm must take into account all the necessary information related to node mobility. In this paper, we propose an Adaptive Weighted Clustering Protocol (AWCP), specially designed for vehicular networks, which takes the highway ID, direction of vehicles, position, speed and the number of neighboring vehicles into account in order to enhance the stability of the network topology. However, the multiple control parameters of our AWCP, make parameter tuning a nontrivial problem. In order to optimize the protocol, we define a

multi-objective problem whose inputs are the AWCP's parameters and whose objectives are: providing stable cluster structures, maximizing data delivery rate, and reducing the clustering overhead. We address this multi-objective problem with the Nondominated Sorted Genetic Algorithm version 2 (NSGA-II). We evaluate and compare its performance with other multi-objective optimization techniques: Multi-objective Particle Swarm Optimization (MOPSO) and Multi-objective Differential Evolution (MODE). The experiments reveal that NSGA-II improves the results of MOPSO and MODE in terms of spacing, spread, ratio of non-dominated solutions, and inverse generational distance, which are the performance metrics used for comparison.

7.4.1.3. Using Road IDs to Enhance Clustering in Vehicular Ad hoc Networks

Vehicular ad hoc networks (VANETs) where vehicles act as mobile nodes is an instance of Mobile Ad hoc NETworks (MANETs), which are essentially developed for intelligent transportation systems. A challenging problem when designing communication protocols in VANETs is coping with high vehicle mobility, which causes frequent changes in the network topology and leads to frequent breaks in communication. The clustering technique is being developed to reduce the impact of mobility between neighboring vehicles. In this paper, we propose an Adaptive Weighted Cluster Protocol for VANETs, which is a road map dependent and uses road IDs and movement direction in order to make the clusters structure as stable as possible. The experimental results reveal that AWCP outperforms four other most commonly used clustering protocols in terms of control packet overhead, the packet delivery ratio, and the average cluster lifetime, which are the most usual metrics used for comparing performance.

7.4.2. Models for VANETs

Participants: Nadjib Achir, Younes Bouchaala, Guy Fayolle, Paul Muhlethaler, Oyunchimeg Shagdar.

7.4.2.1. Model of IEEE 802.11 broadcast scheme with infinite queue

We have analyzed the so-called back-off technique of the IEEE 802.11 protocol in broadcast mode with waiting queues. In contrast to existing models, packets arriving when a station (or node) is in back-off state are not discarded, but are stored in a buffer of infinite capacity. As in previous studies, the key point of our analysis hinges on the assumption that the time on the channel is viewed as a random succession of transmission slots (whose duration corresponds to the length of a packet) and mini-slots during which the back-off of the station is decremented. These events occur independently, with given probabilities. The state of a node is represented by a two-dimensional Markov chain in discrete-time, formed by the back-off counter and the number of packets at the station. Two models are proposed both of which are shown to cope reasonably well with the physical principles of the protocol. The stabillity (ergodicity) conditions are obtained and interpreted in terms of maximum throughput. Several approximations related to these models are also discussed.

7.4.2.2. Model and optimization of CSMA

We have studied the maximum throughput of CSMA in scenarios with spatial reuse. The nodes of our network will be a Poisson Point Process (PPP) of a one or two dimensional space. The one dimensional well fits VANETs. To model the effect of Carrier Sense Multiple Access (CSMA), we give random marks to our nodes and to elect transmitting nodes in the PPP we choose the nodes with the smallest marks in their neighborhood, this is the Matern hardcore selection process. To describe the signal propagation, we use a signal with power-law decay and we add a random Rayleigh fading. To decide whether or not a transmission is successful, we adopt the Signal-over-Interference Ratio (SIR) model in which a packet is correctly received if its transmission power divided by the interference power is above a capture threshold. We assume that each node in our PPP has a random receiver at a typical distance. We choose the average distance to its closest neighbor. We also assume that all the network nodes always have a pending packet. With all these assumptions, we analytically study the density of throughput of successful transmission and we show that it can be optimized with the carrier-sense threshold.

7.4.2.3. Performance analysis of IEEE 802.11 broadcast schemes

We have analyzed different broadcast strategies in IEEE 802.11p Vehicular Ad-hoc NETworks (VANETs). The first strategy is the default IEEE 802.11p strategy. Using a model derived from the Bianchi model, we provide the network performance in terms of throughput and success rate. The second strategy is to use an

acknowledgment technique similar to the acknowledgment with point-to-point traffic. A node will send its broadcast packet as in the default case,but it requires an acknowledgment from a neighbor node. This node may be a random neighbor or may be selected according to precise rules. We analyze this second strategy in terms of throughput and success rate. Somewhat surprisingly, we show that this second strategy improves the delivery ratio of the transmitted packets but reduces the overall throughput. This means that if the CAM messages (Cooperative Awareness Messages) are broadcasted, the total number of packets actually delivered will be greater with the default strategy than with the improved strategy. We propose a third strategy which consists in using the default strategy for normal packets, but we add random redundant transmissions to ensure greater reliability for very important packets. We show that with this simple technique, not only do we obtain suitable reliability, but we also achieve larger global throughput than with the acknowledgment-oriented technique. We have also computed network performance in terms of throughput and success rate with respect to the network parameters and to analyze their impact on performances.

7.5. Models for wireless networks

7.5.1. Interference and SINR coverage in spatial non-slotted Aloha networks

Participants: Bartek Blaszczyszyn, Paul Muhlethaler.

We propose two analytically tractable stochastic-geometric models of interference in adhoc networks using pure (non-slotted) Aloha as the medium access. In contrast the slotted model, the interference in pure Aloha may vary during the transmission of a tagged packet. We develop closed form expressions for the Laplace transform of the empirical average of the interference experienced during the transmission of a typical packet. Both models assume a power-law path-loss function with arbitrarily distributed fading and feature configurations of transmitters randomly located in the Euclidean plane according to a Poisson point process. Depending on the model, these configurations vary over time or are static. We apply our analysis of the interference to study the Signal-to- Interference-and-Noise Ratio (SINR) outage probability for a typical transmission in pure Aloha. The results are used to compare the performance of non-slotted Aloha to the slotted one, which has almost exclusively been previously studied in context of wired ad-hoc networks.

7.5.2. Random linear multihop relaying in a general field of interferers using spatial Aloha

Participants: Bartek Blaszczyszyn, Paul Muhlethaler.

We study a stationary Poisson pattern of nodes on a line embedded in an independent planar Poisson field of interfering nodes. Assuming slotted Aloha and the signal-to-interference-and-noise ratio capture condition, with the usual power-law path loss model and Rayleigh fading, we explicitly evaluate several local and end-to-end performance characteristics related to the nearest-neighbor packet relaying on this line, and study their dependence on the model parameters (the density of relaying and interfering nodes, Aloha tuning and the external noise power). Our model can be applied in two cases: the first use is for vehicular ad-hoc networks, where vehicles are randomly located on a straight road. The second use is to study a “typical” route traced in a (general) planar ad-hoc network by some routing mechanism. The approach we have chosen allows us to quantify the non-efficiency of long-distance routing in “pure ad-hoc” networks and evaluate a possible remedy for it in the form of additional “fixed” relaying nodes, called road-side units in a vehicular network. It also allows us to consider a more general field of interfering nodes and study the impact of the clustering of its nodes on the routing performance. As a special case of a field with more clustering than the Poison field, we consider a Poisson-line field of interfering nodes, in which all the nodes are randomly located on random straight lines. In this case our analysis rigorously (in the sense of Palm theory) corresponds to the typical route of this network. The comparison to our basic model reveals a paradox: clustering of interfering nodes decreases the outage probability of a single (typical) transmission on the route, but increases the mean end-to-end delay

FUN Project-Team

7. New Results

7.1. Data gathering and coverage in WSN

Participants: Nathalie Mitton, Tahiry Razafindralambo, Arunabha Sen.

Data availability is one of the main goals and challenges in Future Ubiquitous Network and especially in Wireless Sensor Networks. Indeed, gathering and collecting data in a mobile environment is a very challenging task. In [12], the authors uses data mules to organize the data collection from the sensor in the field. The results presented in [12] are based on some new and unique assumptions. First, it is assumed that the mules are mobile but also the sensors that generate the data to be collected. Second, we collection time is not the first optimization criteria. The focus of the paper is on minimizing the number of mules given a time contraints. The problem is shown to be NP complete and a transformation of the problem into a minimum flow problem allow the computation of optimal solution using Integer Linear Programming.

The results presented in [5] use the assumption of a mobile objects tracked by some other mobile objects as in [12]. In the case of [5] the focus in on coverage of mobile targets by mobile Unmanned Aerial Vehicle. The paper takes two major assumptions regarding the limited energy of UAV and the observation range. These two constraints are linked with each other since when the UAV increase its altitude, it consumes more energy but also increase its observation range. The problem under consideration is mathematically represented by defining mixed integer non-linear optimization models. Heuristic procedures are defined and they are based on restricted mixed integer programming (MIP) formulation of the problem. A computational study is carried out to assess the behavior of the proposed models and MIP-based heuristics.

7.2. Routing in FUN

Participants: Nathalie Mitton, Mouna Rekik.

Geographic routing is an attractive routing strategy in wireless sensor networks. It works well in dense networks, but it may suffer from the void problem. For this purpose, a recovery step is required to guarantee packet delivery. Face routing has widely been used as a recovery strategy since proved to guarantee delivery. However, it relies on a planar graph not always achievable in realistic wireless networks and may generate long paths. In [25], we propose GRACO, a new geographic routing algorithm that combines a greedy forwarding and a recovery strategy based on swarm intelligence. During recovery, ant packets search for alternative paths and drop pheromone trails to guide next packets within the network. GRACO avoids holes and produces near optimal paths. Simulation results demonstrate that GRACO leads to a significant improvement of routing performance and scalability when compared to the literature algorithms.

GRACO has first been designed in the general case. We then studied its applicability to the Virtual Power Plants and their specific data packets with different priorities [22], [24]. Indeed, the Smart Grid (SG) incorporates communication networks to the conventional electricity system in order to intelligently integrate distributed energy resources (DERs) and allow for demand side management. The move to Smart grid in developing countries has to cope with great disparities of ICT infrastructures even within the same city. Besides, individual DERs are often too small to be allowed access to energy market, likewise power utilities are unable to effectively control and manage small DERs. We propose the use of affordable and scalable wireless communication technology to aggregate geographically sparse DERs into a single virtual power plant. The enrollment of prosumers in the VPP is conditional to financial performance of the plant. Thus, the VPPs are dynamic and are expected to scale up as more and more prosumers are attracted by their financial benefits. the communication network has to follow this progression and therefore to be scalable and rapidly deploy-able. We present a routing algorithm for data communication within the VPP to support centralized, decentralized or fully distributed control of the VPP's DERs.

Based on this study, we adapted GRACO so it can fit the specific cases of Smart Grid [23] and more specifically to the Neighbor Area Networks (NAN) of Smart Grids, or distribution segment of the power system in the smart grid (SG). The deployment of ICT to support conventional grid will solve legacy problems that used to prevent implementation of smart services such as smart metering, demand side management or the integration of Distributed Energy Resources (DERs) within the smart grid. We demonstrate the effectiveness of GRACO in terms of scalability, peer-to-peer routing, end-to-end delay and delivery rate.

7.3. Deployment and Self-Deployment in FUN

Participants: Nathalie Mitton, Valeria Loscri, Tahiry Razafindralambo.

Mobility management is a difficult task in autonomous networks. However, mobility provide a huge advantage when in comes to specific scenario such as emergency-related ones especially when network connection must be restored to provide basic network access to users. [3] investigates the potential of spontaneous networks for providing Internet connectivity over the emergency area through the sharing of resources owned by the end-user devices. Novel and extremely flexible network deployment strategies are required in order to cope with the user mobility, the limited communication capabilities of wireless devices, and the intrinsic dynamics of traffic loads and QoS requirements. In [3], a novel architecture is proposed to take advantage of existing end-user devices and some algorithm, are described to build and efficiently exploit the spontaneous emergency network.

Following the emergency scenario described in [3], [15] and [21] describe an algorithm to minimizes the control trafic generated by specific nodes in the network used repair the network and the deployment of these specific nodes, This nodes, forming a substitution network, in case of emergency, are injected autonomously in the network by the network to restore basic network service. In order to increase the performance of the network, the injected nodes called substitution routers, use their ability to move to change the shape of the network and to increase its performance. These movements needs huge amount of control messages to maintain consistency regarding routers' positions. [15] and [21] give an algorithm for the deployment of these routers and the autoregressive time serie model to reduce the amount of control trafic used for the deployment.

7.4. Smart Cities

Participants: Nathalie Mitton, Valeria Loscri, Riccardo Petrolo.

Smart City represents one of the most promising, prominent and challenging Internet of Things (IoT) applications, but recent ICT trends suggest more and more that cities could also benefit from Cloud computing. The convergence of IoT paradigm and Cloud computing technology, can play a fundamental role for developing of highly level and organized cities form an ICT point of view, but it is of paramount importance to deal a critical analysis to identify the issues and challenges deriving from this synergy. This detailed study has been dealt in [7], where it is shown as the semantic annotation of the sensors in the cloud, and innovative services can be implemented and considered by bridging Cloud and Internet of Things. The Cloud of Things (CoT) paradigm is also considered in [16], where it is shown how the CoT arrives to better distribute resources, putting together and enabling therefore a horizontal integration of various Internet of Things(IoT) platforms. Semantic interoperability of diverse IoT platforms are also a key concept in [18], where the virtualization of different IoT systems in order to model and represent the architecture in accordance with the common standards-based IoT ontologies is applied. The environment comes with a range of visual drag-and-drop tools, which boosts developers' productivity.

7.5. RFID

Participant: Nathalie Mitton.

One of the devices under consideration by the FUN team is RFID. One of the main issues to widely deploy RFID reader is reader-to-reader collision. Indeed, when the electromagnetic fields of the readers overlap, a collision occurs on the tag laying in the overlapping section and cannot be read. In [10], we propose a high adaptive contention-based medium access control (HAMAC) protocol that considerably reduces readers collision problems in a large-scale dynamic RFID system. HAMAC is based only on realistic assumptions that can be experimented and does not require any additional components on RFID reader in order to improve the performance in terms of throughput, fairness and latency. The central idea of the HAMAC is for the RFID reader to use a WSN-like CSMA approach and to set its initial backoff counter to the maximum value that allows to mitigate collision. Then, according to the network congestion on physical channels the reader tries to dynamically control its contention window by linear decreasing on selected physical channel or multiplicative decreasing after scanning all available physical channels. Extensive simulations are proposed to highlight the performance of HAMAC compared to literature's work in large-scale RFID systems where both readers and tags are mobile. Simulation results show the effectiveness and robustness of the proposed anti-collision protocol in terms of network throughput, fairness, coverage and time to read all tags.

7.6. Localization

Participants: Nathalie Mitton, Roudy Dagher, Valeria Loscri, Salvatore Guzzo Bonifacio.

[20] presents our approach to localize a node with the use of only one landmark. It is a passive and non intrusive cross-layer approach that relies on a signal processing of all received signals Results are evaluated by simulation and show good accuracy. To complete the previous study, we developed [11] a novel array-based method to estimate the path loss exponent (PLE). The method is designed as a part of an automatic calibration step, prior to localization of a source transmitting in the near-far field of the array. The method only requires the knowledge of the ranges between the array elements. By making the antenna elements transmit in turn, the array response model in the near-far field is exploited to estimate the current environment PLE. Simulation results show that this method can achieve good performance with one transmission round. The performance of the PLE estimation is investigated in the context of source localization with a sensitivity analysis to the PLE estimation. These works are the purpose of a pending patent (submitted in March 2015).

Alternatively, we derive similar localization schemes to enable a cooperation between mobile robots to locale a target based on RSSI [13]. Received Signal Strength Indicator (RSSI) is commonly considered and is very popular for target localization applications, since it does not require extra-circuitry and is always available on current devices. Unfortunately, target localizations based on RSSI are affected with many issues, above all in indoor environments. In this paper, we focus on the pervasive localization of target objects in an unknown environment. In order to accomplish the localization task, we implement an Associative Search Network (ASN) on the robots and we deploy a real test-bed to evaluate the effectiveness of the ASN for target localization. The ASN is based on the computation of weights, to "dictate" the correct direction of movement, closer to the target. Results show that RSSI through an ASN is effective to localize a target, since there is an implicit mechanism of correction, deriving from the learning approach implemented in the ASN.

7.7. Vehicular Networks

Participants: Nathalie Mitton, Valeria Loscri.

In the framework of our collaboration with Southern University in China, we investigate a specific issue in Vehicular AdHoc Networks (VANET), the information delivery delay analysis for roadside unit deployment in a VANET with intermittent connectivity [9]. A mathematical model is developed to describe the relationship between the average delay for delivering road condition information and the distance between two neighbor RSUs deployed along a road. The derived mathematical model considers a straight highway scenario where two RSUs are deployed at a distance without any direct connection and vehicles are sparsely distributed on the road with road condition information randomly generated between the two neighbor RSUs. Moreover, the model takes into account the vehicle speed, the vehicle density, the likelihood of an incident, and the distance between two RSUs. The effectiveness of the derived mathematical model is verified through simulation results.

Given the information delivery delay constraint of a time-critical application, this model can be used to estimate the maximum distance allowed between two neighbor RSUs, which can provide a reference for the deployment of RSUs in such scenarios.

But Vehicular Networks can also convey social networks. In [30], we survey recent literature on Vehicular Social Networks that are a particular class of vehicular ad hoc networks, characterized by social aspects and features. Starting from this pillar, we investigate perspectives of next generation vehicles under the assumption of social networking for vehicular applications (i.e., safety and entertainment applications). This paper plays a role as a starting point about socially-inspired vehicles, and main related applications, as well as communication techniques. Vehicular communications can be considered as the "first social network for automobiles", since each driver can share data with other neighbors. As an instance, heavy traffic is a common occurrence in some areas on the roads (e.g., at intersections, taxi loading/unloading areas, and so on); as a consequence, roads become a popular social place for vehicles to connect to each other. Human factors are then involved in vehicular ad hoc networks, not only due to the safety related applications, but also for entertainment purpose. Social characteristics and human behavior largely impact on vehicular ad hoc networks, and this arises to the vehicular social networks, which are formed when vehicles (individuals) "socialize" and share common interests. This survey describes the main features of vehicular social networks, from novel emerging technologies to social aspects used for mobile applications, as well as main issues and challenges. Vehicular social networks are described as decentralized opportunistic communication networks formed among vehicles. They exploit mobility aspects, and basics of traditional social networks, in order to create novel approaches of message exchange through the detection of dynamic social structures. An overview of the main state-of-the-art on safety and entertainment applications relying on social networking solutions is also provided.

7.8. FIT

Participants: Nathalie Mitton, Julien Vandaele.

The universal proliferation of intelligent objects is making Internet of Things (IoT) a reality; to operate on a large scale it will critically rely on new, seamless, forms of communications. But how can innovations be validated in a controlled environment, before being massively deployed into the real world? Several platforms have been deployed to address this issue. In [8], we browse a survey of them, highlighting their characteristics and given some tips to choose the most appropriate to our needs.

Our team has contributed to the deployment of the FIT IoT-LAB platform [2], [19], [27], which addresses this challenge by offering a unique open first class service to all IoT developers, researchers, integrators and developers: a large-scale experimental testbed allowing design, development, deployment and testing of innovative IoT applications, in order to test the future and make it safe. One of the specific deployment focuses on the automatic docking of robots for energy recharge. We explain it in [17]. The objective is to achieve long-term autonomous robots within an experiment test-bed. We propose to combine the use of QR codes as landmarks and Infrared distance sensors. The relative size of the lateral edges of the visual pattern is used to position the robot in relation with the dock. Infrared distance sensors are then used to perform different approaching strategies depending on the distance. Experiments show that the proposed solution is fully operational and robust. Not to rely exclusively on visual pattern recognition avoids potential errors induced by camera calibration. Additionally, as a positive side effect, the use of Infrared sensors allows the robot to avoid obstacles while docking. The finality of such an approach is to integrate these robots into the FIT IoT Lab experimental testbed which allows any experimenter to book wireless resources such as wireless sensors remotely and to test their own code. Wifibots holding wireless sensors will be integrated as additional reservable resources of the platform to enlarge the set of possible experimentations with mobile entities.

7.9. New and other communication paradigms

Participants: Nathalie Mitton, Valeria Loscri, Arash Maskooki, Gabriele Sabatino.

Interconnection and self-organized systems are normally populated with heterogeneous and different devices. The differences range from computational capabilities, storage size, etc. Instead of considering the heterogeneity as a limitation, it is possible to "turn it"as a primitive control of the system, in order to realize more robust and more resilient communication systems. Based on these premises, we identify specific situations, where mobile nodes with a plethora of interesting features and sensing capacities, can be exploited by configuring them in such a way to make them playing different roles in respect of them for which they have been initially conceived [4]. The differentiated use of devices, together with a careful analysis of the characteristics and performance requirements of the current and the future networks, allow the adaptation to the exponential growth in demand for high bandwidth applications [26]. This is exactly the philosophy embraced in [28], where Software Defined Radio (SDR) and Cognitive Radio (CR) have been considered and analyzed in a novel context, namely body networked systems. A detailed analysis of body systems as networked systems has also been considered in [6] and [14]. In [6] a novel communication paradigm, namely a molecular communication, has been considered to show how a nanoparticulate system can be suitable to coexist in a biological environment. An experimental analysis to asses the theoretical assumption has been developed in [14]. In order to asses new/alternative communication paradigms, there is the necessity from one side to consider and analyze the specific context and its level of interaction with the communication system and on the other side the correct identification of the specific features of the communication paradigm itself. This type of analysis allowed the design and implementation of an acoustic communication approach [29], where the ultrasound represent the wave carriers of data information. This "unusual" transmission means has been selected as the most suitable in a context as the body, where the aqueous environment makes it not suitable for more "traditional" communication paradigms, e.g. the one based on Radio Frequency (RF) waves.

GANG Project-Team

7. New Results

7.1. Graph and Combinatorial Algorithms

7.1.1. Rainbow matchings in hypergraphs

A rainbow matching for (not necessarily distinct) sets $F_1, ..., F_k$ of hypergraph edges is a matching consisting of k edges, one from each F_i. In [8], we give some order to the multitude of conjectures that relate to this concept, as well as introduce some new conjectures. We also present some partial results on one of these conjectures, that seems central among them – the so-called Ryser-Brualdi-Stein conjecture.

7.1.2. A graph formulation of the union-closed sets conjecture

In 1979, Frankl conjectured that in a finite non-trivial union-closed collection of sets there has to be an element that belongs to at least half the sets. In [7], we show that this is equivalent to the conjecture that in a finite non-trivial graph there are two adjacent vertices, each belonging to at most half of the maximal stable sets. In this graph formulation other special cases become natural. The conjecture is trivially true for non-bipartite graphs and we show that it also holds for the classes of chordal bipartite graphs, subcubic bipartite graphs, bipartite series-parallel graphs and bipartitioned circular interval graphs.

7.1.3. Cops-and-robber games on k-chordal graphs

The cops-and-robber games, introduced by Winkler and Nowakowski (in Discrete Math. 43, 1983) and independently defined by Quilliot (in J. Comb. Theory, Ser. B 38, 1985), concern a team of cops that must capture a robber moving in a graph. In [20], we consider the class of k-chordal graphs, i.e., graphs with no induced (chordless) cycle of length greater than k, $k \geq 3$. We prove that $k-1$ cops are always sufficient to capture a robber in k-chordal graphs. This leads us to our main result, a new structural decomposition for a graph class including k-chordal graphs.

We present a polynomial-time algorithm that, given a graph G and $k \geq 3$, either returns an induced cycle larger than k in G, or computes a tree-decomposition of G, each bag of which contains a dominating path with at most $k-1$ vertices. This allows us to prove that any k-chordal graph with maximum degree Δ has treewidth at most $(k-1)(\Delta-1)+2$, improving the $O(\Delta(\Delta-1)k-3)$ bound of Bodlaender and Thilikos (Discrete Appl. Math. 79, 1997). Moreover, any graph admitting such a tree-decomposition has small hyperbolicity). As an application, for any n-vertex graph admitting such a tree-decomposition, we propose a compact routing scheme using routing tables, addresses and headers of size $O(k \log \Delta + \log n)$ bits and achieving an additive stretch of $O(k \log \Delta)$. As far as we know, this is the first routing scheme with $O(k \log \Delta + \log n)$-routing tables and small additive stretch for k-chordal graphs.

7.1.4. Distinguishing views in symmetric networks

The view of a node in a port-labeled network is an infinite tree encoding all walks in the network originating from this node. In [16], we prove that for any integers $n \geq D \geq 1$, there exists a port-labeled network with at most n nodes and diameter at most D, which contains a pair of nodes whose (infinite) views are different, but whose views truncated to depth $\Omega(D \log (n/D))$ are identical.

7.1.5. Vertex elimination orderings for hereditary graph classes

In [3], we provide a general method to prove the existence and compute efficiently elimination orderings in graphs. This method relies on several tools that were known before, but that were not put together so far: the algorithm LexBFS due to Rose, Tarjan and Lueker, its additional properties discovered by Berry and Bordat, and a local decomposition property of graphs discovered by Maffray, Trotignon and Vušković. We use this method to prove the existence of elimination orderings in several classes of graphs, and to compute them in linear time. Some of the classes have already been studied, namely even-hole-free graphs, square-theta-free Berge graphs, universally signable graphs and wheel-free graphs. Some other classes are new. It turns out that all the classes that we consider can be defined by excluding some of the so-called Truemper configurations. For several classes of graphs, we obtain directly bounds on the chromatic number, or fast algorithms for the maximum clique problem or the coloring problem.

7.1.6. Fast collaborative graph exploration

In [14], we study the following scenario of online graph exploration. A team of k agents is initially located at a distinguished vertex r of an undirected graph. We ask how many time steps are required to complete exploration, i.e., to make sure that every vertex has been visited by some agent. As our main result, we provide the first strategy which performs exploration of a graph with n vertices at a distance of at most D from r in time $O(D)$, using a team of agents of polynomial size $k = Dn^{1+\epsilon} < n^{2+\epsilon}$, for any $\epsilon > 0$. Our strategy works in the local communication model, in which agents can only exchange information when located at a vertex, without knowledge of global parameters such as n or D.

We also obtain almost-tight bounds on the asymptotic relation between exploration time and team size, for large k, in both the local and the global communication model.

7.1.7. Position discovery for a system of bouncing robots

In [11], we consider a scenario in which a collection of n anonymous mobile robots is deployed on a unit-perimeter ring or a unit-length line segment. Every robot starts moving at constant speed, and bounces each time it meets any other robot or segment endpoint, changing its walk direction. We study the problem of position discovery, in which the task of each robot is to detect the presence and the initial positions of all other robots. The robots cannot communicate or perceive information about the environment in any way other than by bouncing nor they have control over their walks which are determined by their initial positions and their starting directions. Each robot has a clock allowing it to observe the times of its bounces. We give complete characterizations of all initial configurations for both the ring and the segment in which no position detection algorithm exists and we design optimal position detection algorithms for all feasible configurations.

7.1.8. Rendezvous of mobile agents in edge-weighted networks

In [15], we introduce a variant of the deterministic rendezvous problem for a pair of heterogeneous agents operating in an undirected graph, which differ in the time they require to traverse particular edges of the graph. Each agent knows the complete topology of the graph and the initial positions of both agents. The agent also knows its own traversal times for all of the edges of the graph, but is unaware of the corresponding traversal times for the other agent. The goal of the agents is to meet on an edge or a node of the graph. In this scenario, we study the time required by the agents to meet, compared to the meeting time T_{OPT} in the offline scenario in which the agents have complete knowledge about each others' speed characteristics. When no additional assumptions are made, we show that rendezvous in our model can be achieved after time $O(nT_{OPT})$ in a n-node graph, and that such time is essentially in some cases the best possible. However, we prove that the rendezvous time can be reduced to $\Theta(T_{OPT})$ when the agents are allowed to exchange $\Theta(n)$ bits of information at the start of the rendezvous process. We then show that under some natural assumption about the traversal times of edges, the hardness of the heterogeneous rendezvous problem can be substantially decreased, both in terms of time required for rendezvous without communication, and the communication complexity of achieving rendezvous in time $\Theta(T_{OPT})$.

7.1.9. Monitoring a graph using faulty mobile robots

In the scenario studied in [27], a team of k mobile robots is deployed on a weighted graph whose edge weights represent distances. The robots perpetually move along the domain, represented by all points belonging to the graph edges, not exceeding their maximal speed. The robots need to patrol the graph by regularly visiting all points of the domain. Here, we consider a team of robots (patrolmen), at most f of which may be unreliable, i.e. they fail to comply with their patrolling duties.

What algorithm should be followed so as to minimize the maximum time between successive visits of every edge point by a reliable patrolmen? The corresponding measure of efficiency of patrolling called idleness has been widely accepted in the robotics literature. We extend it to the case of untrusted patrolmen; we denote by $I_k^f(G)$ the maximum time that a point of the domain may remain unvisited by reliable patrolmen. The objective is to find patrolling strategies minimizing $I_k^f(G)$.

We investigate this problem for various classes of graphs. We design optimal algorithms for line segments, which turn out to be surprisingly different from strategies for related patrolling problems proposed in the literature. We then use these results to study the case of general graphs. For Eulerian graphs G, we give an optimal patrolling strategy with idleness $I_k^f(G) = (f+1)|E|/k$, where $|E|$ is the sum of the lengths of the edges of G. Further, we show the hardness of the problem of computing the idle time for three robots, at most one of which is faulty, by reduction from 3-edge-coloring of cubic graphs — a known NP-hard problem. A byproduct of our proof is the investigation of classes of graphs minimizing idle time (with respect to the total length of edges); an example of such a class is known in the literature under the name of Kotzig graphs.

7.1.10. Limit behavior of the rotor-router system

The rotor-router model, also called the Propp machine, was introduced as a deterministic alternative to the random walk. In this model, a group of identical tokens are initially placed at nodes of the graph. Each node maintains a cyclic ordering of the outgoing arcs, and during consecutive turns the tokens are propagated along arcs chosen according to this ordering in round-robin fashion. The behavior of the model is fully deterministic. Yanovski et al. (Algorithmica, 2003) proved that a single rotor-router walk on any graph with m edges and diameter D stabilizes to a traversal of an Eulerian circuit on the set of all $2m$ directed arcs on the edge set of the graph, and that such periodic behaviour of the system is achieved after an initial transient phase of at most $2mD$ steps.

The case of multiple parallel rotor-routers was studied experimentally, leading Yanovski et al. to the experimental observation that a system of $k > 1$ parallel walks also stabilizes with a period of length at most $2m$ steps. In our work [26] we disprove this observation, showing that the period of parallel rotor-router walks can in fact, be superpolynomial in the size of graph. On the positive side, we provide a characterization of the periodic behavior of parallel router walks, in terms of a structural property of stable states called a subcycle decomposition. This property provides us the tools to efficiently detect whether a given system configuration corresponds to the transient or to the limit behavior of the system. Moreover, we provide polynomial upper bounds of $O(m^4D^2 + mD\log k)$ and $O(m^5k^2)$ on the number of steps it takes for the system to stabilize. Thus, we are able to predict any future behavior of the system using an algorithm that takes polynomial time and space. In addition, we show that there exists a separation between the stabilization time of the single-walk and multiple-walk rotor-router systems, and that for some graphs the latter can be asymptotically larger even for the case of $k = 2$ walks.

7.2. Distributed Computing

7.2.1. Self-stabilizing verification and computation of an MST

In the work [19], we demonstrate the usefulness of distributed local verification of proofs, as a tool for the design of self-stabilizing algorithms. In particular, it introduces a somewhat generalized notion of distributed local proofs, and utilizes it for improving the time complexity significantly, while maintaining space optimality. As a result, we show that optimizing the memory size carries at most a small cost in terms of time, in the context of Minimum Spanning Tree (MST). That is, we present algorithms that are both time and space efficient for

both constructing an MST and for verifying it. This involves several parts that may be considered contributions in themselves.

First, we generalize the notion of local proofs, trading off the time complexity for memory efficiency. This adds a dimension to the study of distributed local proofs, which has been gaining attention recently. Specifically, we design a (self-stabilizing) proof labeling scheme which is memory optimal (i.e., $O(\log n)$ bits per node), and whose time complexity is $O(\log^2 n)$ in synchronous networks, or $O(\Delta \log^3 n)$ time in asynchronous ones, where Δ is the maximum degree of nodes. This answers an open problem posed by Awerbuch and Varghese (FOCS 1991). We also show that $\Omega(\log n)$ time is necessary, even in synchronous networks. Another property is that if f faults occurred, then, within the required detection time above, they are detected by some node in the O(f log n) locality of each of the faults. Second, we show how to enhance a known transformer that makes input/output algorithms self-stabilizing. It now takes as input an efficient construction algorithm and an efficient self-stabilizing proof labeling scheme, and produces an efficient self-stabilizing algorithm. When used for MST, the transformer produces a memory optimal self-stabilizing algorithm, whose time complexity, namely, $O(n)$, is significantly better even than that of previous algorithms. (The time complexity of previous MST algorithms that used $O(\log^2 n)$ memory bits per node was $O(n^2)$, and the time for optimal space algorithms was $O(n|E|)$.) Inherited from our proof labelling scheme, our self-stabilising MST construction algorithm also has the following two properties: (1) if faults occur after the construction ended, then they are detected by some nodes within $O(\log^2 n)$ time in synchronous networks, or within $O(\Delta \log^3 n)$ time in asynchronous ones, and (2) if f faults occurred, then, within the required detection time above, they are detected within the $O(f \log n)$ locality of each of the faults. We also show how to improve the above two properties, at the expense of some increase in the memory.

7.2.2. Clock synchronization and distributed estimation in highly dynamic networks

In [21], we consider the External Clock Synchronization problem in dynamic sensor networks. Initially, sensors obtain inaccurate estimations of an external time reference and subsequently collaborate in order to synchronize their internal clocks with the external time. For simplicity, we adopt the drift-free assumption, where internal clocks are assumed to tick at the same pace. Hence, the problem is reduced to an estimation problem, in which the sensors need to estimate the initial external time. In this context of distributed estimation, this work is further relevant to the problem of collective approximation of environmental values by biological groups.

Unlike most works on clock synchronization that assume static networks, the setting considered here is an extreme case of highly dynamic networks. We do however impose a restriction on the dynamicity of the network. Specifically, we assume a non-adaptive scheduler adversary that dictates an arbitrary, yet independent, meeting pattern. Such meeting patterns fit, for example, with short-time scenarios in highly dynamic settings, where each sensor interacts with only few other arbitrary sensors.

We propose an extremely simple clock synchronization (or an estimation) algorithm that is based on weighted averages, and prove that its performance on any given independent meeting pattern is highly competitive with that of the best possible algorithm, which operates without any resource or computational restrictions, and further knows the whole meeting pattern in advance. In particular, when all distributions involved are Gaussian, the performances of our scheme coincide with the optimal performances. Our proofs rely on an extensive use of the concept of Fisher information. We use the Cramér-Rao bound and our definition of a Fisher Channel Capacity to quantify information flows and to obtain lower bounds on collective performance. This opens the door for further rigorous quantifications of information flows within collaborative sensors.

7.2.3. Wait-freedom with advice

In [13], we motivate and propose a new way of thinking about failure detectors which allows us to define what it means to solve a distributed task wait-free using a failure detector. In our model, the system is composed of computation processes that obtain inputs and are supposed to produce outputs and synchronization processes that are subject to failures and can query a failure detector. Under the condition that correct (never failing) synchronization processes take sufficiently many steps, they provide the computation processes with enough advice to solve the given task wait-free: every computation process outputs in a finite number of its own

steps, regardless of the behavior of other computation processes. Every task can thus be characterized by the weakest failure detector that allows for solving it, and we show that every such failure detector captures a form of set agreement. We then obtain a complete classification of tasks, including ones that evaded comprehensible characterization so far, such as renaming or weak symmetry breaking.

7.2.4. Linear-space bootstrap communication schemes

In [12], we consider a system of n processes with ids that are drawn from a large space. How can these n processes communicate to solve a problem? It is shown that linear number of Multi-Writer Multi-Reader (MWMR) registers are sufficient to solve any read-write wait-free solvable problem and needed to solve some read-write wait-free solvable problem. This contrasts with the existing possible solution borrowed from adaptive algorithms that require $\Theta(n^{3/2})$ MWMR registers.

To obtain the sufficiency result, we show how the processes can non-blockingly emulate a system of n Single-Writer Multi-Reader (SWMR) registers on top of n Multi-Writer Multi-Reader (MWMR) registers. For the necessity result, we show it is impossible to do such an emulation with $n-1$ MWMR registers.

We also presents a wait-free emulation, using $2n-1$ rather than just n registers. The emulation can be used to solve an infinite sequence of tasks that are sequentially dependent (processes need the previous task's outputs in order to proceed to the next task). A non-blocking emulation cannot be used in this case, because it might starve a process forever.

7.2.5. Space complexity of set agreement

The k-set agreement problem is a generalization of the classical consensus problem in which processes are permitted to output up to k different input values. In a system of n processes, an m-obstruction-free solution to the problem requires termination only in executions where the number of processes taking steps is eventually bounded by m. This family of progress conditions generalizes wait-freedom ($m = n$) and obstruction-freedom ($m = 1$). In [29], we prove upper and lower bounds on the number of registers required to solve m-obstruction-free k-set agreement, considering both one-shot and repeated formulations. In particular, we show that repeated k set agreement can be solved using $n + 2m - k$ registers and establish a nearly matching lower bound of $n + m - k$.

7.2.6. Consensus capability of distributed systems

A fundamental research theme in distributed computing is the comparison of systems in terms of their ability to solve basic problems such as consensus that cannot be solved in completely asynchronous systems. In particular, in a seminal work (ACM Trans. Program. Lang. Syst. 13, 1991), Herlihy compares shared-memory systems in terms of the shared objects that they have: he proved that there are shared objects that are powerful enough to solve consensus for n processes, but are too weak to solve consensus for $n + 1$ processes; such objects are placed at level n of a wait-free hierarchy.

Similarly as in that work, in [30] we compare shared-memory systems with respect to their ability to solve consensus for n processes. But instead of comparing systems defined by the shared objects that they have, we compare read-write systems defined by the set of process schedules that can occur in these systems. Defining systems this way can capture many types of systems, e.g., systems whose synchrony ranges from fully synchronous to completely asynchronous, several systems with failure detectors, and "obstruction-free" systems. Here, we consider read-write systems defined in terms of sets of process schedules, and investigate the following fundamental question: Is there a system of $n + 1$ processes such that consensus can be solved for every subset of n processes in the system, but consensus cannot be solved for the $n + 1$ processes of the system? We show that the answer to the above question is "yes", and so these systems can be classified into a hierarchy akin to Herlihy's hierarchy.

7.2.7. Shared whiteboard models of distributed systems

In [4], we study distributed algorithms on massive graphs where links represent a particular relationship between nodes (for instance, nodes may represent phone numbers and links may indicate telephone calls).

Since such graphs are massive they need to be processed in a distributed way. When computing graph-theoretic properties, nodes become natural units for distributed computation. Links do not necessarily represent communication channels between the computing units and therefore do not restrict the communication flow. Our goal is to model and analyze the computational power of such distributed systems where one computing unit is assigned to each node. Communication takes place on a whiteboard where each node is allowed to write at most one message. Every node can read the contents of the whiteboard and, when activated, can write one small message based on its local knowledge. When the protocol terminates its output is computed from the final contents of the whiteboard. We describe four synchronization models for accessing the whiteboard. We show that message size and synchronization power constitute two orthogonal hierarchies for these systems. We exhibit problems that separate these models, i.e., that can be solved in one model but not in a weaker one, even with increased message size. These problems are related to maximal independent set and connectivity. We also exhibit problems that require a given message size independently of the synchronization model.

7.2.8. Discrete Lotka-Volterra population protocols

In [28], we focus on a natural class of population protocols whose dynamics are modeled by the discrete version of Lotka-Volterra equations with no linear term. In such protocols, when an agent a of type (species) i interacts with an agent b of type (species) j with a as the initiator, then b's type becomes i with probability P_{ij}. In such an interaction, we think of a as the predator, b as the prey, and the type of the prey is either converted to that of the predator or stays as is. Such protocols capture the dynamics of some opinion spreading models and generalize the well-known Rock-Paper-Scissors discrete dynamics. We consider the pairwise interactions among agents that are scheduled uniformly at random.

We start by considering the convergence time and show that any Lotka-Volterra-type protocol on a n-agent population converges to some absorbing state in time polynomial in n, w.h.p., when any pair of agents is allowed to interact. By contrast, when the interaction graph is a star, there exist protocols of the considered type, such as Rock-Paper-Scissors, which require exponential time to converge. We then study threshold effects exhibited by Lotka-Volterra-type protocols with 3 and more species under interactions between any pair of agents. We present a simple 4-type protocol in which the probability difference of reaching the two possible absorbing states is strongly amplified by the ratio of the initial populations of the two other types, which are transient, but "control" convergence. We then prove that the Rock-Paper-Scissors protocol reaches each of its three possible absorbing states with almost equal probability, starting from any configuration satisfying some sub-linear lower bound on the initial size of each species. That is, Rock-Paper-Scissors is a realization of a "coin-flip consensus" in a distributed system. Some of our techniques may be of independent value.

7.2.9. Deterministic load-balancing

In [23], 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. (FOCS 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, e.g., 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 $O(d\sqrt{\log n/\mu}, d\sqrt{n})$ 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.

7.2.10. Randomized local network computing

In [32], we have carried on investigating the line of research questioning the power of randomization for the design of distributed algorithms. In their seminal paper, Naor and Stockmeyer [STOC 1993] established that, in the context of network computing, in which all nodes execute the same algorithm in parallel, any construction task that can be solved locally by a randomized Monte-Carlo algorithm can also be solved locally by a deterministic algorithm. This result however holds in a specific context. In particular, it holds only for distributed tasks whose solutions that can be locally checked by a deterministic algorithm. We have extended the result of Naor and Stockmeyer to a wider class of tasks. Specifically, we proved that the same derandomization result holds for every task whose solutions can be locally checked using a 2-sided error randomized Monte-Carlo algorithm. This extension finds applications to, e.g., the design of lower bounds for construction tasks which tolerate that some nodes compute incorrect values. In a nutshell, we have showed that randomization does not help for solving such resilient tasks.

7.2.11. Proof-labeling schemes: randomization and self-stabilization

We have also carried on investigating the power of randomization for the design of proof-labeling schemes. Recall that a proof-labeling scheme, introduced by Korman, Kutten and Peleg [PODC 2005], is a mechanism enabling to certify the legality of a network configuration with respect to a boolean predicate. Such a mechanism finds applications in many frameworks, including the design of fault-tolerant distributed algorithms. In a proof-labeling scheme, the verification phase consists of exchanging labels between neighbors. The size of these labels depends on the network predicate to be checked. There are predicates requiring large labels, of poly-logarithmic size (e.g., MST), or even polynomial size (e.g., Symmetry). In [22], we introduce the notion of randomized proof-labeling schemes. By reduction from deterministic schemes, we show that randomization enables the amount of communication to be exponentially reduced. As a consequence, we show that checking any network predicate can be done with probability of correctness as close to one as desired by exchanging just a logarithmic number of bits between neighbors. Moreover, we design a novel space lower bound technique that applies to both deterministic and randomized proof-labeling schemes. Using this technique, we establish several tight bounds on the verification complexity of classical distributed computing problems, such as MST construction, and of classical predicates such as acyclicity, connectivity, and cycle length.

Next, we have established the formal connections between self-stabilization and proof-labeling scheme. Recall that self-stabilizing algorithms are distributed algorithms supporting transient failures. Starting from any configuration, they allow the system to detect whether the actual configuration is legal, and, if not, they allow the system to eventually reach a legal configuration. In the context of network computing, it is known that, for every task, there is a self-stabilizing algorithm solving that task, with optimal space-complexity, but converging in an exponential number of rounds. On the other hand, it is also known that, for every task, there is a self-stabilizing algorithm solving that task in a linear number of rounds, but with large space-complexity. It is however not known whether for every task there exists a self-stabilizing algorithm that is simultaneously space-efficient and time-efficient. In [24], we make a first attempt for answering the question of whether such an efficient algorithm exists for every task, by focussing on constrained spanning tree construction tasks. We present a general roadmap for the design of silent space-optimal self-stabilizing algorithms solving such tasks, converging in polynomially many rounds under the unfair scheduler. By applying our roadmap to the task of constructing minimum-weight spanning tree (MST), and to the task of constructing minimum-degree spanning tree (MDST), we provide algorithms that outperform previously known algorithms designed and optimized specifically for solving each of these two tasks.

7.2.12. Role of node identifiers in local decision

We have also investigated the role of IDs in network computing. This role 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 such that (1) they cannot be solved without unique identifiers, and (2) unique node identifiers leak a sufficient amount of information such that the problem becomes solvable. In [33] we study 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 multi-set $f(n)$ of n labels; then the adversary assigns the labels to the n nodes in the network. This is a direct generalization of the usual assumption of unique node identifiers. We give a complete characterization 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 behavior, 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.

7.2.13. Geometry on the utility space

In [31], we study the geometrical properties of the utility space (the space of expected utilities over a finite set of options), which is commonly used to model the preferences of an agent in a situation of uncertainty. We focus on the case where the model is neutral with respect to the available options, i.e. treats them, a priori, as being symmetrical from one another. Specifically, we prove that the only Riemannian metric that respects the geometrical properties and the natural symmetries of the utility space is the round metric. This canonical metric allows to define a uniform probability over the utility space and to naturally generalize the Impartial Culture to a model with expected utilities.

7.3. Network Algorithms and Analysis

7.3.1. Information dissemination on social networks

In [17], we model an online social network as a network formation game. We study convergence of selfish dynamics and show that somewhat natural metric assumption enable fast convergence towards an equilibrium with efficient collaborative filtering of content.

7.3.2. Verification of network forwarding tables

In [25], we investigate the problem of verifying forwarding network tables. We show that it is sufficient to test few representative headers when the set of rules applied by routers is complete under intersection.

7.3.3. Refreshing old datasets in a network: LiveRank

In [18], we consider the problem of refreshing a dataset. More precisely, given a collection of nodes gathered at some time (Web pages, users from an online social network) along with some structure (hyperlinks, social relationships), we want to identify a significant fraction of the nodes that still exist at present time. The liveness of an old node can be tested through an online query at present time. We call LiveRank a ranking of the old pages so that active nodes are more likely to appear first. The quality of a LiveRank is measured by the number of queries necessary to identify a given fraction of the active nodes when using the LiveRank order. We study different scenarios from a static setting where the LiveRank is computed before any query is made, to dynamic settings where the LiveRank can be updated as queries are processed. Our results show that building on the PageRank can lead to efficient LiveRanks, for Web graphs as well as for online social networks.

7.3.4. Exploiting user movement for position detection

The major issue of indoor localization system is the trade-off between implementation cost and accuracy. A low-cost system which demands only few hardware devices could save the cost but often it turns out to be less reliable. Aiming at improving classical triangulation method that requires several reference points, we propose in [34] a new method, called Two-Step Movement (2SM), which requires only one reference point (RP) by exploiting useful information given by the position change of a mobile terminal (MT), or the user movement. This method can minimize the number of reference points required in a localization system or

navigation service and reduce system implementation cost. Analytical result shows that the user position can be thus derived and given in simple closed-form expression. Finally, simulation is conducted to demonstrate its effectiveness under noisy environment.

Then, in [35], we build on 2SM. We first improve the positioning performance through multi-sampling technique to combat measurement noise. Secondly, we propose the Generalized Two-Step Movement (G2SM) method for device-to-device (D2D) systems in which both the mobile terminal (MT) and RP can be mobile device. The mobile user's position can be derived analytically and given in simple closed-form expression. Its effectiveness in the presence of noise is shown in simulation results.

7.3.5. Fast diameter and radius computation in real-world graphs

In [5], we propose a new algorithm that computes the radius and the diameter of a weakly connected digraph $G = (V, E)$, by finding bounds through heuristics and improving them until they are validated. Although the worst-case running time is $O(|V||E|)$, we will experimentally show that it performs much better in the case of real-world networks, finding the radius and diameter values after 10–100 runs of Breadth First Search instead of $|V|$ BFS-s (independently of the value of $|V|$), and thus having running time $O(|E|)$ in practice. As far as we know, this is the first algorithm able to compute the diameter of weakly connected digraphs, apart from the naive algorithm, which runs in time $\Omega(|V||E|)$ performing a BFS from each node. In the particular cases of strongly connected directed or connected undirected graphs, we have compared our algorithm with known approaches by performing experiments on a dataset composed by several real-world networks of different kinds. These experiments show that, despite its generality, the new algorithm outperforms all previous methods, both in the radius and in the diameter computation, both in the directed and in the undirected case, both in average running time and in robustness. Finally, as an application example, we have used the new algorithm to determine the solvability over time of the "Six Degrees of Kevin Bacon" game, and of the "Six Degrees of Wikipedia" game. As a consequence, we have computed for the first time the exact value of the radius and the diameter of the whole Wikipedia digraph.

INFINE Team

6. New Results

6.1. Online Social Networks (OSN)

Community detection; bandit algorithms; privacy preservation; reward mechanisms

6.1.1. Community detection

Participants: Laurent Massoulié, Marc Lelarge, Charles Bordenave.

We have progressed in the design of spectral methods for community detection and in the corresponding analysis, in particular by proving the so-called spectral redemption conjecture. This has been published in IEEE FOCS'15. The abstract of the paper is as follows. A non-backtracking walk on a graph is a directed path such that no edge is the inverse of its preceding edge. The nonbacktracking matrix of a graph is indexed by its directed edges and can be used to count non-backtracking walks of a given length. It has been used recently in the context of community detection and has appeared previously in connection with the Ihara zeta function and in some generalizations of Ramanujan graphs. In this work, we study the largest eigenvalues of the non-backtracking matrix of the Erdős-Rényi random graph and of the Stochastic Block Model in the regime where the number ´ of edges is proportional to the number of vertices. Our results confirm the "spectral redemption conjecture" that community detection can be made on the basis of the leading eigenvectors above the feasibility threshold.

6.1.2. Bandit algorithms for active learning of content type at low spam cost

Participants: Laurent Massoulié, Mesrob Ohanessian, Alexandre Proutière.

Progress on "bandit algorithms" for targeted news dissemination. We developed a framework in which to cast the problem, and the so-called "greedy Bayes" algorithm to determine which user to expose to a given content. We proved corresponding optimality properties, and observed that "greedy Bayes" beats the so-called Thompson sampling approach, that is the state-of the-art method in bandit problems. This work was published at ACM Sigmetrics'15.

6.1.3. Clustering and Inference From Pairwise Comparisons

Participants: Rui Wu, Jiaming Xu, Srikant Rayadurgam, Marc Lelarge, Laurent Massoulié, Bruce Hajek.

In a short publication at ACM Sigmetrics'15, we do the following. Given a set of pairwise comparisons, the classical ranking problem computes a single ranking that best represents the preferences of all users. In this paper, we study the problem of inferring individual preferences, arising in the context of making personalized recommendations. In particular, we assume users form clusters; users of the same cluster provide similar pairwise comparisons for the items according to the Bradley-Terry model. We propose an efficient algorithm to estimate the preference for each user: first, compute the net-win vector for each user using the comparisons; second, cluster the users based on the net-win vectors; third, estimate a single preference for each cluster separately. We show that the net-win vectors are much less noisy than the high dimensional vectors of pairwise comparisons, therefore our algorithm can cluster the users reliably. Moreover, we show that, when a cluster is only approximately correct, the maximum likelihood estimation for the Bradley-Terry model is still close to the true preference.

6.2. Spontaneous Wireless Networks and Internet of Things

internet of things; wireless sensor networks; dissemination; resource management

6.2.1. Platform Design for the Internet of Things

Participants: Emmanuel Baccelli, Cedric Adjih, Oliver Hahm, Matthias Waehlisch, Thomas Schmidt, Hauke Petersen.

Within this activity, we have further developed the platforms we champion for the Internet of Things: the open source operating system RIOT and open-access IoT-lab testbeds. RIOT now aggregates open source contributions from 120+ people (and counting) from all over the world, coming both from academia and from industry, and received financial backing from top companies including Cisco and Google in 2015. Revisiting concepts from the early Internet, we have designed and introduced a new software architecture that fits the (memory, CPU, energy) constraints of low-end IoT devices, while being full-featured and easily extensible, thus more future-proof that state of the art. This work was published in ACM MobiSys'15 (IoT-Sys workshop), and released as open source code, integrated in the latest version of RIOT 2015-12. We have also designed a distributed test framework which supports advanced continuous integration techniques, allows for the integration of project contributors to volunteer hardware and software resources to the test system, and can function as a permanent distributed plugtest for network interoperability testing. This work was published in ACM MobiSys'15 (IoT-Sys workshop). Concerning IoT-lab, we have contributed to the completion of the design and the roll-out of IoT-lab testbeds in multiple sites in France and started deploying an additional one in Berlin. Description of completed work and design was published in IEEE IoT-WF'15.

6.2.2. Standards for Spontaneous Wireless Networks

Participant: Emmanuel Baccelli.

Within this activity, we have contributed to new network protocol standards for spontaneous wireless networking, applied to ad hoc networks and the Internet of Things. In particular, collaborating with Fraunhofer, we have published Directional Airtime Metric (DAT), a new wireless metric standard targeting wireless mesh networks. The standard is in the RFC editor's queue (which means the corresponding IETF standard, an RFC, will be published within weeks). Furthermore, collaborating with ARM and Sigma Designs, we published RFC 7733, which provides guidance in the configuration and use of protocols from the RPL protocol suite to implement the features required for control in building and home environments. In collaboration with various industrial partners, with have also published a number of other Internet drafts, including an analysis of the characteristics of multi-hop ad hoc wireless communication between interfaces in the context of IP networks, and an analysis of the challenges of information-centric networking in the Internet of Things.

6.3. Resource and Traffic Management

Traffic offloading; infrastructure deployment; opportunistic routing; traffic modeling; intermittently connected networks.

6.3.1. On the Interaction between Content Caching and Routing

Participants: Kolar Purushothama Naveen, Laurent Massoulié, Emmanuel Baccelli, Aline Carneiro Viana, Don Towsley.

Nowadays Internet users are mobile over 60% of their time online, and mobile data traffic is expected to increase by more than 60% annually to reach 15.9 exabytes per month by 2018. This evolution will likely incur durably congested wireless access at the edge despite progress in radio technologies. To alleviate congestion at the Internet edge, one promising approach is to target denser deployments of wireless access points. As a result, mobile users are potentially within radio reach of several access points (AP) from which content may be directly downloaded. In this context, distinct AP's can have very different bandwidth and memory capacities. Such differences raise the following question: When requests can be sent to several such access points, how to optimize performance through both load balancing and content replication?

In this work, we introduce formal optimization models to address this question, where bandwidth availability is represented via a cost function, and content availability is represented either by a cost function or a sharp constraint. For both formulations we propose dynamic caching and request assignment algorithms. Crucially our request assignment scheme is based on a server price signal jointly reflecting content and bandwidth availability. Using mean field approximation and Lyapunov functions techniques, we prove that our algorithms are optimal and stable in a limiting fluid regime with large arrival rates and content chunking. Through simulations we exhibit the efficacy of our request assignment strategy in comparison to the common practices

of assigning requests purely based on either bandwidth or content availability. Finally, using the popular LRU (Least Recently Used) strategy instead for cache replacements, we again demonstrate the superior performance of our request assignment strategies. This work was published in the ACM SIGCOMM'15 workshop on All Things Cellular.

6.3.2. From Routine to Network Deployment for Data Offloading in Metropolitan Areas

Participants: Eduardo Mucceli, Aline Carneiro Viana.

Smartphone sales are booming, nearly half billion were sold in 2011; more smartphones, more mobile data traffic, and Currently, 3G cellular networks in metropolitan areas are struggling to attend the recent boost up of mobile data consumption. Carefully deploying WiFi hotspots allow to maximize WiFi offloading and can both be cheaper than upgrade the cellular network structure and concede substantial improvement in the network capacity. In this context, in this work, we first propose a new way to map into a graph the *people behavior* (i.e., mobility context) in an urban scenario. Our proposed behavior-to-graph solution is simple, take into consideration the restrictions imposed by transportation modes to traffic demand, the space-time interaction between people and urban locations, and finally, is powerful to be used as input to any popular area identification problem (key points for an efficient network planning). Secondly, we propose a metric to identify locations more capable of providing coverage for people and consequently, more suitable for receiving hotspots. Deploying a small percentage of hotspots ranked by the herein proposed metric provides high percentages of coverage time for people moving around in the city. Using a real-life metropolitan trace, we show our routine-based strategy guarantees higher offload ratio than current approaches in the literature while using a realistic traffic model. This work, including new characterization results of the used trace and new analysis of space-traffic correlation, is under submission in a trasaction.

6.3.3. Mobile Data Traffic Modeling: Revealing Hidden Facets

Participants: Eduardo Mucceli, Aline Carneiro Viana, Kolar Purushothama Naveen, Carlos Sarraute.

Smartphone devices provide today the best means of gathering users information about content consumption behavior on a large scale. In this context, the literature is rich in work studying and modeling users mobility, but little is publicly known about users content consumption patterns. The *understanding of users' mobile data traffic demands* is of fundamental importance when looking for solutions to manage the recent boost up of mobile data usage and to improve the quality of communication service provided. Hence, the definition of a *usage pattern* can allow telecommunication operators to better foreseen future demanded traffic and consequently, to better (1) deploy data offloading hotspots or (2) timely plan network resources allocation and then, set subscription plans.

Using a large-scale dataset collected from a major 3G network in a big metropolitan area, in this work, we present the first detailed measurement-driven modeling of mobile data traffic usage of smartphone subscribers. Our main outcome is a synthetic measurement-based mobile data traffic generator, capable of imitating traffic-related activity patterns of different categories of subscribers and time periods of a routinary normal day in their lives. For this, we first characterize individual subscribers routinary behaviour, followed by the detailed investigation of subscribers' usage pattern (i.e., "when" and "how much" traffic is generated). Broadly, our observations bring important insights into network resource usage. We then classify the subscribers into six distinct profiles according to their usage pattern and model these profiles according to two different journey periods: peak and non-peak hours. We show that the synthetic trace generated by our data traffic model consistently imitates different subscriber profiles in two journey periods, when compared to the original dataset. We discuss relevant issues in traffic demands and describe implications in network planning and privacy. This work, including a new characterization results of the used trace, including analysis correlating age and gender to traffic demands, as well as new profiling results, is under submission in a transaction.

6.3.4. Data Delivery in Opportunistic and Intermittently Connected Networks

Participants: Ana Cristina Vendramin, Anelise Munaretto, Myriam Delgado, Aline Carneiro Viana, Mauro Fonseca.

The pervasiveness of computing devices and the emergence of new applications and cloud services are factors emphasizing the increasing need for adaptive networking solutions. In most cases, this adaptation requires the design of interdisciplinary approaches as those inspired by nature, social structures, games, and control systems. The approach presented in this work brings together solutions from different, yet complementary domains, i.e., networking, artificial intelligence, and complex networks, and is aimed at addressing the problem of efficient data delivery in intermittently connected networks.

As mobile devices become increasingly powerful in terms of communication capabilities, the appearance of opportunistic and intermittently connected networks referred to as Delay Tolerant Networks (DTNs) is becoming a reality. In such networks, contacts occur opportunistically in corporate environments such as conferences sites, urban areas, or university campuses. Understanding node mobility is of fundamental importance in DTNs when designing new communication protocols that consider opportunistic encounters among nodes. This work proposes the Cultural Greedy Ant (CGrAnt) protocol to solve the problem of data delivery in opportunistic and intermittently connected networks. CGrAnt is a hybrid Swarm Intelligence-based forwarding protocol designed to address the dynamic and complex environment of DTNs. CGrAnt is based on: (1) Cultural Algorithms (CA) and Ant Colony Optimization (ACO) and (2) operational metrics that characterize the opportunistic social connectivity between wireless users. The most promising message forwarders are selected via a greedy transition rule based on local and global information captured from the DTN environment. Using simulations, we first analyze the influence of the ACO operators and CA knowledge on the CGrAnt performance. We then compare the performance of CGrAnt with the PROPHET and Epidemic protocols (two well known related protocols in the literature) under varying networking parameters. The results show that CGrAnt achieves the highest delivery ratio (gains of 99.12% compared with PROPHET and 40.21% compared with Epidemic) and the lowest message replication (63.60% lower than PROPHET and 60.84% lower than Epidemic). This work is under submission to an international journal.

6.3.5. Designing Adaptive Replication Schemes in Distributed Content Delivery Networks

Participants: Mathieu Leconte, Marc Lelarge, Laurent Massoulié.

In a paper published at the ITC'15 conference, we address the problem of content replication in large distributed content delivery networks, composed of a data center assisted by many small servers with limited capabilities and located at the edge of the network. We aim at optimizing the placement of contents on the servers to offload the data center as much as possible. We model the sub-system constituted by the small servers as a loss network, each loss corresponding to a request to the data center. Based on large system / storage behavior, we obtain an asymptotic formula for the optimal replication of contents and propose adaptive schemes to attain it by reacting to losses, as well as faster algorithms which can react before losses occur. We show through simulations that our adaptive schemes outperform significantly standard replication strategies both in terms of loss rates and adaptation speed.

6.3.6. Vehicular Network under a Social Perception

Participants: Felipe D. Cunha, Aline Carneiro Viana, Raquel A. F. Mini, Antonio A.f. Loureiro.

Vehicular Mobility is strongly influenced by the speed limits, destinations, traffic conditions, period of the day, and direction of the public roads. At the same time, the driver's behavior produces great influences in vehicular mobility. People tend to go to the same places, at the same day period, through the same trajectories, which le ad them to the appearance of driver's daily routines. These routines lead us to the study of mobility in VANETs under a social perspective and to investigate how effective is to explore social interactions in this kind of network. In this work, we thus characterize and evaluate social properties of a realistic vehicular trace found in literature. Our aim is to study the vehicles' mobility in accordance to social behaviors. Social metrics are computed and the obtained results are compared to random graphs. With our analysis, we could verify the existence of regularity and common interests among the drivers in vehicular networks.

After having identified routine in vehicles mobility patterns and their correlation with the period of the day, we then leverage the identified social aspects to design a *Socially Inspired Broadcast Data Dissemination* for VANETs. We claim that protocols and applications designed for Vehicular Ad Hoc Networks need to adapt to

vehicles routines in order to provide better services. With this issue in mind, we designed a data dissemination solution for these networks that considers the daily road traffic variation of large cities and the relationship among vehicles. The focus of our approach is to select the best vehicles to rebroadcast data messages according to social metrics, in particular, the clustering coefficient and the node degree. Moreover, our solution is designed in such a way that it is completely independent of the perceived road traffic density. Simulation results show that, when compared to related protocols, our proposal provides better delivery guarantees, reduces the network overhead and possesses an acceptable delay.

6.3.7. Design and Analysis of an Efficient Friend-to-Friend Content Dissemination System

Participants: Kanchana Thilakarathna, Aline Carneiro Viana, Aruna Seneviratne, Henrik Petander.

In this work, we focus on dissemination of content for delay tolerant applications/services, (i.e. content sharing, advertisement propagation, etc.) where users are geographically clustered into communities. Due to emerging security and privacy concerns, majority of users are becoming more reluctant to interact with strangers and are only willing to share information/content with the users who are previously identified as friends. As a result, despite its promise, opportunistic communications systems have not been widely adopted. In addition, in this environment, opportunistic communication will not be effective due to the lack of known friends within the communication range. We thus propose a novel architecture which combines the advantages of distributed decentralized storage and opportunistic communications. The proposed system addresses the trust and privacy concerns of opportunistic communications systems, and enables the provision of efficient distributed mobile social networking services. We exploit the fact that users will trust their friends, and the friends will help in disseminating content by temporarily storing and forwarding content. This can be done by replicating content on friends' devices who are likely to consume that content and provide the content to other friends when the device has access to low cost networks. The fundamental challenge then is to minimize the number of replicas, to ensure high and timely availability. We provide a formal definition of this content replication problem, and show that it is NP hard. Then, we propose a community based greedy heuristic algorithm with novel dynamic centrality metrics that replicates the content on a minimum number of friends' devices, and maximizes the availability of content. Using both real world and synthetic traces, we validate effectiveness of the proposed scheme. In addition, we demonstrate the practicality of the the proposed system, through an implementation on Android smartphones. This work is under submission in an international transaction.

6.3.8. Telling Apart Social and Random Relationships in Dynamic Networks

Participants: Pedro Olmo Vaz de Melo, Aline Carneiro Viana, Marco Fiore, Katia Jaffrès-Runser, Frédéric Le Mouël, Antonio A. F. Loureiro, Lavanya Addepalli, Guangshuo Chen.

Recent studies have analyzed data generated from mobile individuals in urban regions, such as cab drivers or students in large campuses. Particular attention has been paid to the dynamics of user movement, whose real-world complexity cannot be fully captured through synthetic models. Indeed, understanding user mobility is of fundamental importance when designing new communication protocols that exploit opportunistic encounters among users. In this case, the problem mainly lies in correctly forecasting future contacts. To that end, the regularity of daily activities comes in handy, as it enforces periodic (and thus predictable) space-time patterns in human mobility. Although human behavior is characterized by an elevated rate of regularity, random events are always possible in the routines of individuals. Those are hardly predictable situations that deviate from the regular pattern and are unlikely to repeat in the future.

We argue that the ability to accurately spot random and social relationships in dynamic networks is essential to network applications that rely on a precise description of human routines, such as recommendation systems, forwarding strategies and opportunistic dissemination protocols. We thus propose a strategy to analyze users' interactions in mobile networks where users act according to their interests and activity dynamics. Our strategy, named *Random rElationship ClASsifier sTrategy (RECAST)*, allows classifying users' wireless interactions, separating random interactions from different kinds of social ties. To that end, RECAST observes how the real system differs from an equivalent one where entities' decisions are completely random. We evaluate the effectiveness of the RECAST classification on five real-world user contact datasets collected in diverse networking contexts. Our analysis unveils significant differences among the dynamics of users' wireless

interactions in the datasets, which we leverage to unveil the impact of social ties on opportunistic routing. We show that, for such specific purpose, the relationships inferred by classifier are more relevant than, e.g., self-declared friendships on Facebook. This work was published in 2015 at the Performance Evaluation Elsevier Journal [9].

MADYNES Project-Team

7. New Results

7.1. Monitoring

7.1.1. Anonymous networks monitoring

Participants: Thibault Cholez [contact], Isabelle Chrisment, Olivier Festor.

In 2015, we pursued our collaboration with Juan Pablo Timpanaro a former team's PhD student and published a new paper [47] on the I2P anonymous network (http://i2p2.de). More precisely, we monitored I2P's decentralised directory, known as the netDB, and produced two contributions. On the one hand, we conducted arguably the first *churn* study of the I2P network, showing that I2P users are more stable than non-anonymous peer-to-peer users. On the other hand, we analysed the design of the netDB and compared it against the popular KAD design, demonstrating that the former is more vulnerable to different attacks, specially to Eclipse attacks, which can be mitigated by applying some safer design choices of the latter. We lately showed the positive impact on performance of including KAD's DHT configuration into the netDB in terms of bandwidth, storage and messages overhead.

7.1.2. Smartphone usage monitoring

Participants: Vassili Rivron [contact], Mohammad Irfan Khan, Simon Charneau [Inria], Isabelle Chrisment.

In [39] we presented some results from our study based on a combination of crowdsending and survey. We discussed some technical problems we faced and some lessons learned during our crowdsensing experiment. Furthermore we showed how information regarding social context can be used for better interpretation of crowdsensed data. Next we selected some questions from the multiple choice survey questionnaire and combined the responses with crowdsensed data to analyze users' perception about their smartphone usage and discussed cognitive factors associated with reporting information on questionnaires. Moreover we showed that combining sensing with survey can improve both the techniques and the combination has important use cases such as helping users to have a better understanding and control of their technology usage.

7.1.3. Active Monitoring

Participants: Abdelkader Lahmadi [contact], Jérôme François, Valentin Giannini, Frederic Beck [LHS], Bertrand Wallrich [LHS].

The main motivation of this work was to assess the exposition of industrial systems in the Internet, especially by measuring how many SCADA systems are accessible. To do so, we built an IPv4 methodology which is able to scan the entire IPv4 address space by maximizing the distance between consecutive IP addresses. It thus avoids colateral effect of overloading targeted networks and being blacklisted. We thus extend the Zmap tool (zmap.io) by also including other functionalities such as distributed scans, indexation and visualisation of the results [63]. First experiences have been performed and are under evaluation.

7.1.4. Sensor networks monitoring

Participants: Rémi Badonnel, Isabelle Chrisment, Olivier Festor, Abdelkader Lahmadi [contact], Anthea Mayzaud.

This year, our work on security-oriented monitoring has been centered on building a distributed architecture that supports passive monitoring in the Internet of Things using the RPL protocol [37]. A particular interest has been given to advanced metering infrastructure (AMI) networks, where higher order devices are expected to form the backbone infrastructure, to which more constrained nodes would connect. Our distributed architecture exploits the capabilities of these higher order devices to perform network monitoring tasks, and takes benefits from properties inherent to that protocol, such as DODAG building and multi-instance routing mechanisms, in order to passively monitor the environment with a minimal impact on constrained nodes.

We have also consolidated our taxonomy on security attacks in these networks [8]. In addition, we have pursued our work on topological inconsistency attacks [9]. It is evident from the experiments that we have conducted that mitigating such attacks is critical to avoid channel congestion and high resource usage. Our initial adaptive threshold (AT) strategy to mitigate the effects of such attacks has been further improved. The new strategy dynamically takes into account network characteristics in order to infer an appropriate threshold for counteracting these attacks.

7.2. Security

7.2.1. Security analytics

Participants: Jérôme François [contact], Abdelkader Lahmadi, Manobala Nirmala, Vincent Noyalet.

During the year 2015, we have extended our monitoring platform dedicated to Android environments [69] with more analytics features. The monitoring platform is dedicated to the collection, storage, analysis and visualization of logs and network flow data of mobile applications. The platform relies on a set of on-device probes to monitor network and system activities of these applications. The data are collected from these probes and parsed through generic and flexible collectors relying on Flume agents that we have adapted and extended. We are storing the collected data using a column oriented Hbase storage engine (Hadoop database). Finally, after being parsed, the data are made available within the Elasticsearch engine to search and visualize them using the Kibana tool. We have also presented the building blocks of the platform in a lab session within the conference AIMS 2015 [70].

We have also maintained an IETF draft [75] to promote a standardization effort towards the extension of IP Flow-based monitoring with geographic information. Associating Flow information with their measurement geographic locations will enable security applications to detect anomalous activities. In the case of mobile devices, the characterization of communication patterns using only time and volume is not enough to detect unusual location-related communication patterns.

Besides, we looked at aggregating flows collected at the High Security Lab since a single attack is represented by multiple flows. For example, a DDoS or a scan is a sequence of similar parallel flows coming from the same or distributed machines. As attacks occur very frequently and even at the same time, grouping flows occurring in a pre-defined time window is not a valid approach. Two approaches have been investigated and are actually dependent of the sources of collected flows. First, we analyzed collected Netflow data from the Darknet which is basically a sinkhole without any services running or announced. Hence, all traffic is considered as abnormal and is limited to a set of predefined attacks. Indeed, since no packets can be sent back, complex attacks with different steps cannot be caught. Therefore, scanning, flooding-based denial-of-service and backscatter are the main types of anomalies we can observe. Flows are thus grouped and labeled regarding certain criteria (common IP addresses/subnets, ports, co-occurrence) thanks to a pre-established decision process [58]. The final goal was to compare data collected in Nancy and in Tokyo. Secondly, we assume flow data without specific knowledge about the type of traffic it embeds. In such a case, the goal is to automatically extract recurrent patterns. The initial approach consisted in representing flows as nodes in a graph and linking them when sharing some properties (IP addresses, ports). Major subsequent problems have been faced like indexation, split flows in multiple files and visualization [59].

7.2.2. Management of HTTPS traffic

Participants: Thibault Cholez [contact], Shbair Wazen, Jérôme François, Isabelle Chrisment.

We previously investigated the latest technique for HTTPS traffic filtering that is based on the Server Name Indication (SNI) field of TLS and which has been recently implemented in many firewall solutions. We showed that SNI has two weaknesses, regarding (1) backward compatibility and (2) multiple services using a single certificate and we implemented a proof of concept of these vulnerabilities as a web browser extension (Escape). This work was published in the IFIP/IEEE IM'15 conference [44].

This led us to the development of new reliable methods to investigate the increasing number of HTTPS traffic that may hold security breaches but without relying on decryption at any step, in order to respect users' privacy (no HTTPS proxy). Many approaches already identify the main type of an application (Web, P2P, SSH,..) running in secure tunnels, and others identify a couple of specific encrypted web pages through website fingerprinting.

In this context, we developed a better technique to precisely identify the services run within HTTPS connections, i.e. to name the services, without relying on specific header fields that can be easily altered. We have defined dedicated features for HTTPS traffic that are used as input for a multi-level identification framework based on machine learning algorithms. Our evaluation based on real traffic shows that we can identify encrypted web services with a high accuracy. This work will be published next year in the IFIP/IEEE Network Operations and Management Symposium (NOMS 2016).

7.2.3. Configuration security automation

Participants: Rémi Badonnel [contact], Gaetan Hurel, Abdelkader Lahmadi, Olivier Festor.

Our work during year 2015 was mainly focused on the orchestration of security functions in the context of mobile smart environments [35]. Most of current security approaches for these environments are provided in the form of applications or packages to be directly installed on the devices themselves. Such approaches may be qualified as on-device. However, on-device approaches generally induce significant local resource consumption leading to the significant reduction of battery lifetime. In the meantime, current cloud-based approaches for mobile security attempt to deal with this issue by offloading most of the workload on a remote server, but may introduce significant additional latency. In that context, we have pursued the efforts on our strategy for dynamically outsourcing and composing security functions in the cloud, considering software-defined networking. The architecture relies on a set of security functions that are activated, configured and orchestrated according to the current contexts and risks, while a dedicated modelling has been introduced for supporting the evaluation of security compositions and their properties. The chaining of security functions is performed dynamically in order to fit with the security requirements of mobile devices at runtime. In particular, we have proposed in [35] to analyze and cluster applications running on the mobile devices based on their network behaviors, in order to drive the selection and deployment of adequate security compositions that may be fully outsourced or split between in-cloud and on-device.

We have also investigated in [23] to what extent security automation, more specifically in the context of vulnerability management, might be supported by conceptual knowledge discovery. The intended extension might be a mean to cope with the increasing dynamics and complexity of networked environments. Most current security solutions still seem to work under certain boundaries that prevent them to act intelligently and flexibly, i.e. strictly sticked to the available security information in order to analyze, report and eventually remediate found problems. Our purpose is to exploit methods and techniques coming from formal concept and knowledge discovery in databases, in order to provide high-level automation based on mechanisms capable of understanding, reasoning about, and anticipating the surrounding environment and its vulnerabilities.

7.3. Experimentation, Emulation, Reproducible Research

This section covers our work on experimentation on testbeds (mainly Grid'5000), on emulation (mainly on Distem), and on Reproducible Research.

7.3.1. Grid'5000 design and evolutions

Participants: Jérémie Gaidamour, Arthur Garnier, Lucas Nussbaum [contact], Clément Parisot.

The team was again heavily involved in the evolutions and the governance of the Grid'5000 testbed.

In the context of ADT LAPLACE, Jérémie Gaidamour adapted and configured the CiGri middleware on Grid'5000. CiGri enables the execution of large campaigns of *best-effort* jobs (low priority, interruptible jobs). It is expected that this work will allow the remaining free time slots to be filled by tasks from other research communities such as natural language processing.

Jérémie Gaidamour also greatly improved *stats5k*, our tool to generate metrics about the testbed (usage, resources availability, etc.), available at https://intranet.grid5000.fr/stats/.

Arthur Garnier added the testing of Grid'5000 tutorials to our continuous integration installation, enabling the earlier detection of problems on the testbed. He then led the migration to PostgreSQL as the backend for the OAR batch scheduler – a behind-the-scenes but major migration.

In addition to daily administratrive duties and to his work on Kwapi described below (section 7.3.2), Clément Parisot added support for *production* workloads to Grid'5000, extending the scope of the testbed to make it more suitable for additional user communities. He then managed the installation of the new clusters at Nancy, purchased in the context of OIP Grid'5000 and CPER CyberEntreprises.

Finally, in addition to his roles in the *bureau*, *comité d'architectes* and *comité des responsables de sites* of Grid'5000, Lucas Nussbaum managed the purchase of the new clusters at Nancy mentioned above, and gave several presentations about the testbed, at *Journées SUCCES* [14], at *Retour d'expéRiences sur la Recherche Reproductible* [15], and at *École Cumulo Numbio*.

7.3.2. A unified monitoring framework for energy consumption and network traffic

Participants: Lucas Nussbaum [contact], Clément Parisot.

Providing experimenters with deep insight about the effects of their experiments is a central feature of testbeds, that Grid'5000 was only partially addressing. We designed Kwapi, a framework that unifies measurements for both energy consumption and network traffic. Because all measurements are taken at the infrastructure level (using sensors in power and network equipment), using this framework has no dependencies on the experiments themselves. Initially designed for OpenStack infrastructures, the Kwapi framework allows monitoring and reporting of energy consumption of distributed platforms. In this work, we extended Kwapi to network monitoring, and overcame several challenges: scaling to a testbed as large as Grid'5000 while still providing high-frequency measurements; providing long-term loss-less storage of measurements; handling operational issues when deploying such a tool on a real infrastructure.

This work was published at Tridentcom [31] and presented in a GENI/FIRE collaboration workshop [12]. It is now in production as the default monitoring framework on Grid'5000.

7.3.3. Comparison of HPC and Clouds testbeds

Participant: Lucas Nussbaum [contact].

Given the recent launch of two large NSF-funded projects that share similar goals as Grid'5000 (CloudLab and ChameleonCloud), we worked on analyzing the design choices made so far by those projects, comparing them with Grid'5000. Preliminary results were presented at REPPAR [17] and at a GENI/FIRE collaboration workshop [13].

7.3.4. Emulation with Distem

Participants: Emmanuel Jeanvoine, Lucas Nussbaum [contact], Cristian Ruiz.

Several improvements have been made around Distem, mostly in the context of ADT COSETTE.

During the internship of Arthur Carcano, we tried to use Distem to experiment on NDN infrastructures. We obtained promising results, especially in terms of scale. We plan to continue this work and publish it in 2016.

We also submitted, to CCGRID, a paper demonstrating the use of Distem to evaluate fault tolerance and load balancing strategies implemented in Charm++. This submission is still pending evaluation.

Finally, in an effort to validate Distem performance, we studied the performance of Container-based virtualization technologies such as LXC or Docker, as most of the underlying technology is also shared with Distem. We studied their performance in the context of HPC, and showed that containers technology has matured over the years, and that performance issues are being solved. This work has been published at VHPC [43].

7.3.5. Management of large-scale experiments

Participants: Emmanuel Jeanvoine, Lucas Nussbaum [contact], Cristian Ruiz.

Following our survey of experiment management tools [7] accepted at FGCS at the end of 2014 and published early this year, we worked on Ruby-Cute, a library that aggregates various useful functionality in the context of such tools. We hope that it will be useful as a basis for future tools, and ease testing of new ideas in that field. The library is available on http://ruby-cute.github.io/.

7.3.6. Tracking provenance in experiment control tools

Participants: Tomasz Buchert, Lucas Nussbaum [contact].

In the context of our work on XPFlow, we worked on the collection of provenance during experiments. We surveyed provenance collection in various domains of computer science, introduced a new classification of provenance types suited to distributed systems experiments, and proposed a design of a provenance system inspired by this classification. This work has been published at REPPAR [29].

7.3.7. Reproducible Research

Participant: Lucas Nussbaum [contact].

Lucas Nussbaum gave a presentation on Reproducible Research[16] at the ICube laboratory seminar (Strasbourg). A shorter version of the talk was given to the Inria *Comité des projets* in Nancy.

Lucas Nussbaum also co-organized the second edition of REPPAR, a workshop on Reproducibility in Parallel Computing, held in conjunction with Euro-Par'2015.

7.4. Routing

7.4.1. Routing in Wireless Sensor Networks

Participants: Emmanuel Nataf [contact], Patrick-Olivier Kamgueu, Nesrine Khelifi.

We have formalized our previous work on the routing protocol for wireless sensor network by fuzzy logic specifications. The rules of routing metric composition are now valid for any network depth and we demonstrated its quality by real experimentation [36]. This work is done in the context of the associated team we build with the Cameroun and the Inria international lab LIRIMA.

For potentially very large wireless sensor network, our routing or any other routing, can not limit traffic bottleneck near the network root. Network depth should also be reduced as hop by hop communication is a factor which strongly increases data loss rate. Considering theses problems Nesrine Khelifi PhD student of the Manouba University in Tunisia spent 3 months within the Madynes team trying to limit the depth of the network by splitting it under the supervision of network quality observers we had to define.

7.4.2. Operator calculus based routing in Wireless Sensor Networks

Participants: Evangelia Tsiontsiou, René Schott, Stacey Staples [Southern Illinois University Edwardsville], Jamilla Benslimane, Bilel Nefzi, Ye-Qiong Song [contact].

Recently, Operator calculus (OC) has been developed by Schott and Staples with whom we collaborate. We make use of OC methods on graphs to solve path selection in the presence of multiple constraints. Based on OC, we developed a distributed algorithm for path selection in a graph. This approach has been applied to efficiently solve a joint routing, channel and time slot scheduling optimization problem in UWB wireless sensor networks [6]. We also designed a new routing protocol which makes use of this algorithm: the Operator Calculus based Routing Protocol (OCRP). In OCRP, a node selects the set of eligible next hops based on the given constraints and the distance to the destination. It then sends the packet to all eligible next hops. The protocol is implemented in Contiki OS (Rime profile) and emulated for TelosB motes using Cooja. We compared its performance against tree and directional flooding routing and showed the advantages of our technique [28]. Our ongoing work consists in its comparison with RPL to show its practical contribution to handle simultaneously several IETF ROLL routing metrics. This work is part of Lorraine AME Satelor project granted by Lorraine Region.

7.4.3. Probabilistic Energy-Aware Routing for Wireless Sensor Networks

Participants: Evangelia Tsiontsiou, Bernardetta Addis, Alberto Ceselli [Universita degli Studi di Milano], Ye-Qiong Song [contact].

Healthcare applications are considered as promising fields for Wireless Sensor Networks (WSNs). Thanks to WSNs, patients can be monitored in hospitals or smart home environments, providing health improvement, or emergency care. A key issue is the limited battery of sensors; indeed, current WSN research trends for healthcare applications include energy efficient routing and network lifetime guarantee mechanisms, among others. One of our ongoing work consists in designing a Smart Probabilistic Energy-Aware Routing Protocol (SPEAR) for WSNs which aims at maximizing the network lifetime by keeping low energy consumption and balancing network traffic between nodes. Our experimental campaign reveals that our SPEAR protocol outperforms the popular Energy Aware Routing Protocol (EAR) from the literature, proving to be more effective in extending the network lifetime. This work has resulted in a conference submission. It is part of Lorraine AME Satelor project granted by Lorraine Region.

7.4.4. Energy-aware IP networks management

Participants: Bernardetta Addis [contact], Giuliana Carello [DEIB, Politecnico di Milano, Italy], Antonio Capone [DEIB, Politecnico di Milano, Italy], Luca Gianoli [Polytechnique de Montreal, Canada], Sara Mattia [IASI, CNR, Roma, Italy], Brunide Sansò [Polytechnique de Montreal, Canada].

The focus of our research is to minimize the energy consumption of the network through a management strategy that selectively switches off devices according to the traffic level. We consider a set of traffic scenarios and jointly optimize their energy consumption assuming a per-flow routing. We propose a traffic engineering mathematical programming formulation based on integer linear programming that includes constraints on the changes of the device states and routing paths to limit the impact on quality of service and the signaling overhead.

A very important issue that may be affected by green networking techniques is resilience to node and link failures. We thus extended the optimization models to guarantee network survivability. Results show that significant savings, up to 30%, may be achieved even when both survivability and robustness are fully guaranteed.

Computational cost of proposed models can be very high when dealing with large size instances (network size and/or number of demands). For this reason, we proposed and tested different problem formulations with the aim of solving larger size instances at optimality. We focus on a particular form of shared protection mechanism, where energy consumption is associated only to active devices during normal functioning. We propose a standard and a projected formulation, with additions of cuts and valid inequalities. Computational results show that the projected formulation is very effective [20]. We plan to extend the work to consider multiperiod scenarios.

7.4.5. Virtual Network Functions Placement and Routing Optimization

Participants: Bernardetta Addis [contact], Dallal Belabed [LIP6, Univ Paris 06, France], Mathieu Bouet [Thales Communications & Security, France], Stefano Secci [LIP6, Univ Paris 06, France].

Network Functions Virtualization (NFV) is incrementally deployed by Internet Service Providers (ISPs) in their carrier networks, by means of Virtual Network Function (VNF) chains, to address customers' demands. The motivation is the increasing manageability, reliability and performance of NFV systems, the gains in energy and space granted by virtualization, at a cost that becomes competitive with respect to legacy physical network function nodes. From a network optimization perspective, the routing of VNF chains across a carrier network implies key novelties making the VNF chain routing problem unique with respect to the state of the art: the bitrate of each demand flow can change along a VNF chain, the VNF processing latency and computing load can be a function of the demands traffic, VNFs can be shared among demands, etc. We started our work providing an integer linear programming model for Virtual Network Functions Placement and demand rerouting. By extensive simulation on realistic ISP topologies, we draw conclusions on the trade-offs achievable between legacy Traffic Engineering (TE) ISP goals and novel combined TE-NFV goals [19].

7.4.6. Composing IoT protocols with Named-Data Networking

Participants: Salvatore Signorello [University of Luxembourg], Olivier Festor [contact], Radu State [University of Luxembourg].

With the emergence of IoT, many layer 2 protocols have been proposed with each of them its own characteristics, advantages and drawbacks. Choosing a protocol often depends on the global context, as for example number of users, time of the day... Although devices can now be fitted with multiple interfaces, using always the same specific layer 2 protocol is not efficient, in particular if we assume that connected devices are retrieving or exchanging similar contents. For example, assuming that WiFi is the most usable interface to download some files in Internet through an access point may not be ideal if a close-by device accessible by Bluetooth already has it. To accommodate so multiple layer 2 protocols, we propose to leverage the Named-Data Networking (NDN) paradigm which allows to explore in parallel multiple paths for retrieving content independently of the underlying protocol. Our first results [46] show that such a theoretical solution cannot work practically. Indeed, applying NDN in a blind mode over multiple layer 2 protocols does not assume the corresponding specificities like for example various collision rates depending on the underlying protocols, which have to be taken into account.

7.5. Multi-modeling and co-simulation

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Yannick Presse, Victorien Elvinger, Julien Vaubourg, Alexandre Tan, Benjamin Segault, Emmanuel Nataf.

Vincent Chevrier (former Maia team, Dep 5, LORIA) is a collaborator and the correspondant for the MS4SG project, Benjamin Camus, and Christine Bourjot (former MAIA team, Dep 5, LORIA) are collaborators for AA4MM/MECSYCO. Julien Vaubourg's PhD is under the co-direction of V. Chevrier and L. Ciarletta.

In Pervasive or Ubiquitous Computing, a growing number of communicating/computing devices are collaborating to provide users with enhanced and ubiquitous services in a seamless way.

These systems, embedded in the fabric of our daily lives, are complex: numerous interconnected and heterogeneous entities are exhibiting a global behavior impossible to forecast by merely observing individual properties. Firstly, users physical interactions and behaviors have to be considered. They are influenced and influence the environment. Secondly, the potential multiplicity and heterogeneity of devices, services, communication protocols, and the constant mobility and reorganization also need to be addressed. Our research on this field is going towards both closing the loop between humans and systems, physical and computing systems, and taming the complexity, using multi-modeling (to combine the best of each domain specific model) and co-simulation (to design, develop and evaluate) as part of a global conceptual and practical toolbox.

We proposed the AA4MM meta-model [76] that solves the core challenges of multimodeling and simulation coupling in an homogeneous perspective. In AA4MM, we chose a multi-agent point of view: a multi-model is a society of models; each model corresponds to an agent and coupling relationships correspond to interaction between agents. In the MS4SG (Multi Simulation for Smart Grids) projet which involves some members of the former MAIA team, Madynes and EDF R&D on smart-grid simulation, we developed a proof of concepts for a smart-appartment case that serves as a basis for building up use cases.

In 2015 we worked on the following research topics:

- Assessment and evaluation of complex systems.
- Cyber Physical Systems

  We have led the design and implementation of the Aetournos platform at Loria. The collective movements of a flock of flying communicating robots / UAVs, evolving in potentially perturbed environment constitute a good example of a Cyber Physical System. Applying co-simulation technique we plan to develop a hybrid "network-aware flocking behavior" / "behavior aware routing protocol".

We have provided a working set of tools: multi-simulation behavior / network / physics and generic software development using ROS (Robot Operating System). The UAVs carry a set of sensors for location awareness, their own computing capabilities and several wireless networks.

The effort put in the UAVs gathers academic and research ressources from the Aetournos platform, the R2D2 ADT and the 6PO project, while applied, industrial and more R&D projects have been pursued this year (Outback Joe Search and Rescue Challenge, Alerion, Hydradrone) .

- MS4SG to link multi-simulations tools such as HLA (High Level Architecture) and FMI (Functional Mockup Interface) thanks to our AA4MM framework. We have so far successfully applied our solution to the simulation of smart apartment complex and to combine the electrical and networking part of a Smart Grid. The AA4MM software has seen major improvements in 2015 thanks to the ressources provided by the MS4SG project and a Carnot engineer financed thanks to Inria. It has been renamed as MECSYCO (http://www.mecsyco.com).

  Starting from domain specific and heterogenous models and simulators, the MECSYCO suite allows for multi *systems* integration at several levels: conceptual, formal and software. A couple of visualization tools have been developed as proof of concepts both at run-time and post-mortem.

  We have developed software components and plugins that interconnects within MECSYCO heterogeneous simulators from different domains: FMU (working with the 1.0 and 2.0 FMI standard for CoSimulation) ou non-FMU such as NS3 or Omnet++.

  Several EDF oriented use cases have demonstrated multi-simulations.

In additon to technical reports, several publications have been accepted in 2015 on these subjects [51], [49] and [48].

7.6. Pervasive or Ubiquious Computing

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Emmanuel Nataf, Thomas Paris, Quentin Houbre, Benjamin Segault, Jonathan Arnault, Eric Perlinski, Antoine Richard.

In Pervasive or Ubiquitous Computing, a growing number of communicating/computing devices are collaborating to provide users with enhanced and ubiquitous services in a seamless way.

These systems, incresingly numerous and heterogeneous, are embedded in the fabric of our daily lives. Our initial subject of interest is to study them with regards to their complexity: Those numerous interconnected and heterogeneous entities are exhibiting a global behavior impossible to forecast by merely observing individual properties.

Firstly, users physical interactions and behaviors have to be considered. They are influenced and influence their surroundings and the environment. Secondly, the potential multiplicity and heterogeneity of devices, services, communication protocols, and the constant mobility and reorganization also need to be addressed.

Our research on this field is going towards both closing the loop between humans and systems, physical and computing systems, and taming the complexity, using multi-modeling (to combine the best of each domain specific model) and co-simulation (to design, develop and evaluate) as part of a global conceptual and practical toolbox.

During some exploratory work, we have seen the potential of these Pervasive Computing ressources in the (Very Serious) Gaming area.

In 2015 we worked on the following topics:

- Cyber Physical Systems

  We pursuied the design and implementation of the Aetournos platform at Loria. The collective movements of a flock of flying communicating robots / UAVs, evolving in potentially perturbed environment constitute a good example of a Cyber Physical System. Eventually, we applied co-simulation technique and plan to develop a hybrid "network-aware flocking behavior" / "behavior aware routing protocol".

We developed a working set of tools: multi-simulation behavior / network / physics and generic software development using ROS (Robot Operating System). The UAVs carry a set of sensor for location awareness, their own computing capabilities and several wireless networks.

The effort put in the UAVs gathers academic and research ressources from the Aetournos platform, the Inria ADT R2D2 and the 6PO project, while applied, industrial and more R&D projects have been pursued this year (Medical Express / Outback Joe Search and Rescue Challenge, Alerion, Hydradrone, and a CIFRE PhD with Thales) .

- Smart * (MS4SG)

  We have studied scientific problems around model and simulator composition. We have also looked into practical and implementation issues in the frame of our MECSYCO /AA4MM solutions. We have added to our Smart Grid scenarios a smart appartment complex use case.

- (Very Serious) Gaming: Starburst Gaming

7.7. Quality-of-Service

7.7.1. Self-adaptive MAC protocol for both QoS and energy efficiency

Participants: Kévin Roussel, Shuguo Zhuo, Olivier Zendra, Ye-Qiong Song [contact].

Three main contributions have been made this year. Firstly iQueue-MAC has been extended to work on both single channel mode and multi-channel mode, improving its throughput performance [11]. Secondly, S-CoSenS and iQueue-MAC our previously designed protocols have been implemented on RIOT OS over MSP430-based motes. Our contribution consists in developing a port of RIOT OS on the MSP430 micro-controller and demonstrating that RIOT OS offers rich and advanced real-time features, especially the simultaneous use of as many hardware timers as the underlying platform (micro-controller), which are fundamental features to implement high performance MAC protocols [41]. The Cooja/MSPSim network simulation framework is widely used for developing and debugging, but also for performance evaluation of WSN projects. Our third contribution shows that Cooja is not limited only to the simulation of the Contiki OS based systems and networks, but can also be extended to perform simulation experiments of other OS based platforms, especially that with RIOT OS. Moreover, when performing our own simulations with Cooja and MSPSim, we observed timing inconsistencies with identical experimentations made on actual hardware. Such inaccuracies clearly impair the use of the Cooja/MSPSim framework as a performance evaluation tool, at least for time-related performance parameters. The detailed results of our investigations on the inaccuracy problems, as well as the consequences of this issue, and possible ways to fix or avoid it are available in [42]. Part of this work has been supported by PIA LAR project.

7.7.2. End-to-end delay modeling and evaluation in wireless sensor networks

Participants: François Despaux, Abdelkader Lahmadi, Ye-Qiong Song [contact].

Probabilistic end-to-end performance guarantee may be required when dealing with real-time applications. As part of ANR QUASIMODO project, we are dealing with Markov modeling of multi-hop networks running duty-cycled MAC protocols. One of the problems of the existing Markovian models resides in their strong assumptions that may not be directly used to assess the end-to-end delay in practice. In particular, realistic radio channel, capture effect and OS-related implementation factors are not taken into account. We proposed to explore a new approach combining code instrumentation and Markov chain analysis. In [32] we propose a novel approach to obtain the Markov chain model of sensor nodes by means of Process Mining techniques through the analysis of MAC protocol execution traces for a given traffic scenario. End to end delay is then computed based on this Markov chain. Experimentations were done using IoT-LAB testbed platform. Comparisons in terms of delay have been presented for two different metrics of the RPL protocol (hop count and ETX). The overall approach and its generalization using non-linear regression techniques in terms of traffic rate are detailed in the PhD thesis of François Despaux defended in September 2015 [1].

7.7.3. Dynamic resource allocation in network virtualization

Participants: Mohamed Said Seddiki, Mounir Frikha [SupCom, Tunis, Tunisie], Ye-Qiong Song [contact].

This work has been carried out as part of a co-supervised PhD thesis between University of Lorraine and SupCom Tunis.

The objective of this research topic is to develop different resource allocation mechanisms in Network Virtualization, for increasing the QoS guarantee. Firstly, we demonstrated the potential of SDN in the QoS management of a virtualized home network (VN). We proposed and implemented "FlowQoS", a mechanism that can be deployed by an Internet Service Provider in the last-mile hop or in the home gateway. Performance measurements show that this solution can share bandwidth between applications according to user-defined configuration to guarantee QoS for each active traffic. The second contribution is the modeling of the interaction between service providers and infrastructure providers using game theory framework to offer dynamic sharing of physical infrastructure across multiple VN with different QoS requirements. We presented a set of non-cooperative games to model the negotiation phase and the dynamic allocation of nodes and physical links for each deployed VN[10]. Finally we proposed a predictive approach that allows an adaptive control of bandwidth allocation in order to reduce the packet delays for a given VN on each physical link. The last two contributions offer dynamic sharing models of physical infrastructure resources while guaranteeing the QoS for each VN.

The overall approach is detailed in the PhD thesis of Said Seddiki defended in April 2015 [2].

7.7.4. QoS and fault-tolerance in distributed real-time systems

Participants: Florian Greff, Laurent Ciarletta, Arnauld Samama [Thales TRT], Eric Dujardin [Thales TRT], Ye-Qiong Song [contact].

The QoS must be guaranteed when dealing with real-time distributed systems interconnected by a network. Not only task schedulability in processors, but also message schedulability in networks should be analysed for validating the system design. Fault-tolerance is another critical issue that one must take into account. In collaboration with Thales TRT industrial partner as part of a CIFRE PhD work, we started a study on the real-time dependability of distributed multi-criticity systems interconnected by an embedded mesh network (RapidIO). For easing the QoS specification at the higher level, DDS middleware is used. We postulate that enhancing QoS for real-time applications entails the development of a cross-layer support of high-level requirements, thus requiring a deep knowledge of the underlying networks. This year, we proposed and implemented a new simulation/emulation/experimentation framework called ERICA, for designing such a feature. ERICA integrates both a network simulator (Ptolemy) and an actual hardware network to allow implementation and evaluation of different QoS-guaranteeing mechanisms. It also supports real-software-in-the-loop, i.e. running of real applications and middleware over these networks. Each component can evolve separately or together in a symbiotic manner, also making teamwork more flexible [68], [33].

7.7.5. Wireless sensor and actuator networks

Participants: Lei Mo, Xiufang Shi [Zhejiang University], Jiming Chen [Zhejiang University], Ye-Qiong Song [contact].

Wireless sensor and actuator networks provide a key technology for fully interacting within a CPS (Cyber-Physical System). However, the introduction of the mobile actuator nodes in a network rises some new challenging issues. In this context, we addressed two important issues: the multiple target tracking using both fixed and mobile sensors and the optimal scheduling of mobile wireless energy chargers (actuators) for fixed sensor nodes.

In our work, the data association problem in multiple target tracking is investigated. To reduce the computational complexity of traditional Joint Probabilistic Data Association (JPDA) algorithm, a modified JPDA algorithm is proposed to execute data association in multiple target tracking by utilizing the information of occlusion conditions, which is identified by a three-step algorithm. Simulation results show that the proposed algorithm has good tracking performance but low computational complexity [45].

We also investigated the multiple mobile chargers coordination problem that is minimizing the energy expenditure of the mobile chargers while guaranteeing the perpetual operation of the wireless sensor network. We formulated this problem as a mixed-integer linear program (MILP). To solve this problem efficiently, we proposed a novel decentralized method which is based on Benders decomposition. The multiple mobile chargers coordination problem is then decomposed into a master problem (MP) and a slave problem (SP), with the MP for mobile chargers scheduling and the SP for mobile chargers moving and charging time allocation. The convergence of proposed method is analyzed theoretically. Simulation results demonstrated the effectiveness and scalability of the proposed method [38].

7.7.6. Big Data-oriented networking

Participants: Jérôme François [contact], Lautaro Dolberg [University of Luxembourg], Thomas Engel [University of Luxembourg], Raouf Boutaba [University of Waterloo], Reaz Ahmed [University of Waterloo], Shihabur Rahman Chowdhury [University of Waterloo].

Performances of Big Data applications are tightly coupled with the performance of the network in supporting large data transfers. Deploying high-performance networks in data centers is thus vital but configuration and performance management as well as the usage of the network are of paramount importance. We thus surveyed helpful approaches in a book chapter [55]. This chapter starts by discussing the problem of virtual machine placement and its solutions considering the underlying network topology. It then provides an analysis of alternative topologies highlighting their advantages from the perspective of Big Data applications needs. In this context, different routing and flow scheduling algorithms are discussed in terms of their potential for using the network most efficiently. In particular, Software-Defined Networking relying on centralized control and the ability to leverage global knowledge about the network state is propounded as a promising approach for efficient support of Big Data applications.

7.8. Advanced Cache Management in Content-centric Networks

Participants: Thomas Silverston [contact], Cholez Thibault, Bernardini César, Aubry Elian, Chrisment Isabelle, Olivier Festor.

Information Centric Networking (ICN) has become a promising new paradigm for the future Internet architecture. It is based on named data, where content address, content retrieval and the content identification is led by its name instead of its physical location. One of the ICN key concepts relies on in-network caching to store multiple copies of data in the network and serve future requests, which helps reducing the load on servers, congestion in the network and enhances end-users delivery performances. Thus, the efficiency of the CCN architecture depends drastically on performances of caching strategies at each node. To date, there has been a lot of studies proposing new caching strategies to improve the performances of CCN. However, among all these strategies, it is still unclear which one performs better as there is a lack of common environment to compare these strategies. To this end, we compared the performances of CCN caching strategies within the same simulation environment. We build a common evaluation scenario and we compare via simulation five relevant caching strategies: Leave Copy Everywhere (LCE), Leave Copy Down (LCD), ProbCache, Cache “Less” For More and MAGIC. We analyze the performances of all the strategies in terms of Cache Hit, Stretch, Diversity and Complexity, and determine the cache strategy that fits the best with every scenario. This work has been published in IEEE Globecom 2015 [26].

At the meantime, CCN architecture uses *Interest* and *Data* messages to request and receive the data, and there has been no routing scheme to match a request to a specific content, as it is currently the case in the Internet. Indeed, CCN relies on flooding, which is a limitation for a future deployment at the Internet-scale. To this end, we proposed a Routing Scheme for CCN based on the softwarization (SDN). In our scheme SRSC, a controller gets knowledge of the network it administers as well as the content, and each node request the next hop to forward the Interest to their controller, until it reaches the closer Content Stores with the requested content. Nodes use a communication channel with the controller that relies only CCN messages and does not use the traditional SDN communication channel protocol Openflow over IP. The rationale is to help having CCN as a stand-alone new networking stack and to enforce its deployement without the IP infrastructure. This research work has been published in IEEE Netsoft 2015 [22] and Algotel 2015 [21].

MAESTRO Project-Team

7. New Results

7.1. Network Science

Participants: Eitan Altman, Konstantin Avrachenkov, Arun Kadavankandy, Jithin Kazhuthuveettil Sreedharan, Hlib Mykhailenko, Philippe Nain, Giovanni Neglia, Yonathan Portilla, Alexandre Reiffers-Masson.

7.1.1. Posting behavior in Social Networks and Content Active Filtering

In [57], Alexandre Reiffers-Masson and Eitan Altman in collaboration with Yezekael Hayel (UAPV), model the posting behavior in Social Networks in topics which have negative externalities, and propose content active filtering in order to increase content diversity. By negative externalities, it is meant that when the quantity of posted contents about some topic increases the popularity of posted contents decreases. They introduce a dynamical model to describe the posting behavior of users taking into account these externalities. Their model is based on stochastic approximations and sufficient conditions are provided to ensure its convergence to a unique rest point. They provide a closed form expression for this rest point. Content Active Filtering (CAF) are actions taken by the administrator of the Social Network in order to promote some objectives related to the quantity of contents posted in various topics. As objective of the CAF they consider maximizing the diversity of posted contents.

7.1.2. Network centrality measures

Recent papers studied the control of spectral centrality measures of a network by manipulating the topology of the network. In [56], Alexandre Reiffers-Masson, Eitan Altman and Yezekael Hayel (UAPV) extend these works by focusing on a specific spectral centrality measure, the Katz-Bonacich centrality. The optimization of the Katz-Bonacich centrality using a topological control is called the Katz-Bonacich optimization problem. The authors first prove that this problem is equivalent to a linear optimization problem. Thus, in the context of large graphs, one can use state-of-the-art algorithms. The authors provide a specific applications of the Katz-Bonacich centrality minimization problem based on the minimization of gossip propagation and make some experiments on real networks which validate the model assumptions.

Betweenness centrality is one of the basic concepts in the analysis of social networks. The initial definition for the betweenness of a node in a graph is based on the fraction of the number of geodesics (shortest paths) between any two nodes that this given node lies on, to the total number of the shortest paths connecting these nodes. This method has quadratic complexity and does not take into account indirect paths. In [45] K. Avrachenkov in collaboration with V. Mazalov (Korelian Institute of Applied Mathematical Research, Russia) and B. Tsynguev (Transbaikal State Univ., Russia) propose a new concept of betweenness centrality for weighted networks, called beta current flow centrality, based on Kirchhoff's law for electric circuits. In comparison with the original current flow centrality and alpha current flow centrality, this new measure can be computed for larger networks. The results of numerical experiments for some examples of networks, in particular, for the popular social network VKontakte as well as the comparison with PageRank method are presented.

PageRank has numerous applications in information retrieval, reputation systems, machine learning, and graph partitioning. In [44], K. Avrachenkov and A. Kadavankandy in collaboration with L.O. Prokhorenkova and A. Raigorodskii (both from Yandex Research) study PageRank in undirected random graphs with expansion property. The Chung-Lu random graph represents an example of such graphs. The authors show that in the limit, as the size of the graph goes to infinity, PageRank can be represented by a mixture of the restart distribution and the vertex degree distribution.

7.1.3. Mining social networks

Social Networks became a major actor in information propagation. Using the Twitter popular platform, mobile users post or relay messages from different locations. The tweet content, meaning and location show how an event-such as the bursty one "JeSuisCharlie" happened in France in January 2015 is comprehended in different countries. In [75], [76] researchers from UAPV and Inria (Mohamed Morchid, Yonathan Portilla, Didier Josselin, Richard Dufour, Eitan Altman, Marc El-Beze, Jean-Valère Cossu, Georges Linarès, Alexandre Reiffers-Masson), studied clustering of the tweets according to the co-occurrence of their terms, including the country, and forecasting the probable country of a non located tweet, knowing its content. First, they present the process of collecting a large quantity of data from the Twitter website. The dataset consists of 2.189 located tweets about "Charlie", from the 7th to the 14th of January. The authors then describe an original method adapted from the Author-Topic (AT) model based on the Latent Dirichlet Allocation method (LDA). They define a homogeneous space containing both lexical content (words) and spatial information (country). During a training process on a part of the sample, the authors provide a set of clusters (topics) based on statistical relations between lexical and spatial terms. During a clustering task, they evaluate the method effectiveness on the rest of the sample that reaches up to 95% of good assignments. It shows that the model is pertinent to foresee tweet location after a learning process.

7.1.4. Analysis of Internet Memes

Memes have been defined by R. Dawkins as cultural phenomena that propagate through non genetic ways. In [42], Eitan Altman and Yonathan Portilla examine three very popular Internet Memes and study their impact on the society in mediterranean countries. the authors use existing software tools (such as Google Trends) as well as tools that they develop in order to quantify the impact of the Memes on the mediterranean societies. The authors obtain quite different results with the different tools they use, which they explain based on some propagation characteristic of each one of the Memes. The analysis shows the extent to which these Memes cross borders and thus contribute to the creation of a globalized culture. The authors finally identify some of the impacts of the globalization of culture.

7.1.5. Trend detection in social networks using Hawkes processes

In [52], Julio Cesar Louzada Pinto and Tijani Chahed (Telecom SudParis) in collaboration with Eitan Altman propose a new trend detection algorithm, designed to find trendy topics being disseminated in a social network. The authors assume that the broadcasts of messages in the social network is governed by a self-exciting point process, namely a Hawkes process, which takes into consideration the real broadcasting times of messages and the interaction between users and topics. The authors formally define trendiness and derive trend indices for each topic being disseminated in the social network. These indices take into consideration the time between the detection and the message broadcasts, the distance between the real broadcast intensity and the maximum expected broadcast intensity, and the social network topology. The proposed trend detection algorithm is simple and uses stochastic control techniques in order to calculate the trend indices. It is also fast and aggregates all the information of the broadcasts into a simple one-dimensional process, thus reducing its complexity and the quantity of data necessary to the detection.

7.1.6. Study of the Youtube recommendation system

The Youtube recommendation system is one the most important view source of a video. In [54], Yonathan Portilla, Alexandre Reiffers-Masson, Eitan Altman in collaboration with Rachid El-Azouzi (UAPV) study the role of recommendation systems in boosting the popularity of videos. The authors first construct a graph that captures the recommendation system in Youtube and study empirically the relationship between the number of views of a video and the average number of views of the videos in its recommendation list. The authors then consider a random walker on the recommendation graph, i.e. a random user that browses through videos such that the video it chooses to watch is selected randomly among the videos in the recommendation list of the previous video it watched. The authors study the stability properties of this random process and show that the trajectory obtained does not contain cycles if the number of videos in the recommendation list is small (which is the case if the computer's screen is small).

7.1.7. Average consensus protocols

In [22] M. El Chamie (Univ. of Washington, USA), G. Neglia and K. Avrachenkov study the weight optimization problem for average consensus protocols by reformulating it as a Schatten norm minimization with parameter p. They show that as p approaches infinity, the optimal solution of the Schatten norm induced problem recovers the optimal solution of the original problem. Moreover, by tuning the parameter p in the proposed minimization, it is possible to trade-off the quality of the solution (i.e., the speed of convergence) for communication/computation requirements (in terms of number of messages exchanged and volume of data processed). They then propose a distributed algorithm to solve the Schatten norm minimization and show that it outperforms the other distributed weight selection methods.

7.1.8. Estimation techniques

The estimation of a large population's size by means of sampling procedures is a key issue in many networking scenarios. Their application domains span from RFID systems to peer-to-peer networks; from traffic analysis to wireless sensor networks; from multicast networks to WLANs. In [14], N. Accettura (Univ. of California Berkeley, USA), G. Neglia and L. A. Grieco (Politecnico di Bari, Italy) illustrate and classify in a coherent framework the main approaches proposed so far in the computer networks literature to deal with such a problem. In particular, starting from the methodologies proposed in ecological studies since the last century, they survey their counterparts in the computer network domain, finding that many lessons can be gained from this insightful investigation. Capture-Recapture techniques are deeply analyzed to allow the reader to exactly understand their pros, cons, and applicability bounds. Finally, they discuss some open issues that deserve further investigations and could be relevant to afford estimation problems in next generation Internet.

Online social networks (OSN) contain extensive amount of information about the underlying society that is yet to be explored. One of the most feasible technique to fetch information from OSN, crawling through Application Programming Interface (API) requests, poses serious concerns over the the guarantees of the estimates. In [70] J. Sreedharan and K. Avrachenkov in collaboration with B. Ribeiro (Purdue University, USA) focus on making reliable statistical inference with limited API crawls. Based on regenerative properties of the random walks, they propose an unbiased estimator for the aggregated sum of functions over edges and proved the connection between variance of the estimator and spectral gap. In order to facilitate Bayesian inference on the true value of the estimator, they derive the approximate posterior distribution of the estimate. Later the proposed ideas are validated with numerical experiments on inference problems in real-world networks.

7.1.9. Percolation in multilayer networks

In [79], P. Nain and his co-authors (S. Guha and P. Basu from Raytheon BB Technologies, D. Towsley from the Univ. of Massachusetts, C. Capar from Ericsson Research, A. Swami from the US Army Research Lab.) consider multiple networks formed by a common set of users connected via M different means of connectivity, where each user (node) is active, independently, in any given network with probability q. They show that when q exceeds a threshold $q_c(M)$, a giant connected component appears in the M-layer network—thereby enabling faraway users to connect using 'bridge' nodes that are active in multiple network layers, even though the individual layers may only have small disconnected islands of connectivity. They show that $q_c(M) \leq \sqrt{\log(1-p_c)}/\sqrt{M}$, where p_c is the bond percolation threshold of the underlying connectivity graph G, and $q_c(1) \equiv q_c$ is its site percolation threshold. The threshold $q_c(M)$ is found explicitly when G is a large random network with an arbitrary node-degree distribution and numerically for various regular lattices. Finally, an intriguingly close connection between this multilayer percolation model and the well-studied problem of site-bond percolation is revealed, in the sense that both models provide a smooth transition between the traditional site and bond percolation models. This connection is used to translate analytical approximations of the site-bond critical region developed in the 1990s, which are functions only of p_c and q_c of the respective lattice, to excellent general approximations of $q_c(M)$.

7.1.10. Extreme Value Theory for Complex Networks

In [20] J. Sreedharan and K. Avrachenkov in collaboration with N. Markovich (Institute of Control Sciences, Moscow) explore the dependence structure in the sampled sequence of complex networks. They consider randomized algorithms to sample the nodes and study extremal properties in any associated stationary sequence of characteristics of interest like node degrees, number of followers, or income of the nodes in online social networks, which satisfy two mixing conditions. Several useful extremes of the sampled sequence like the kth largest value, clusters of exceedances over a threshold, and first hitting time of a large value are investigated. The dependence and the statistics of extremes is abstracted into a single parameter that appears in extreme value theory, called the Extremal Index. The authors derive this parameter analytically and also estimate it empirically. They propose the use of the Extremal Index as a parameter to compare different sampling procedures. As a specific example, degree correlations between neighboring nodes are studied in detail with three prominent random walks as sampling techniques.

7.1.11. Random Matrix Theory for Complex Networks

In [68] A. Kadavankandy and K. Avrachenkov in collaboration with L. Cottatellucci (Eurecom) consider an extension of Erdős-Rényi graph known in the literature as the Stochastic Block Model (SBM). They analyze the limiting empirical distribution of the eigenvalues of the adjacency matrix of a SBM. They derive a fixed point equation for the Stieltjes transform of the limiting eigenvalue empirical distribution function (e.d.f.), concentration results on both the support of the limiting e.d.f. and the extremal eigenvalues outside the support of the limiting e.d.f. Additionally, they derive analogous results for the normalized Laplacian matrix and discuss potential applications of the general results in epidemics and random walks.

In [40], the same authors continue with the analysis of eigenvectors of a Stochastic Block Model. The eigenvalue spectrum of the adjacency matrix of a SBM consists of two parts: a finite discrete set of dominant eigenvalues and a continuous bulk of eigenvalues. They characterize analytically the eigenvectors corresponding to the continuous part: the bulk eigenvectors. For symmetric SBM adjacency matrices, the eigenvectors are shown to satisfy two key properties. A modified spectral function of the eigenvalues, depending on the eigenvectors, converges to the eigenvalue spectrum. Its fluctuations around this limit converge to a Gaussian process different from a Brownian bridge. This latter fact disproves that the bulk eigenvectors are Haar distributed.

7.2. Wireless Networks

Participants: Eitan Altman, Abdulhalim Dandoush.

7.2.1. A General SDN-based IoT Framework with NVF Implementation

The emerging technologies of IoT (Internet of Things), SDN (Software Defined Networking), and NFV (Network Function Virtualization) have a great potential for the information service innovation in the Cloud and big data era. In [26], Jie Li (Tsukuba Univ.) in cooperation with Eitan Altman and with Corinne Touati (Inria Grenoble-Rhône-Alpes) have studied architecture issues in Internet of Things based on SDN with NFV implementation. The contribution of the paper is in providing a view point for integrating these technologies based on their existing standards.

7.2.2. Self-Organizing Network (SON)

Self-Organizing Network (SON) technology aims at autonomously deploying, optimizing and repairing the Radio Access Networks. In [31], Abdoulaye Tall, Zwi Altman (Orange, Issy les Moulineaux) and Eitan Altman showed that in certain cases, it is essential to take into account the impact of the backhaul state in the design of the SON algorithm. They revisit the Base Station load definition taking into account the backhaul state. They provide an analytical formula for the load along with a simple estimator for both elastic and guaranteed bit-rate traffic. They incorporate the proposed load estimator in a self-optimized Load Balancing algorithm. Simulation results for a backhaul constrained heterogeneous network illustrate how the correct load definition can guarantee a proper operation of the SON algorithm.

SON is further studied by these authors in [58], [59] where the Vertical Sectorization (VS) is adapted. VS consists in creating vertically separated sectors in the original cell using an Active Antenna Systems (AAS) supporting two distinct beams with different downtilts. The total transmit power is split between the two sectors, while the frequency bandwidth can be reused by each sector, creating additional interference between the two sectors. For low traffic demand, VS may lead to performance degradation, while for high traffic demand in both sectors, VS is likely to bring about important capacity gains. Hence intelligent activation policy of VS is needed to fully benefit from this feature. The authors propose an approach taking advantage of the more focused downtilted beam. A dynamic alpha fair bandwidth sharing is proposed for low and medium load. It is autonomously replaced by full bandwidth reuse for high load scenarios using a threshold-based controller. A flow-level dynamic simulator is used to numerically validate the proposed mechanisms.

7.2.3. Automated Dynamic Offset for Network Selection in Heterogeneous Networks

Complementing traditional cellular networks with the option of integrated small cells and WiFi access points can be used to further boost the overall traffic capacity and service level. Small cells along with WiFi access points are projected to carry over 60% of all the global data traffic by 2015. With the integration of small cells on the radio access network levels, there is a focus on providing operators with more control over small cell selection while reducing the feedback burden. Altogether, these issues motivate the need for innovative distributed and autonomous association policies that operate on each user under the network operator's control, utilizing only partial information, yet achieving near-optimal solutions for the network. In [25], Majed Haddad (UAPV), Piotr Wiecek (Institute of Mathematics and Computer Science, Wroclaw), Saidi Habib (Inria project-team DYOGENE) and Eitan Altman propose a load-aware network selection approach applied to automated dynamic offset in heterogeneous networks. In particular, they investigate the properties of a hierarchical (Stackelberg) Bayesian game framework, in which the macro cell dynamically chooses the offset about the state of the channel in order to guide users to perform intelligent network selection decisions between macro cell and small cell networks. The authors effectively address the problem of how to intelligently configure a dynamic offset which optimizes network's global utility while users maximize their individual utilities.

7.2.4. Localization in ad-hoc wireless sensors networks

Range-based localization algorithms in wireless sensor networks are more accurate but also more computationally complex than the range-free algorithms. The work on this topic by M. S. Elgamel (Arab Academy for Science, Technology & Maritime Transport, Egypt) and A. Dandoush, previously reported, has been published in [23].

7.3. Network Engineering Games

Participants: Eitan Altman, Konstantin Avrachenkov, Giovanni Neglia.

7.3.1. Matching games and the association problem

In [33], Mikaël Touati, Jean-Marc Kélif (Orange Labs), Rachid El-Azouzi (UAPV), Marceau Coupechoux (Telecom ParisTech) and Eitan Altman propose two new algorithms for finding stable structures in ordinal coalition potential games. The first one is enumerative and it performs on a graph. The second one is a modified Deferred Acceptance Algorithm using counter-proposals. It finds a many-to-one matching. The authors illustrate with the example of video caching from a content creator's servers to a service provider's servers.

This is applied to the association of mobiles to IEEE 802.11-based WLANs in populated areas where many mobile terminals are covered by several Access Points (APs) [32]. These mobiles have the possibility to associate to the AP with the strongest signal (best-RSSI association scheme). This can lead to poor performances and overloaded APs. Moreover, the well-known anomaly in the protocol at the MAC layer may also lead to very unpredictable performances and affect the system throughput due to the presence of heterogeneous data rate nodes and the shared nature of the 802.11 medium. In [61], the same authors solve the joint resource allocation and mobile user association after modeling it as a matching game with complementarities, peer effects and selfish players.

7.3.2. Normalized Nash Equilibria for power control with correlated constraints

When correlated constraints are introduced to a game (i.e. the set of actions of a player depends on the policies of other players) there may exist infinitely many Nash equilibria. Assume one wishes to select a particular one u. According to the Karush Kuhn Tucker theorem, there exist Lagrange mutipliers such that the best response when all players use their equilibrium policy is the same as that obtained by optimizing the corresponding Lagrangian of that player. The Lagrange multipliers can be interpreted as marginal costs such that if they are imposed on the player as some tax to pay then this induces the player to use Nash equilibrium. The following question arises: does there exist an equilibrium u for which the corresponding Lagrange multipliers are player independent. If the answer is positive then this would make in many cases the billing scalable and simple to implement. An equilibrium u for which the corresponding Lagrange multipliers are player independent is called a normalized Nash equilibrium (NNE). In [39], [50] and [24], Arnob Ghosh (Univ. of Pennsylvania), Laura Cottatellucci (Eurecom) and Eitan Altman provide new conditions for existence and uniqueness of NNE and apply this for power control games arising in cognitive radio [24] and in heterogeneous networks [39], [50].

7.3.3. Admission control to an infinite server queue

In [36], Eitan Altman studies in collaboration with Piotr Wiecek (Wrocław Univ. of Technology) and Arnob Ghosh (Univ. of Pennsylvania) a mean field approximation of the M/M/∞ queueing system. The problem they consider is quite different from standard games of congestion as they consider the case in which higher congestion results in smaller costs per user. This is motivated by a situation in which some TV show is broadcast so that the same cost is needed no matter how many users follow the show. Using a mean-field approximation, they show that this results in multiple equilibria of threshold type which is explicitly computed. The authors further derive the social optimal policy and compute the price of anarchy, and show that the mean-field approximation becomes tight as the workload increases, thus the results obtained for the mean-field model well approximate the discrete one.

7.3.4. Posting Time of Content over a Temporally-Ordered Shared Medium

In [17], Eitan Altman in collaboration with Nahum Shimkin (Technion) consider a game of timing between a random number of content creators, who compete for position and exposure time over a shared medium such as an on-line classified list. Contents (such as ads, messages, multimedia items or comments) are ordered according to their submission times, with more recent submissions displayed at the top (and better) positions. The instantaneous effectiveness of each ad depends on its current display position, as well as on a time-dependent exposure function common to all. Each content creator may choose the submission time of her content within a finite time interval, with the goal of maximizing the total exposure of this content. The authors formulate the problem as a non-cooperative game, analyze its symmetric equilibrium, characterize it in terms of a differential boundary value problem and devise a numerical scheme for its computation.

7.3.5. Routing Games

A central question in routing games has been to establish conditions for the uniqueness of the equilibrium, either in terms of network topology or in terms of costs. This question is well understood in two classes of routing games. The first is the non-atomic routing introduced by Wardrop on 1952 in the context of road traffic in which each player (car) is infinitesimally small; a single car has a negligible impact on the congestion. Each car wishes to minimize its expected delay. Under arbitrary topology, such games are known to have a convex potential and thus a unique equilibrium. The second framework is splitable atomic games: there are finitely many players, each controlling the route of a population of individuals (let them be cars in road traffic or packets in the communication networks). In [64], Eitan Altman and Corinne Touati (Inria Grenoble-Rhône-Alpes) study two other frameworks of routing games in which each of several players has an integer number of connections (which are population of packets) to route and where there is a constraint that a connection cannot be split. Through a particular game with a simple three link topology, they identify various novel and surprising properties of games within these frameworks. The authors show in particular that equilibria are non unique even in the potential game setting of Rosenthal with strictly convex link costs. They further show that non-symmetric equilibria arise in symmetric networks.

7.3.6. Resilience of Routing in Parallel Link Networks

Aniruddha Singhal, Corinne Touati (both from Inria Grenoble-Rhône-Alpes) in collaboration with Eitan Altman and Jie Li (Univ. of Tsukuba) revisit in [63], the resilience problem of routing traffic in a parallel link network model with a malicious player using a game theoretic framework. Consider that there are two players in the network: the first player wishes to split its traffic so as to minimize its average delay, which the second player, i.e., the malicious player, tries to maximize. The first player has a demand constraint on the total traffic it routes. The second player controls the link capacities: it can decrease by some amount the capacity of each link under a constraint on the sum of capacity degradation. The authors first show that the average delay function is convex both in traffic and in capacity degradation over the parallel links and thus does not have a saddle point. They identify best responses strategies of each player and compute both the max-min and the min-max values of the game. One is especially interested in the min-max strategy as it guarantees the best performance under worst possible link capacity degradation. It thus allows to obtain routing strategies that are resilient and robust. The authors compare the results of the min-max to those obtained under the max-min strategies. They provide stable algorithms for computing both max-min and min-max strategies as well as for best responses.

7.3.7. The Social Medium Selection Game

In [72], Fabrice Lebeau (ENS Lyon) Corinne Touati and Nof Abuzainab (Inria Grenoble-Rhône-Alpes) in collaboration with Eitan Altman, consider competition of content creators in routing their content through various media. The routing decisions may correspond to the selection of a social network (e.g. twitter versus facebook or linkedin) or of a group within a given social network. The utility for a player to send its content to some medium is given as the difference between the dissemination utility at this medium and some transmission cost. The authors model this game as a congestion game and compute the pure potential of the game. In contrast to the continuous case, they show that there may be various equilibria. The authors show that the potential is M-concave which allows them to characterize the equilibria and to propose an algorithm for computing it. They then give a learning mechanism which leads to an efficient algorithm to determine an equilibrium. The authors finally determine the asymptotic form of the equilibrium and discuss the implications on the social medium selection problem.

7.3.8. Activation Games in Online Dating Platforms

In [41], Eitan Altman in collaboration with Francesco De Pellegrini (CREATE-NET, Trento) and Huijuan Wang (Delft Univ. of Technology) describe a model for the activation level of users in online dating platforms (ODPs). Such popular systems are conceived in order to match individuals from two groups of potential mates. The business of such platforms pivots around the customers' expectancy to get in contact with their future dates: upon the payment of a fee to the platform owner, ODPs provide specific tools to improve reach and visibility. However, ODPs require a critical number of active users in order to sustain their operations (and their business). Customers of the platform trade off on the price for being more visible and attract mates' contacts. A user becomes inactive if he or she is not contacted by others for some time: being contacted by potential mates acts as an activation signal. The aim of the analysis is to propose a game theoretical framework to capture such a complex activation problem in strategic form. The authors unveil the structure of Nash equilibria and further derive a Stackelberg formulation. The latter is a hierarchical game where the platform owner aims at maximizing profits while preserving the ODP activity level above a critical epidemic threshold.

7.3.9. Epidemics in Networks

Stojan Trajanovski, Huijuan Wang, Piet Van Mieghem (all from Delft Univ. of Technology), in collaboration with Yezekael Hayel (UAPV) and Eitan Altman have pursued their work in the Congas European project concerning malware attacks modeled as SIS (for Susceptible-Infected-Susceptible) epidemics in networks. In [34], the authors consider decentralized optimal protection strategies when a virus is propagating over a network. they assume that each node in the network can fully protect itself from infection at a constant cost, or the node can use recovery software, once it is infected. They model the system using a game theoretic framework and find pure, mixed equilibria, and the Price of Anarchy (PoA) in several network topologies.

Further, they propose both a decentralized algorithm and an iterative procedure to compute a pure equilibrium in the general case of a multiple communities network. Finally, the authors evaluate the algorithms and give numerical illustrations of all results.

They then considered the game-formation problem while balancing multiple, possibly conflicting objectives like cost, high performance, security and resiliency to viruses. In [60], Stojan Trajanovski, Fernando Antonio Kuiper and Piet Van Mieghem (all from Delft Univ. of Technology) in collaboration with Yezekael Hayel (UAPV) and Eitan Altman use a game-formation approach to network design where each player (node), aims to collectively minimize the cost of installing links, of protecting against viruses, and of assuring connectivity. In the game, minimizing virus risk as well as connectivity costs results in sparse graphs. They show that the Nash Equilibria are trees that, according to the Price of Anarchy (PoA), are close to the global optimum, while the worst-case Nash Equilibrium and the global optimum may significantly differ for small infection rate and link installation cost. Moreover, the types of trees, in both the Nash Equilibria and the optimal solution, depend on the virus infection rate, which provides new insights into how viruses spread: for a high infection rate, the path graph is the worst- and the star graph is the best-case Nash Equilibrium. However, for small and intermediate infection rates, trees different from the path and star graphs may be optimal.

7.3.10. Retrial games

In [46] K. Avrachenkov in collaboration with E. Morozov and R. Nekrasova (both from Petrozavodsk State Univ., Russia) consider a single-server retrial system with one and several classes of customers. In the case of several classes, each class has its own orbit for retrying customers. The retrials from the orbits are generated with constant retrial rates. In the single class case, the objective is finding an optimal retrial rate. Whereas in the multi-class case, a game theoretic framework is used and equilibrium retrial rates are found. The performance criteria balance the number of retrials per retrying customer with the number of unhappy customers.

7.3.11. Cooperative Network Design

The Network Design problem has received increasing attention in recent years. Previous works have addressed this problem considering almost exclusively networks designed by selfish users, which can be consistently suboptimal. In [18] K. Avrachenkov, J. Elias (Univ. Paris Descartes, France), F. Martignon (Univ. Paris Sud, France), G. Neglia and L. Petrosyan (St. Petersburg State Univ.) address the network design issue using cooperative game theory, which permits to study ways to enforce and sustain cooperation among users. Both the Nash bargaining solution and the Shapley value are widely applicable concepts for solving these games. However, the Shapley value presents several drawbacks in this context. For this reason, they solve the cooperative network design game using the Nash bargaining solution (NBS) concept. More specifically, they extend the NBS approach to the case of multiple players and give an explicit expression for users' cost allocations. They further provide a distributed algorithm for computing the Nash bargaining solution. Then, they compare the NBS to the Shapley value and the Nash equilibrium solution in several network scenarios, including real ISP topologies, showing its advantages and appealing properties in terms of cost allocation to users and computation time to obtain the solution.

Numerical results demonstrate that the proposed Nash bargaining solution approach permits to allocate costs fairly to users in a reasonable computation time, thus representing a very effective framework for the design of efficient and stable networks.

7.4. Green Networking and Smart Grids

Participants: Sara Alouf, Eitan Altman, Alberto Benegiamo, Alain Jean-Marie, Giovanni Neglia.

7.4.1. Energy efficiency and management in wireless networks

In [35], Rodrigo A. Vaca Ramirez and John S. Thompson (Univ. of New England), in collaboration with Eitan altman and Victor Ramos Ramos (UAM - Univ. Autonoma Metropolitana Unidad Iztapalapa) consider a low complexity virtual Multiple-input Multiple-output (MIMO) coalition formation algorithm. The goal is to obtain improvements in energy efficiency by forming multi-antenna virtual arrays for information

transmission in the uplink. Virtual arrays are formed by finding a stable match between single antenna devices such as mobile station (MS) and relay stations (RS) by using a game theoretic approach derived from the concept of the college admissions problem. They focus on enhancing the MS performance by forming virtual coalitions with the RSs. Thus, power savings are obtained through multi-antenna arrays by implementing the concepts of spatial diversity and spatial multiplexing for uplink transmission. They focus on optimizing the overall consumed power rather than the transmitted power of the network devices. Furthermore, it is shown analytically and by simulations that when overall consumed power is considered as an optimization metric, the energy efficiency of the single antennas devices is not always improved by forming a virtual MIMO array. Hence, single antenna devices may prefer to transmit on their own when channel conditions are favorable. In addition, the simulation results show that the proposed framework provides comparable energy savings and a lower implementation complexity when compared to a centralized exhaustive search approach that is coordinated from the Base Station.

Sara Alouf, Ioannis Dimitriou (now at Univ. Patras, Greece) and Alain Jean-Marie had worked on the modeling of wireless communication base stations with autonomous energy supply (solar, wind). They had proposed a versatile 5-dimensional Markov model of the device, and shown that the Quasi Birth-Death framework is adequate for solving the model. This work has been completed with a companion product-form model based on E. Gelenbe's modeling of energy networks with signals [48].

7.4.2. Stochastic Geometric Models for Green Networking

In [16], Eitan Altman in collaboration with Cengis Hasan, Manjesh Kumar Hanawal (IIT Mumbai), Shlomo Shamai (Technion), Jean-Marie Gorce (Inria project-team SOCRATE), Rachid El-Azouzi (UAPV) and Laurent Roullet (Alcatel Lucent Bell Labs) study both the uplink and downlink energy efficiency based on the assumption that base stations are distributed according to an independent stationary Poisson point process. This type of modeling allows to make use of the property that the spatial distribution of the base stations after thinning (switching-off) is still a Poisson process. This implies that the probability of the SINR can be kept unchanged when switching-off base stations provided that one scales up the transmission power of the remaining base stations. The authors then solve the problem of optimally selecting the switch-off probabilities so as to minimize the energy consumptions while keeping unchanged the SINR probability distribution. They then study the trade-off in the uplink performance involved in switching-off base stations. These include energy consumption, the coverage and capacity, and the impact on amount of radiation absorbed by the transmitting user.

7.4.3. Direct Load Control

Energy demand and production need to be constantly matched in the power grid. The traditional paradigm to continuously adapt the production to the demand is challenged by the increasing penetration of more variable and less predictable energy sources, like solar photovoltaics and wind power. An alternative approach is the so called direct control of some inherently flexible electric loads to shape the demand. Direct control of deferrable loads presents analogies with flow admission control in telecommunication networks: a request for network resources (bandwidth or energy) can be delayed on the basis of the current network status in order to guarantee some performance metrics. In [53] G. Neglia, in collaboration with G. Di Bella (Telecom Italia, Italy), L. Giarré and I. Tinnirello (Univ. of Palermo, Italy) go beyond such an analogy, showing that usual teletraffic tools can be effectively used to control energy loads. In particular they propose a family of control schemes which can be easily tuned to achieve the desired trade-off among resources usage, control overhead and privacy leakage.

7.4.4. Charge of Electric Vehicles

The massive introduction of Electric Vehicles (EVs) will make fleet managers spend a significant amount of money to buy electric energy. If energy price changes over time, accurate scheduling of recharging times may result in significant savings. In [29] C. Rottondi (IDSIA Dalle Molle Institute for Artificial Intelligence, Switzerland), G. Neglia and G. Verticale (Politecnico di Milano, Italy) evaluate the complexity of the optimal scheduling problem considering a scenario with a fleet manager having full knowledge of the customers'

traveling needs at the beginning of the scheduling horizon. They prove that the problem has polynomial complexity and provide complexity lower and upper bounds. Moreover, they propose an online sub-optimal scheduling heuristic that schedules the EVs' recharge based on historical travelling data. They compare the performance of the optimal and sub-optimal methods to a benchmark online approach that does not rely on any prior knowledge of the customers' requests, in order to evaluate whether the additional complexity required by the proposed strategies is worth the achieved economic advantages. Numerical results show up to of 35% cost savings with respect to the benchmark approach.

7.5. Content-Oriented Systems

Participants: Sara Alouf, Eitan Altman, Konstantin Avrachenkov, Alain Jean-Marie, Philippe Nain, Giovanni Neglia.

7.5.1. Modeling modern DNS caches

Sara Alouf and Nicaise Choungmo Fofack (former PhD student at MAESTRO, currently at Ingima) have thoroughly revised their study of the modern behavior of DNS caches. In particular the closure properties of the class of distributions called *diagonal matrix-exponential* are fully derived, hence the analytic models presented in [78] to study tree of caches with general caching durations are extended to the case of polytrees [15].

7.5.2. Data placement and retrieval in distributed/peer-to-peer systems

In previous years, Alain Jean-Marie and collaborators from the Univ. Montpellier have defined a family of combinatorial designs that minimize the variance in the availability of replicated documents in unreliable infrastructures. Then with Jean-Claude Bermond (CNRS, with the Inria project-team COATI), Dorian Mazauric (now with Inria project-team ABS) and Joseph Yu (UFV Vancouver), it was shown that *well-balanced families* solve the problem, and such families were constructed for small numbers of replicas. This work is now published in [21]. During the internship of Mikhail Grigorev, several methods for generating at random good solutions have been investigated.

7.5.3. Fairness in caching systems

Data offloading from the cellular network to lowcost WiFi has been the subject of several research works in the last years. In-network caching has also been studied as an efficient means to further reduce cellular network traffic. In [49] M. El Chamie (Univ. of Washington, USA), C. Barakat (Inria project-team DIANA) and G. Neglia consider a scenario where mobile users can download popular contents (e.g., maps of a city, shopping information, social media, etc.) from WiFi-enabled caches deployed in an urban area. They study the optimal distribution of contents among the caches (i.e., what contents to put in each cache) to minimize users' access cost in the whole network, and argue that this optimal distribution does not necessarily provide geographic fairness, i.e., users at different locations can experience highly variable performance. In order to mitigate this problem, they propose two different cache coordination algorithms based on gossiping. These algorithms achieve geographic fairness while preserving the minimum access cost for end users.

In [43] K. Avrachenkov in collaboration with V.S. Borkar (IIT Mumbai, 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 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 Whittle index is derived, together with its theoretical justification.

7.6. Advances in Methodological Tools

Participants: Eitan Altman, Konstantin Avrachenkov, Ilaria Brunetti.

7.6.1. Control theory

In [19] K. Avrachenkov in collaboration with O. Habachi (UAPV) and A. Piunovskiy and Y. Zhang (both from Univ. of Liverpool, UK) investigate infinite horizon deterministic optimal control problems with both gradual and impulsive controls, where any finitely many impulses are allowed simultaneously. Both discounted and long run time average criteria are considered. They establish very general and at the same time natural conditions, under which the dynamic programming approach results in an optimal feedback policy. The established theoretical results are applied to the Internet congestion control, and by solving analytically and nontrivially the underlying optimal control problems, they obtain a simple threshold-based active queue management scheme, which takes into account the main parameters of the transmission control protocols, and improves the fairness among the connections in a given network.

7.6.2. Game theory

7.6.2.1. Evolutionary games

Standard Evolutionary Game framework is a useful tool to study large interacting systems and to understand the strategic behavior of individuals in such complex systems. Adding an individual state to model a local feature of each player in this context, allows one to study a wider range of problems in various application areas as networking, biology, etc. In [47], Ilaria Brunetti and Eitan Altman in collaboration with Yezekael Hayel (UAPV) introduce such an extension of evolutionary game framework and particularly, focus on the dynamical aspects of this system. Precisely, the authors study the coupled dynamics of the strategies and the individual states inside a population of interacting individuals. They consider here a two-strategies evolutionary game. They first obtain a system of combined dynamics and they show that the rest-points of this system are equilibria of the evolutionary game with individual state. Second, by assuming two different time scales between states and strategy dynamics, one can compute explicitly the equilibria. Then, by transforming the evolutionary game with individual states into a standard evolutionary game, the authors obtains an equilibrium which is equivalent, in terms of occupation measure, to the previous one. Finally, they show a generalization of the model. All the results are illustrated with numerical results.

7.6.2.2. Stochastic Games

Motivated by uncertainty in the value of the interest rate, in [62] K. Avrachenkov in collaboration with A. Varava (KTH, Sweden) study discounted zero-sum stochastic games with and arbitrary discount factor. Their general goal is to obtain a power series expansion of the value of the game with respect to the discount factor around its nominal value. They consider a specific but important class of stochastic games – completely mixed stochastic games. As an illustrative example they take a tax evasion model.

MUSE Team

6. New Results

6.1. Home Network or Access Link? Locating Last-mile Downstream Throughput Bottlenecks

Participants: Srikanth Sundaresan (ICSI), Nick Feamster (Princeton), Renata Teixeira

As home networks see increasingly faster downstream throughput speeds, a natural question is whether users are benefiting from these faster speeds or simply facing performance bottlenecks in their own home networks. In our paper recently accepted for publication in PAM'16, we studied whether downstream throughput bottlenecks occur more frequently in their home networks or in their access ISPs. We identified lightweight metrics that can accurately identify whether a throughput bottleneck lies inside or outside a user's home network and developed a detection algorithm that locates these bottlenecks. We validated this algorithm in controlled settings and characterized bottlenecks on two deployments, one of which included 2,652 homes across the United States. We found that wireless bottlenecks are more common than access-link bottlenecks—particularly for home networks with downstream throughput greater than 20 Mbps, where access-link bottlenecks are relatively rare.

6.2. On the Reliability of Profile Matching Across Large Online Social Networks

Participants: Oana Goga and Krishna Gummadi (MPI-SWS), Patrick Loiseau (EURECOM), Robin Sommer (ICSI), Renata Teixeira

Matching the profiles of a user across multiple online social networks brings opportunities for new services and applications as well as new insights on user online behavior, yet it raises serious privacy concerns. Prior literature has showed that it is possible to accurately match profiles, but their evaluation focused only on sampled datasets. In our KDD'15 paper [2], we study the extent to which we can reliably match profiles in practice, across real-world social networks, by exploiting public attributes, i.e., information users publicly provide about themselves. Today's social networks have hundreds of millions of users, which brings completely new challenges as a reliable matching scheme must identify the correct matching profile out of the millions of possible profiles. We first define a set of properties for profile attributes–Availability, Consistency, non-Impersonability, and Discriminability (ACID)–that are both necessary and sufficient to determine the reliability of a matching scheme. Using these properties, we propose a method to evaluate the accuracy of matching schemes in real practical cases. Our results show that the accuracy in practice is significantly lower than the one reported in prior literature. When considering entire social networks, there is a non-negligible number of profiles that belong to different users but have similar attributes, which leads to many false matches. Our paper sheds light on the limits of matching profiles in the real world and illustrates the correct methodology to evaluate matching schemes in realistic scenarios.

6.3. Exploiting crowd sourced reviews to explain movie recommendation

Participants: Sara El Aouad, Christophe Dupuy, Francis Bach, and Renata Teixeira (Inria), Christophe Diot (Technicolor)

Streaming services such as Netflix, M-Go, and Hulu use advanced recommender systems to help their customers identify relevant content quickly and easily. These recommenders display the list of recommended movies organized in sublists labeled with the genre or some more specific labels. Unfortunately, existing methods to extract these labeled sublists require human annotators to manually label movies, which is time-consuming and biased by the views of annotators. In our work [6], we design a method that relies on crowd sourced reviews to automatically identify groups of similar movies and label these groups. Our method takes the content of movie reviews available online as input for an algorithm based on Latent Dirichlet Allocation (LDA) that identifies groups of similar movies. We separate the set of similar movies that share the same combination of genre in sublists and personalize the movies to show in each sublist using matrix factorization. The results of a side-by-side comparison of our method against Technicolor's M-Go VoD service are encouraging.

6.4. Characterizing Home Device Usage From Wireless Traffic Time Series

Participants: Katsiaryna Mirylenka (IBM Research - Zurich), Vassilis Christophides, Themis Palpanas (Paris Descartes University), Ioannis Pefkianakis (Hewlett Packard Labs), Martin May (Technicolor).

The analysis of *temporal behavioral patterns* of home network users can reveal important information to Internet Service Providers (ISPs) and help them to optimize their networks and offer new services (e.g., remote software upgrades, troubleshooting, energy savings). Our study [4] uses time series analysis of continuous traffic data from wireless home networks, to extract traffic patterns recurring within, or across homes, and *assess the impact of different device types (fixed or portable) on home traffic*. Traditional techniques for time series analysis are not suited in this respect, due to the limited stationary and evolving distribution properties of wireless home traffic data. We propose a novel framework that relies on a *correlation-based similarity* measure of time series, as well as a notion of *strong stationarity* to define recurring motifs and dominant devices. Using this framework, we analyze the wireless traffic collected from 196 home gateways over two months. Our framework goes beyond existing application-specific analysis techniques, such as analysis of wireless traffic, which mainly rely on data aggregated across hundreds, or thousands of users. It enables the extraction of recurring patterns from traffic time series of individual homes, leading to a much more fine-grained analysis of the behavior patterns of the users. We also determine the best time aggregation policy w.r.t. to the number and statistical importance of the extracted motifs, as well as the device types dominating these motifs and the overall gateway traffic. Our results show that ISPs can exceed the simple observation of the aggregated gateway traffic and better understand their networks.

6.5. On Continuous Top-k Queries with Real-Time Scoring Functions

Participants: Nelly Vouzoukidou (Google, France), Bernd Amann (LIP6), Vassilis Christophides.

Modern news sharing and social media platforms allow millions of users to *produce and consume information in real-time*. To assess relevance of published information in this new setting, batch scoring based on content similarity, link centrality or page views is no longer sufficient. Instead, streams of events like "replies" (for posting comments), "likes" (for rating content) or "retweets" (for diffusing information) explicitly provided by users represent valuable online feedback on published information that has to be exploited in order to adjust in real-time any available score of information items. Note that in the future Internet of Things (IoT), not only digital, but also physical objects will be expected to be ranked in a fully automated way with respect to real-time human activities (viewing concentration), vital signals (emotional arousal), etc.

Rather than indexing as quickly as possible information items to re-evaluate *snapshot queries*, publish/subscribe systems index *continuous queries* and update on the fly their results each time a new matching item arrives. Existing publish/subscribe systems rely on two alternative continuous filetring semantics, namely *predicate-based* filtering or *similarity-based top-k* filtering. In predicate-based systems, incoming items that match the filtering predicates are simply added to the result list of continuous queries, while in similarity-based top-k publish/subscribe systems, matching items have also to exhibit better relevance w.r.t. the items already appearing as the top-k results of the continuous query. In top-k publish/subscribe systems the relevance of an item remains constant during a pre-specified time window, and once its lifetime exceeds the

item simply expires. Only recently, information recency has become part of the relevance score of continuous queries. Clearly, when information relevance decays as time passes both (a) results lists maintenance and (b) early pruning of the query index traversal are challenged. While these problems have been studied for (textual or spatio-textual) content scoring functions with time decay, non-homogeneous scoring functions accommodating various forms of *query-dependent* and *query-independent* information relevance with time decay is supported only by MeowsReader. In this work we are going beyond this general form of *time-decayed static scores* and consider continuous queries featuring *real-time scoring functions* under the form of *time decaying positive user feedback* for millions of online social media events per minute and millions of user queries.

RAP Project-Team

4. New Results

4.1. Random Graphs

Participant: Nicolas Broutin.

And/Or trees for random Boolean functions

For some time, a number of teams have tried to devise natural probability distributions on Boolean functions. Indeed, the most natural one, the uniform one, is not quite satisfactory: almost all Boolean functions have maximal complexity, while it is extremely difficult to construct some with high complexity. One approach consists in generating functions by seeing them as "expressions" encoded as a tree of computation. We generalize and unify the previous approaches that are restricted to very specific cases by looking at the distributions induced on the Boolean function by large computation trees that are arbitrary, except for the fact they the neighborhoods of the root (where the computation concentrates) stabilizes in distribution as the sizes of the tree increases [12].

4.2. Resource Allocation in Large Data Centres

Participants: Christine Fricker, Philippe Robert, Guilherme Thompson.

With the exponential increase in internet data transmission volume over the past years, efficient bandwidth allocation in large data centres has become crucial. Illustrating examples are the rapid spread of cloud computing technology, as well as the growth of the demand for video streaming, both of which were quasi non-existent 10 years ago.

Currently, most systems operate under decentralised policies due to the complexity of managing data exchange on large scales. In such systems, customer demands are served respecting their initial service requirements (a certain video quality, amount of memory or processing power etc.) until the system reaches saturation, which then leads to the blockage of subsequent customer demands. Strategies that rely on the scheduling of tasks are often not suitable to address this load balancing problem as the users expect instantaneous service usage in real time applications, such as video transmission and elastic computation. Our research goal is to understand and redesign its algorithms in order to develop decentralised policies that can improve global performance using local instantaneous information. This research is made in collaboration with Fabrice Guillemin, from Orange Labs.

In a first approach to this problem, we examined offloading schemes in fog computing context, where one data centres are installed at the edge of the network. We analyse the case with one data centre close to user which is backed up by a central (bigger) data centre. When a request arrives at an overloaded data centre, it is forwarded to the other data centre with a given probability, in order to help coping with saturation and reducing the rejection of requests. In [16], we have been able to show that the performance of such a system can be expressed in terms of the invariant distribution of a random walk in the quarter plane. As a consequence we have been able to assess the behaviour and performance of these systems, proving the effectiveness of such an offloading arrangement.

In a second step, we investigated allocation schemes which consist in reducing the bandwidth of arriving requests to a minimal value when the system is close to saturation. We analysed the effectiveness of such a downgrading policy, which, if the system is correctly designed, will reduce the fraction of rejected transmissions. We developed a mathematical model which allows us to predict system behaviour under such a policy and calculate the ideal threshold (in the same scale as the resource) after which downgrading should be initiated, given system parameters. We proved the existence of a unique equilibrium point, around which we have been able to determine the probability of the system being above or under the threshold. We found that system blockage can be almost surely eliminated. This policy finds a natural application in the context of video streaming services and other real time applications, such as MPEG-DASH. A document is being written to further publication.

Finally, with those results, we now try to extend our research towards more complex systems, investigating the behaviour of multiple resource systems (such as a Cloud environment, where computational power is provided using unities of CPU and GB of RAM) and other offloading schemes, such as the compulsory forwarding of a request when it's blocked at the edge server, but keeping a trunk reservation to protect the service originally assigned to the big data centre.

4.3. Resource allocation in vehicle sharing systems

Participants: Christine Fricker, Plinio Santini Dester, Hanene Mohamed, Yousra Chabchoub.

This is a collaboration with Danielle Tibi, Université Denis Diderot.

Vehicle sharing systems are becoming an urban mode of transportation, and launched in many cities, as Velib' and Autolib' in Paris. One of the major issues is the avail ability of the resources: vehicles or free slots to return them. These systems became a hot topic in Operation Research and now the importance of stochasticity on the system behavior is commonly admitted. The problem is to understand the system behavior and how to manage these systems in order to provide both resources to users. Our stochastic model is the first one taking into account the finite number of spots at the stations.

With Danielle Tibi, we use limit local theorems to obtain the asymptotic stationary joint distributions of several station states when the system is large (both numbers of stations and bikes), in the case of finite capacities of the stations. This gives an asymptotic independence property for node states. This widely extends the existing results on heterogeneous bike-sharing systems.

Recently we investigate some network load balancing algorithms to improve the bike sharing system behavior. We focus on the choice of the least loaded station among two t o return the bike. A problem is the influence of the delay between the choice time (the beginning of the trip) and the time the station is joined (the end of the trip). However the main challenge is to deal with the choice between two neighboring stations. For that, a system of infinite queues is studied in light traffic. For a bike-shar ing homogeneous model, we restrict our study to a deterministic cooperation of two by two stations. It relies on new results for the classical system of two queues under the join-the-shortest-queue policy.

JC Decaux provides us data describing Velib' user trips. These data are useful to measure the system parameters, validate our models and test our algorithms. Indeed, we use these data to investigate load balancing algorithms such as two-choice policies.

4.4. Scaling Methods

Participants: Philippe Robert, Wen Sun.

4.4.1. Fluid Limits in Wireless Networks

This is a collaboration with Amandine Veber (CMAP, École Polytechnique). The goal is to investigate the stability properties of wireless networks when the bandwidth allocated to a node is proportional to a function of its backlog: if a node of this network has x requests to transmit, then it receives a fraction of the capacity proportional to $\log(1+x)$, the logarithm of its current load. This year we completed the analysis of a star network topology with multiple nodes. Several scalings were used to describe the fluid limit behaviour.

4.4.2. The Time Scales of a Transient Network

A large distributed system where users' files are duplicated on unreliable data servers is investigated. Due to a server breakdown, a copy of a file can be lost, it can be retrieved if another copy of the same file is stored on other servers. In the case where no other copy of a given file is present in the network, it is definitely lost. In order to have multiple copies of a given file, it is assumed that each server can devote a fraction of its processing capacity to duplicate files on other servers to enhance the durability of the system.

A trade-off is necessary between the bandwidth and the memory used for this back-up mechanism and the data loss rate. Back-up mechanisms already exist and have been studied thanks to simulation. To our knowledge, no theoretical study exists on this topic. With a very simple centralized model, we have been able to emphasise a trade-off between capacity and life-time with respect to the duplication rate. From a mathematical point of view, we are currently studying different time scales of the system with an averaging phenomenon.

We have used scaling methods with different time scales to derive some asymptotic results on the decay of a simplified network: it is assumed that any copy of a given file is lost at some fixed rate and the total processing capacity of the system is devoted to duplicate the file with least number of copies. We start from the optimal initial state: each file has the maximum number of copies. Due to random losses, the state of the network is transient and all files will be eventually lost. There is a stability assumption for the system having a critical time scale of decay. When the stability condition is not satisfied, i.e. when it is initially overloaded, we have shown that the state of the network converges to an interesting local equilibrium. We are currently studying a more general case which the duplication depends on the structure of the system. See [7].

4.5. Stochastic Models of Biological Networks

Participants: Renaud Dessalles, Sarah Eugene, Philippe Robert.

4.5.1. Stochastic Modelling of self-regulation in the protein production system of bacteria

This is a collaboration with Vincent Fromion from INRA Jouy-en-Josas, which started on December 2014.

In prokaryotic cells (e.g. E. Coli. or B. Subtilis) the protein production system has to produce in a cell cycle (i.e. less than one hour) more than 10^6 molecules of more than 2500 kinds, each having different level of expression. The bacteria uses more than 85% of its resources to the protein production. Gene expression is a highly stochastic process: bacteria sharing the same genome, in a same environment will not produce exactly the same amount of a given protein. Some of this stochasticity can be due to the system of production itself: molecules, that take part in the production process, move freely into the cytoplasm and therefore reach any target in the cell after some random time; some of them are present in so much limited amount that none of them can be available for a certain time; the gene can be deactivated by repressors for a certain time, etc. We study the integration of several mechanisms of regulation and their performances in terms of variance and distribution. As all molecules tends to move freely into the cytoplasm, it is assumed that the encounter time between a given entity and its target is exponentially distributed.

4.5.1.1. Feedback model

We have also investigated the production of a single protein, with the transcription and the translation steps, but we also introduced a direct feedback on it: the protein tends to bind on the promoter of its own gene, blocking therefore the transcription. The protein remains on it during an exponential time until its detachment caused by thermal agitation.

The mathematical analysis aims at understanding the nature of the internal noise of the system and to quantify it. We tend to test the hypothesis usually made that such feedback permits a noise reduction of protein distribution compared to the “open loop” model. We have made the mathematical analysis of the model (using a scaling to be able to have explicit results), it appeared that reduction of variance compared to an “open loop” model is limited: the variance cannot be reduced for more than 50%.

We proposed another possible effect of the feedback loop: the return to equilibrium is faster in the case of a feedback model compared to the open loop model. Such behaviour can be beneficial for the bacteria to change of command for a new level of production of a particular protein (due, for example, to a radical change in the environment) by reducing the respond time to reach this new average. This study has been mainly performed by simulation and it has been shown that the feedback model can go 50% faster than the open loop results. See [13]..

4.5.1.2. Transcription-translation model for all proteins

The other model that has been studied integrates the production of all the proteins. Each gene has to be transcribed in mRNA (using RNA-Polymerase molecules) and each mRNA has to be translated in protein (using ribosome molecules). Experiments (as the one from Taniguchi et al. (2010)) have shown that protein production is subject to high variability especially for highly expressed proteins. Our goal is to determine what in the protein production mechanism is responsible for the noise.

We already made simulations that takes into amount of RNA-Polymerases and Ribosomes and that genes and mRNAs sequester these molecules during the whole the time of elongation. This global sharing of Ribosomes/RNA-Polymerases reproduce only a part of the unknown noise experimentally seen. We are developing Python simulations that extends this model and take into account other feature that might be responsible for the noise in protein production. This new simulation will include new features such as:

- The volume of the cell. We consider it as proportional to the total number of proteins, and will increase as the cell grows. Transcription and translation initiation are then depending on the concentration of respectively free RNA-polymerase and free ribosomes.
- The division of the cell. At division, all components have an equal chance to go in either one of the two daughter cell.
- DNA replication. At some point in the cell cycle, the genome duplicates, doubling therefore the copy number of each gene

The simulation parameters will be fit with the data of Taniguchi et al. (2010) and the goal is to compare our result to see if which aspects of the protein production are responsible for the noise of the proteins.

4.5.2. Stochastic Modelling of Protein Polymerization

This is a collaboration with Marie Doumic, Inria MAMBA team.

The first part of our work focuses on the study of the polymerization of protein. This phenomenon is involved in many neurodegenerative diseases such as Alzheimer's and Prion diseases, e.g. mad cow. In this context, it consists in the abnormal aggregation of proteins. Curves obtained by measuring the quantity of polymers formed in in vitro experiments are sigmoids: a long lag phase with almost no polymers followed by a fast consumption of all monomers. Furthermore, repeating the experiment under the same initial conditions leads to somewhat identical curves up to translation. After having proposed a simple model to explain this fluctuations, we studied a more sophisticated model, closer to the reality. We added a conformation step: before being able to polymere, proteins have to misfold. This step is very quick and remains at equilibrium during the whole process. Nevertheless, this equilibrium depends on the polymerization which is happening on a slower time scale. The analysis of these models involves stochastic averaging principles.

The second part concerns the study of telomeres. This work is made in collaboration with Zhou Xu, Teresa Teixeira, from IBCP in Paris.
In eukaryotic cells, at each mitosis, chromosomes are shortened, because the DNA polymerase is not able to duplicate one ending of the chromosome. To prevent loss of genetic information- which could be catastrophic for the cell-chromosomes are equipped with telomeres at their endings. These telomeres do not contain any genetic information; they are a repetition of the sequence T-T-A-G-G-G thousands times. At each mitosis, there is therefore a loss of telomere. As it has a finite length, when the telomeres are too short, the cell cannot divide anymore: they enter in replicative senescence. Our model tries to captures the two phases of the shortening of telomeres: first, the initial state of the cells, when the telomerase is still active to repair the telomeres. Second, when the telomerase is inhibited, we try to estimate the senescence threshold, when the replication of the cells stops.

SOCRATE Project-Team

6. New Results

6.1. Flexible Radio Front-End

The contributions of members of this axis are mainly on four topics: Wake-Up Radio,Full-Duplex transceivers, SDR Gateways for Urban Networks, and Channel Estimation. In the global concept of enhancing wireless communications, those four topics are complimentary, addressing the reduction of energy consumption, the increase of throughput and/or flexibility of the transmission and the performance evaluation.

6.1.1. Wake-Up Radio

The last decades have been really hungry in new ways to reduce energy consumption. That is especially true when talking about wireless sensor networks in general and home multimedia networks in particular, since electrical energy consumption is the bottleneck of the network. One of the most energy-consuming functional block of an equipment is the radio front end, and methods to switch it off during the time intervals where it is not active must be implemented. This study [10] has proposed a wake-up radio circuit which is capable of both addressing and waking up not only a more efficient but also more energy-consuming radio front end. By using a frequency footprint to differentiate each sensor, awaking all the sensors except for the one of interest is avoided. The particularity of the proposed wake-up receiver is that the decision is taken in the radio-frequency part and no baseband treatment is needed. The global evaluation in theory and in simulation was performed, and a first testbed of this technology was fabricated.

6.1.1.1. Full-Duplex

This work studies [8] a Full-Duplex Dual-Band (FDDB) OFDM radio architecture that enables the radio transceiver to be more flexible and provides a viable radio link capacity gain. A simple but practical I/Q imbalance estimation and compensation method, based on the frequency-flat-fading behavior of the self-interference channel, is proposed. The performance of the proposed I/Q imbalance compensation method is evaluated by system level simulations conducted with ADS and Matlab. The co-simulation results show that the proposed radio transceiver could potentially increase the physical layer transmission rate by four times compared to the conventional radio link at the cost of tolerable loss of BER performance. The I/Q imbalance compensation method can effectively compensate both high and low I/Q imbalance without the problem of algorithm convergence. Application of this technique for physical layer security has already been proposed.

6.1.1.2. SDR for SRDs

The technologies employed in urban sensor networks are permanently evolving, and thus the gateways of these networks have to be regularly upgraded. The existing method to do so is to stack-up receivers dedicated to one communication protocol. However, this implies to have to replace the gateway every time a new protocol is added to the network. A more practical way to do this is to perform a digitization of the full band and to perform digitally the signal processing, as done in Software-Defined Radio (SDR). The main hard point in doing this is the dynamic range of the signals: indeed the signals are emitted with very different features because of the various propagation conditions. It has been proved that the difference of power between two signals can be so important that no existing Analog-to-Digital Converter (ADC) is able to properly digitize the signals. We propose a solution to reduce the dynamic range of signals before digital conversion. In this study [28], the assumption is made that there is one strong signal, and several weak signals. This assumption is made from the existing urban sensor networks topology. A receiver architecture with two branches is proposed with a “Coarse Digitization Path” (CDP) and a “Fine Digitization Path” (FDP). The CDP allows to digitize the strong signal and to get data on it that is used to reconfigure the FDP. The FDP then uses a notch filter to attenuate the strong signal (and then to reduce the dynamic range of the signals) and digitizes the rest of the band. Another way to relax these specifications on ADCs is an analog processing, such as companding, that should be performed before digitization. The companding technique is usually employed on one signal (and

not on multiple signals that are only separated on the frequency domain). This work [36], [29] studies three companding laws to test their efficiency in relaxing the digitization constraints with multiple signals. A μ-law, a Piecewise-Linear (PL) law and a Piecewise-Linear, Constant Gain with Offsets (PLCGO) law are tested. We have described how to use a PLCGO approach to reduce ADC's complexity, and two implementations of the compressing law are proposed.

6.1.1.3. Channel Estimation

In modern mobile telecommunications, shadow fading has to be modeled by a two-dimensional (2D) correlated random variable since shadow fading may present both cross-correlation and spatial correlation due to the presence of similar obstacles during the propagation. In our study, 2D correlated random shadowing is generated based on the multi-resolution frequency domain ParFlow (MR-FDPF) model. The MR-FDPF model is a 2D deterministic radio propagation model, so a 2D deterministic shadowing can be firstly extracted from it. Then, a 2D correlated random shadowing can be generated by considering the extracted 2D deterministic shadowing to be a realization of it. Moreover, based on the generated 2D correlated random shadowing, a complete 2D semi-deterministic path loss model can be proposed. The proposed methodology [5] can be implemented into system-level simulators where it will be very useful due to its ability to generate realistic shadow fading.

[23] presents the first implementation on software defined radio nodes in the large scale testbed CorteXlab of a radio link estimation technique based on OFDM transmissions. The purpose of this large scale testbed is to offer to the whole scientific community an open tool to test new techniques for multiuser , cooperative and cognitive radio networks in a controlled environment. As the experimentation room was defined in order to offer reproducible measurements, it is important to be able to characterize each radio link between all transceivers. Therefore, we present here the development of a channel sounder directly implemented on the software radio nodes. This paper presents the first implementation on software defined radio nodes in the large scale testbed called CorteXlab of a radio link estimation technique based on OFDM transmissions. The purpose of this large scale testbed is to offer to the whole scientific community an open tool to test new techniques for multiuser , cooperative and cognitive radio networks in a controlled environment. As the experimentation room was defined in order to offer reproducible measurements, it is important to be able to characterize each radio link between all transceivers. Therefore, we proposed the development of a channel sounder directly implemented on the software radio nodes.

6.2. Agile Radio Resource Sharing

This axis addresses the challenges relative to the network perspective of software radio. While the two other axes have their focus on the design of the software radio nodes, axis 2 deals with coexistence and cooperation in a multi-user communications perspective.

A first research direction concerns theoretical limits of different reference scenarios where trade-offs between spectral efficiency, energy efficiency, stability and/or fairness are analyzed. This work exploits multi-users information theory, game theory and stochastic geometry. This year, a particular focus has been put on the interference channel with feedback and on dense wireless networks. New problems have been also investigated with the simultaneous energy and information transmission problem for energy harvesting and some specific attacks in smart grids.

In parallel our research activities are also driven by applicative frameworks. Concerning 4G RAN, a new interference alignment scheme has been proposed, simulated and implemented on CortexLab. This work has been presented as one of the promising technologies proposed by Greentouch. IoT has been identified as a new challenge for 5G with the objective of serving a very large number of nodes per cell, in a connectionless manner and with very small packets. The original transmission technology using ultra narrow band modulation and proposed by Sigfox for large area of IoT nodes has been investigated. A multiband CSMA strategy has been also evaluated in collaboration with CEA-Leti for dense Wifi like IoT access networks. Body area networks (BANs) represent also a very challenging applicative framework, with strong dynamics, interference environments, and low energy requirements. In partnership with Euromedia and Hikob, our studies focused on

dynamic algorithms for information gathering in a sport event broadcast system. Additionnally, localization capabilities at the body scale may offer interesting perspetives but require specific MAC protocols.

6.2.1. Fundamental Limits

6.2.1.1. Energy efficiency - Spectral Efficiency (EE-SE) Tradeoffs in Wireless RANs

The spectral and energy efficiency (SE-EE) trade-off in cellular networks has attracted significant recent interest in the wireless community [1]. The work in [7] studies this fundamental limit with a simple and effective method. The proposed theoretical framework is based on an optimal radio resource allocation of transmit power and bandwidth for the downlink direction, applicable for an orthogonal cellular network. The analysis is initially focused on a single cell scenario, for which in addition to the solution of the main SE-EE optimization problem, it is proved that a traffic repartition scheme can also be adopted as a way to simplify this approach. By exploiting this interesting result along with properties of stochastic geometry, this work is extended to a more challenging multi-cell environment, where interference is shown to play an essential role and for this reason several interference reduction techniques are investigated. Special attention is also given to the case of low signal to noise ratio (SNR) and a way to evaluate the upper bound of EE in this regime is provided. This methodology leads to tractable analytical results under certain common channel properties, and thus allows the study of various models without the need for demanding system level simulations.

6.2.1.2. Interference Channels with Feedback

The capacity region of the two-user linear deterministic (LD) interference channel with noisy output feedback (IC-NOF) is fully characterized in [35], [26]. This result allows the identification of several asymmetric scenarios in which implementing channel-output feedback in only one of the transmitter-receiver pairs is as beneficial as implementing it in both links, in terms of achievable individual rate and sum-rate improvements w.r.t. the case without feedback. In other scenarios, the use of channel-output feedback in any of the transmitter-receiver pairs benefits only one of the two pairs in terms of achievable individual rate improvements or simply, it turns out to be useless, i.e., the capacity regions with and without feedback turn out to be identical even in the full absence of noise in the feedback links. As a byproduct, the exact conditions on the signal to noise ratios on the feedback links to observe an improvement on either a single rate, both single rates, or the sum-rate capacity, for any IC-NOF are also fully described in [41].

6.2.1.3. Simultaneous Energy and Information Transmission

The fundamental limits of simultaneous information and energy transmission in the two-user Gaussian multiple access channel (G-MAC) with and without feedback are fully characterized in [33], [9]. All the achievable information and energy transmission rates (in bits per channel use and energy-units per channel use respectively) are identified. Thus, the information-energy capacity region is defined in both cases. In the case without feedback, an achievability scheme based on power-splitting and successive interference cancelation is shown to be optimal. Alternatively, in the case with feedback (G-MAC-F), a simple yet optimal achievability scheme based on power-splitting and Ozarow's capacity achieving scheme is presented. Three of the most important observations in this work are: (a) The capacity-energy region of the G-MAC without feedback is a proper subset of the capacity-energy region of the G-MAC-F; (b) Feedback can at most double the energy rate for a fixed information rate; and (c) Time-sharing with power control is strictly suboptimal in terms of sum-rate in the G-MAC without feedback.

6.2.1.4. Multiple Access Channel and Broadcast Channel with Linear Feedback Schemes

In [11], it is shown that for the two-user Gaussian broadcast channel with correlated noises and perfect feedback the largest region that can be achieved by linear-feedback schemes equals the largest region that can be achieved over a dual multi-access channel when in this latter the channel inputs are subject to a "non-standard" sum-power constraint that depends on the BC-noise correlation. Combining this new duality result with Ozarow's MAC-scheme gives an elegant achievable region for the Gaussian BC with correlated noises. A constructive iterative coding scheme is then presented for the non-symmetric Gaussian BC with uncorrelated noises that is sum-rate optimal among all linear-feedback schemes. This coding scheme shows that the connection between the MAC and the BC optimal schemes is tighter than what is suggested by our duality result on achievable rates. In fact, it is linear-feedback sum-rate optimal to use Ozarow MAC-encoders and MAC-decoders— rearranged—to code over the BC.

6.2.2. Low Complexity Receivers for Massive MIMO Systems

In wireless communications, Multi-user massive MIMO network is a scenario that has been recently proposed, where many mobile terminals are served by a Base Station (BS) equipped with a very high number of antennas. In such a scenario, the detection in the uplink remains a challenge, since the BS is required to detect signals transmitted from all users while trying to exploit full received diversity. The optimal detection criterion that fulfills the diversity requirement is the Maximum-Likelihood (ML) joint detection which has been proposed to detect jointly the transmitted signals. However, such a criterion is not applicable to the addressed multi-user massive MIMO scenario due to its computational complexity that increases exponentially with the number of signals to be detected. In our work paper, we have proposed a relaxed ML detector based on an iterative decoding strategy that reduces the computational cost. We exploit the fact that the transmit constellation is discrete, and remodel the channel as a MIMO channel with sparse input belonging to the binary $\{0, 1\}$. The sparsity property allows us to relax the ML problem as a quadratic minimization under linear and $\ell 1$-norm constraint. We then prove the equivalence of the relaxed problem to a convex optimization problem solvable in polynomial time. Simulation results illustrate the efficiency of the low-complexity proposed detector compared to other existing ones in very large and massive MIMO context.

6.2.3. Distributed Radio Resource Management

6.2.3.1. Interference Alignment in Cellular Networks with no-Explicit Coordination

Current networks aim to support high data rates for end users by increasing the spectral efficiency in bits-per-Hertz, at the expense of the energy efficiency of the network. Indeed, an important part of the energy consumption of mobile networks is proportional to the radiated energy, which relies on the frequency bandwidth and the transmission power. Any energy efficient transmission scheme should exploits the whole system bandwidth by allocating the entire available spectrum to each base station. Such an approach, however, leads to significant interference increase and performance degradation for mobiles located at the cell edges. The key challenge is to balance interference avoidance and spectrum use to reach an optimal spectral efficiency – energy efficiency (EE-SE) trade-off. The work achieved in the framework of Greentouch collaboration is based on the non classical interference alignement scheme proposed by Suh and Tse in dowlink mode. The key contribution relies on users scheduling with a unique criteria based as well on QoS priorities and orthogonality of precoding directions. The spectral efficiency is improved by a factor 2 for edge users and a energy saving of about 30% is made possible. This scheme has been evaluated on simulation scenarios as defined by Greentouch partners and a simplified version has been implemented on FIT/CorteXlab and demontrastrated during the final event of Greentouch (New-York, June 2015).

6.2.4. RANs for IoT : Dense and Connectionless Solutions

Internet of Things (IoT) is going to take a major place in the telecommunications market as announced in technical and public medias. The paradigm of IoT relies on the deployment of billions of objects having the capability of transmitting information about their context and environment and to create a real-time, secured and efficient interaction between the real and the virtual worlds, pushing them to evolve from the state of cousins to the state of Siamese twins. IoT revealed to be a key technology for solving societal issues such as digital cities, intelligent transportation, green environment monitoring or medical care and elderly person monitoring.

IoT has strong connections with machine-to-machine (M2M), and sometimes in literature, both terms refer to the same idea. From our point of view, IoT covers a broader scope including as well the technology and the applications. On the opposite, M2M refers to the technologies that allow machines or objects to communicate.

In any case, from the technical point of view, the main challenge of this new paradigm is to let a huge number of machine type devices (MTDs) be connected to the Internet at a low cost, with a limited infrastructure and featuring a very long life time with very small battery or energy needs [4].

In this global picture, we may consider different technical issues. M2M has first been defined to connect MTDs in their vicinity. The proposed solutions extensively rely on the research results produced over the last 20 years for ad-hoc and wireless sensor networks. Initiated 20 years ago from theoretical concepts, this very active research area has gone up to the definition of full standards (802.15.4, 802.15.6, Zigbee, Bluetooth) which have already found a market.

More recently, the IoT paradigm has been extended to the problem of connecting all these MTDs to the Internet, and through Internet to anyone or anything. The massive connection of objects spread over the world is a challenge that has some similarities with the paradigm of cellular networks which aimed at connecting people. This similarity attracted the interest of mobile network providers, to exploit such attractive potential market and IoT has been identified as a target for the future 5G.

6.2.4.1. Performance of Ultra-NarrowBand Techniques

The Ultra-narrow-band technology is an appealing solution for the low throughput wireless sensor networks (10b/s - 1 kb/s). It is complementary to the classical cellular networks thanks to its low energy consumption and very long range communication (up to 50 km in free-space) [4]. This technology has already been deployed and is proved to be ultra-efficient for point-to-point communications in Sigfoxs' network. Nodes are transmitting at a random time and random frequency carrier (random frequency division multiple access schemes : R-FDMA), so the uplink is exposed to interference. In our approach, we have proposed to model this interference for the UNB network when taking into account the path-loss and Rayleigh effects, with stochastic geometry tools. The obtained model allows us to estimate the system performance, and its capacity in terms of maximum number average of simultaneous nodes in a unique cell [37]. We have also considered the replication mechanism, and identified the optimum number of replications.

6.2.4.2. Multiband CSMA for Dense Wireless Networks in Uplink

In this approach, the objective is to mitigate the degradation of the throughput and delay performance in wireless local area networks (WLAN) that employ carrier sense multiple access collision avoidance (CSMA/CA) protocol with request to send and clear to send (RTS/CTS) mechanism, when a large number of IoT like nodes are deployed. In our approach, the overhead is reduced with a modified handshake mechanism. The medium access control (MAC) overhead caused by the RTS and CTS messages is high comparing to the total duration of successful transmission. In order to reduce the MAC overhead we propose in this work a new strategy to serve many users successively. This strategy consists on sending many RTS in parallel by different stations on different frequency sub-bands. Once the RTS messages do not collide with each other, there will be no need to resend the RTS and wait for a CTS to gain the channel access [21].

6.2.5. Algorithms and Protocols for BANs

6.2.5.1. Information Gathering in a Group of Mobile Users

Distributed decisions within any group of agents, is a very active research area and theoretical results as well as efficient algorithms have already been proposed but in the context of wireless networks, the task is made harder due to possible transmission errors, channel asymmetry, dynamic behaviour of the channel and node mobility. In this work, we consider a group of mobile agents moving roughly in a common direction. We study different algoritimic solutions allowing each agent to periodically discover its neighbours: one-hop neighbours as well as multi-hop neighbors. The reference scenario is a bike race, during which groups are susceptible to split or merge. The objective is a live gathering of information about who is present in a group for live TV broadcasting. For that, we need a fully distributed approach allowing every agent to discover with a consensus algorithm the list of neighbours participating to the same pack. This study may be of interest for various other applications such as group navigation support in crowded environments, autonomous navigation of a fleet of robots. . . This problem exhibits some similarities with a clustering problem. However, a clustering problem aims at exploiting the structure of a graph and to form some subgroups to ensure a good structure of the network for further communications while our objective is rather to estimate the groups naturally formed in the real world. Hence, we have focused on distributed decision algorithms, which are widely present in the literature. Max-consensus problem has been much less studied than average consensus. The proposed algorithms are based on the N-dimension generalization of the Random Broadcast Max-Consensus algorithm,

allowing each agent to build and share the list of its muli-hop neighbors. We extend this approach to a dynamic context where the group information needs to be updated according to possible group merge or split. Experimental validation has been done in the context of a cycling race with 10 agents, equipping each bicycle with a wireless sensor node to assess the interactions between the racers and to provide a live monitoring of the dynamic evolution of the cyclists groups that form during the race.

6.2.5.2. MAC Protocols and Algorithms for Localization at the Body Scale

The purpose of this work is to evaluate the impact of the node speed on the ranging estimation for location applications with Wireless Body Area Networks (WBAN). While estimated with the 3-Way ranging protocol (3-WR), this distance between two nodes placed on the body can be affected by the human movements [30], [17]. Thus, we study theoretically the ranging error with the 3-WR, based on a perfect channel, a MAC layer based on TDMA using two scheduling strategies (Single node localization (P2P-B) and Aggregated & Broadcast (A&B)) and a PHY layer based on Ultra Wideband (IR-UWB) [31]. We demonstrate the accuracy of the model, and show that the distance error is highly correlated with the speed of nodes [16], while the associated mobility model has an impact on the design of MAC strategies by simulation [18].

6.2.6. Other Topics

6.2.6.1. Data Injection Attacks in Smart Grids

Multiple attacker data injection attack construction in electricity grids with minimum-mean-square-error (MMSE) state estimation is studied for centralized and decentralized scenarios in [34]. A performance analysis of the trade-off between the maximum distortion that an attack can in- troduce and the probability of the attack being detected by the network operator is considered. Within this setting, optimal centralized attack construction strategies are studied. The decentralized case is examined in a game-theoretic setting. A novel utility function is proposed to model this trade-off and it is shown that the resulting game is a potential game. The existence and cardinality of the corresponding set of Nash Equilibria (NE) in the game is analyzed. For the particular case of two attackers, numerical results based on IEEE test systems are presented. These results suggest that attackers perform better when they seize control of power flow measurements instead of power injection measurements.

6.3. Software Radio Programming Model

6.3.1. Data Flow Programming

Streaming languages have been proven to be a natural and efficient approach for taking advantage of the intrinsic parallelism of modern CPU architectures. The focus of many previous work has been to improve the throughput of streaming programs. In [27], we rather focus on satisfying quality-of-service requirements of streaming applications executed alongside non-streaming processes. We monitor synchronous dataflow (SDF) programs at runtime both at the application and system levels, in order to identify violations of quality-of-service requirements. Our monitoring requires the programmer to provide the expected throughput of its application (e.g 25 frames per second for a video decoder), then takes full benefit from the compilation of the SDF graph to detect bottlenecks in this graph and identify causes among processor or memory overloading. It can then be used to perform dynamic adaptations of the applications in order to optimize the use of computing and memory resources.

6.3.2. Smart Sensors

The article [19] presents the development of a wireless wearable sensor for the continuous, long-term monitoring of cardiac activity. Heart rate assessment, as well as heart rate variability parameters are computed in real time directly on the sensor, thus only a few parameters are sent via wireless communication for power saving. Hardware and software methods for heart beat detection and variability calculation are described and preliminary tests for the evaluation of the sensor are presented. With an autonomy of 48 hours of active measurement and a Bluetooth Low Energy radio technology, this sensor will form a part of a wireless body network for the remote mobile monitoring of vital signals in clinical applications requiring automated collection of health data from multiple patients.

6.3.3. Cryptography

For security applications in wireless sensor networks (WSNs), choosing best algorithms in terms of energy-efficiency and of small memory requirements is a real challenge because the sensor networks are composed of low-power entities. Previous works benchmarked 12 block-ciphers on an ATMEL AVR ATtiny45 8-bit microcontroller. In [2], most of the recent lightweight block cipher proposals, as well as some conventional block ciphers, are studied on the Texas Instruments MSP430 16-bit microcontroller . The chosen block ciphers are described with a security and an implementation summary. Implementations are then evaluated on a dedicated platform.

6.3.4. Hardware Arithmetic

6.3.4.1. Hardware Implementations of Fixed-Point Atan2

The atan2 function computes the polar angle arctan(x/y) of a point given by its cartesian coordinates. It is widely used in digital signal processing to recover the phase of a signal. The article [14] studies for this context the implementation of atan2 with fixed-point inputs and outputs. It compares the prevalent CORDIC shift-and-add algorithm to two multiplier-based techniques. The first one reduces the bivariate atan2 function to two functions of one variable: the reciprocal, and the arctangent. These two functions may be tabulated, or evaluated using bipartite or polynomial approximation methods. The second technique directly uses piecewise bivariate polynomial approximations, in degree 1 and degree 2. It requires larger tables but has the shortest latency. Each of these approaches requires a relevant argument reduction, which is also discussed. All the algorithms are described with the same accuracy target (faithful rounding) and implemented with similar care in FloPoCo. Based on synthesis results on FPGAs, their relevance domains are discussed.

6.3.4.2. Fixed-Point Implementations of the Reciprocal, Square Root and Reciprocal Square Root Functions

Implementations of the reciprocal, square root and reciprocal square root often share a common structure. The article [39] is a survey and comparison of methods for computing these functions. It compares classical methods (direct tabulation, multipartite tables, piecewise polynomials, Taylor-based polynomials, Newton-Raphson iterations). It also studies methods that are novel in this context: the Halley method and, more generally, the Householder method. The comparisons are made in the context of the same accuracy target (faithful rounding) and of an arbitrary fixed-point format for the inputs and outputs (precisions of up to 32 bits). Some of the methods discussed might require some form of range reduction, depending on the input range. The objective of the article is to optimize the use of fixed-size FPGA resources (block multipliers and block RAMs). The discussions and conclusions are based on synthesis results for FPGAs.

6.3.4.3. Fixed-Point Hardware Polynomials

Polynomial approximation is a general technique for the evaluation of numerical functions of one variable such as atan, reciprocal and square roots studied above. The article [38] addresses the automatic construction of fixed-point hardware polynomial evaluators. By systematically trying to balance the accuracy of all the steps that lead to an architecture, it simplifies and improves the previous body of work covering polynomial approximation, polynomial evaluation, and range reduction. This work is supported by an open-source implementation in FloPoCo.

6.3.5. Software Elementary Functions

6.3.5.1. Code Generators for Mathematical Functions

A typical floating-point environment includes support for a small set of about 30 mathematical functions such as exponential, logarithms and trigonometric functions. These functions are provided by mathematical software libraries (libm), typically in IEEE754 single, double and quad precision. The article [13] suggests to replace this libm paradigm by a more general approach: the on-demand generation of numerical function code, on arbitrary domains and with arbitrary accuracies. First, such code generation opens up the libm function space available to programmers. It may capture a much wider set of functions, and may capture even standard functions on non-standard domains and accuracy/performance points. Second, writing libm code requires fine-tuned instruction selection and scheduling for performance, and sophisticated floating-point

techniques for accuracy. Automating this task through code generation improves confidence in the code while enabling better design space exploration, and therefore better time to market, even for the libm functions. This article discusses, with examples, the new challenges of this paradigm shift, and presents the current state of open-source function code generators.

6.3.5.2. Computing Floating-Point Logarithms with Fixed-Point Operations

Elementary functions from the mathematical library input and output floating-point numbers. However it is possible to implement them purely using integer/fixed-point arithmetic. This option was not attractive between 1985 and 2005, because mainstream processor hardware supported 64-bit floating-point, but only 32-bit integers. Besides, conversions between floating-point and integer were costly. This has changed in recent years, in particular with the generalization of native 64-bit integer support. The purpose of the article [40] is therefore to reevaluate the relevance of computing floating-point functions in fixed-point. For this, several variants of the double-precision logarithm function are implemented and evaluated. Formulating the problem as a fixed-point one is easy after the range has been (classically) reduced. Then, 64-bit integers provide slightly more accuracy than 53-bit mantissa, which helps speed up the evaluation. Finally, multi-word arithmetic, critical for accurate implementations, is much faster in fixed-point, and natively supported by recent compilers. Novel techniques of argument reduction and rounding test are introduced in this context. Thanks to all this, a purely integer implementation of the correctly rounded double-precision logarithm outperforms the previous state of the art, with the worst-case execution time reduced by a factor 5. This work also introduces variants of the logarithm that input a floating-point number and output the result in fixed-point. These are shown to be both more accurate and more efficient than the traditional floating-point functions for some applications.

URBANET Team

7. New Results

7.1. Characterizing and deploying urban networks

Participants: Ahmed Boubrima, Angelo Furno, Diala Naboulsi, Patrice Raveneau, Walid Bechkit, Marco Fiore, Hervé Rivano, Razvan Stanica.

7.1.1. Collection and Analysis of Mobile Phone Data

Cellular communications are undergoing significant evolutions in order to accommodate the load generated by increasingly pervasive smart mobile devices. At the same time, recent generations of mobile phones, embedding a wide variety of sensors, have fostered the development of open sensing applications, while cellular operators are looking for new services they can provide using the data collected on their side, in the access or the core network.

The analysis of operator-side data is a recently emerged research field, and, apart a few outliers, relevant works cover the period from 2005 to date, with a sensible densification over the last three years. In [9], we provided a thorough review of the multidisciplinary activities that rely on mobile traffic datasets, identifying major categories and sub-categories in the literature, so as to outline a hierarchical classification of research lines and proposing a complete introductory guide to the research based on mobile traffic analysis. The usage of these datasets in the design of new networking solutions, in order to achieve the so-called *cognitive networking* paradigm, is discussed in detail in the PhD thesis of Diala Naboulsi [2], where the examples of green networking and virtualized radio access networks are given.

When constructing a social network from interactions among people (e.g., phone calls, encounters), a crucial task is to define the threshold that separates social from random (or casual) relationships. The ability to accurately identify social relationships becomes essential to applications that rely on a precise description of human routines, such as recommendation systems, forwarding strategies and opportunistic dissemination protocols. We thus proposed a strategy to analyze users' interactions in dynamic networks where entities act according to their interests and activity dynamics [10]. Our strategy allows classifying users interactions, separating random ties from social ones, and unveils significant differences among the dynamics of users' wireless interactions in the datasets.

Furthermore, mobile traffic data has been recently used to characterize the urban environment in terms of urban fabric profiles. While showing promising results, the existing urban fabric detection solutions are built without a clear understanding of the detection process chain. In [16], we distinguished and analyzed the different steps common to all urban profiling techniques. By evaluating the impact of each step of the process, we were able to propose a new solution that outperforms the state of the art techniques. Our approach uses the weekly periodicity of human activities, as well as a median-based filtering technique, resulting in a better clustering in terms of both coverage and entropy, as shown by results obtained on two large scale mobile traffic datasets covering the urban areas of Milan and Turin, in Italy. The solution proposed in this work was selected among the 10 finalists of the Telecom Italia Big Data challenge.

A second source of mobile data is the smartphone itself. In the context of the PrivaMov project, funded by the Labex IMU, we developed and deployed a data collection platform on more than 100 Android devices. A first step in the study of this enormous dataset (more than 50 Gb have been collected to date) was presented in [21], with a focus on the extraction of user mobility information and Wi-Fi mapping. This led us to the study of Wi-Fi tracking, a method relying on signals emitted by portable devices to track individuals for commercial, security or surveillance purposes. Wi-Fi tracking has the potential to passively track a large fraction of the population and is therefore an ideal population surveillance technology and a serious privacy threat. In [19], we argue that Wi-Fi routers make an ideal building block to create a large scale Wi-Fi tracking system, showing how they can be easily turned into Wi-Fi tracking devices through software modification. We

provided a first evaluation of the tracking capabilities of an hypothetical Wi-Fi tracking system through a set of simulations based on real-world datasets. Results showed that the spatial distribution of Wi-Fi routers is such that compromising even a small fraction of Wi-Fi routers is sufficient to track people for a large fraction of the time.

Preservation of user privacy is therefore paramount in the publication of datasets that contain fine-grained information about individuals. The problem is especially critical in the case of mobile traffic datasets collected by cellular operators, as discussed above, as they feature high subscriber trajectory uniqueness and they are resistant to anonymization through spatiotemporal generalization. In [17], we first unveiled the reasons behind such undesirable features of mobile traffic datasets, by leveraging an original measure of the anonymizability of users' mobile fingerprints. Building on such findings, we proposed GLOVE, an algorithm that grants k-anonymity of trajectories through specialized generalization. We evaluated our methodology on two nationwide mobile traffic datasets, and show that it achieves k-anonymity while preserving a substantial level of accuracy in the data.

7.1.2. Deployment of Wireless Sensor Networks for Pollution Monitoring

Recently, air pollution monitoring emerged as one of the main services of smart cities because of the increasing industrialization and the massive urbanization. Wireless Sensor Networks are a suitable technology for this purpose, thanks to their substantial benefits including low cost and autonomy. Minimizing the deployment cost is one of the major challenges in the design of such networks, therefore sensors positions have to be carefully determined. In [13], we proposed two integer linear programming formulations based on real pollutants dispersion modeling to deal with the minimum cost sensor network deployment for air pollution monitoring. We illustrated the concept by applying our models on real world data, namely the Nottingham City street lights. We compared the two models in terms of execution time and showed that the second flow-based formulation is much better. We finally conducted extensive simulations to study the impact of some parameters and derive some guidelines for efficient urban sensor deployment for air pollution monitoring.

7.2. Technology specific solutions

Participants: Jin Cui, Walid Bechkit, Khaled Boussetta, Hervé Rivano, Fabrice Valois.

7.2.1. Temperature-Aware Algorithms for Wireless Sensor Networks

Temperature variations have a significant effect on low power wireless sensor networks as wireless communication links drastically deteriorate when temperature increases. A reliable deployment should take temperature into account to avoid network connectivity problems resulting from poor wireless links when temperature increases. A good deployment needs also to adapt its operation and save resources when temperature decreases and wireless links improve. Taking into account the probabilistic nature of the wireless communication channel, in [12] we investigated the effect of temperature on percolation-based connectivity in large scale wireless sensor networks and showed that more energy can be saved by allowing some nodes to go to deep sleep mode when temperature decreases and links improve. Based on this result, we proposed a simple, yet efficient, Temperature-Aware MAC plugin (TA-MAC), which can be potentially used with any MAC protocol, enabling it to dynamically adapt the network effective density in order to allow further energy savings, while maintaining network connectivity. We carried out simulations and demonstrated that sate of the art protocols augmented with the TA-MAC plugin allow a significant energy efficiency improvement.

Going one step further, we developed a mathematical model that provides the most energy efficient deployment in function of temperature without compromising the correct operation of the network by preserving both connectivity and coverage [3]. We used our model to design three temperature-aware algorithms that seek to save energy (i) by putting some nodes in hibernate mode as in the SO (Stop-Operate) algorithm in TA-MAC, or (ii) by using transmission power control as in PC (Power-Control), or (iii) by doing both techniques as in SOPC (Stop-Operate Power-Control). All proposed algorithms are fully distributed and solely rely on temperature readings without any information exchange between neighbors, which makes them low overhead and robust. Our results identified the optimal operation of each algorithm and showed that a significant amount of energy can be saved by taking temperature into account.

7.2.2. Resilience in Wireless Sensor Networks

The concept of resilience for routing protocols in wireless sensor networks has been proposed and developed in the team in the last few years. In our previous works, a general overview of the resilience, including definition, metric and resilient techniques based on random behavior and data replication have been proposed. Following these previous methods, in [6] we proposed a new resilient solution based on network coding techniques, to improve resilience in wireless sensor networks for smart metering applications. More precisely, using our resilience metric based on a performance surface, we compared several variants of a well-known gradient based routing protocol with the previous methods (random routing and packet replications) and the new proposed methods (two network coding techniques). The proposed methods outperformed the previous methods in terms of data delivery success even in the presence of high attack intensity.

We also continued to study the resilience of routing protocols against malicious insiders willing to disrupt network communications. Previously, the simulation results showed that introducing randomness in routing protocols increases uncertainty for an adversary, making the protocols unpredictable. When combined with data replication, it permits route diversification between a source and a destination, thus enhancing the resilience. In [15], we proposed a theoretical framework to quantify analytically the performance of random protocols against attacks based on biased random walks on a torus lattice. The objective is to evaluate analytically the influence of bias and data replication introduced to random walks. The bias allows to decrease the route length by directing random walks toward the destination, thus reducing the probability of a data packet to meet a malicious insider along the route; however, it decreases also the degree of randomness (entropy). When random protocols are combined with data replication, the reliability is improved thanks to route diversity despite an additional overhead in terms of energy consumption.

7.2.3. Data aggregation in Wireless Sensor Networks

Aggregation functions are intended to save energy and capacity in Wireless Sensor Networks, by avoiding unnecessary transmissions. Aggregation functions take benefit from spatial and/or temporal correlations to forecast or to compress the real data which are collected. Although several works have focused on data aggregation in Wireless Sensor Networks, there is a lack of a formal unified framework that can compare several aggregation functions suitable for a given network topology, a given application and a target accuracy. In [14], we address this question by proposing a Markov Decision Process that can help to evaluate the performances of aggregation functions. The performances are expressed using two new proposed metrics, which can assess the energy and capacity savings of aggregation functions. As illustrative examples, we use our Markov Decision Process to evaluate and analyze the performances of basic aggregation functions (e.g. average) and more complex ones (time series, polynomial functions).

7.2.4. Data Gathering in Mesh Networks

In the gathering problem in mesh networks, a particular node in a graph, the base station, aims at receiving messages from some nodes in the graph. At each step, a node can send one message to one of its neighbors (such an action is called a call). However, a node cannot send and receive a message during the same step. Moreover, the communication is subject to interference constraints, more precisely, two calls interfere in a step, if one sender is at distance at most d_I from the other receiver. Given a graph with a base station and a set of nodes having some messages, the goal of the gathering problem is to compute a schedule of calls for the base station to receive all messages as fast as possible, i.e., minimizing the number of steps (called makespan). The gathering problem is equivalent to the personalized broadcasting problem where the base station has to send messages to some nodes in the graph, with same transmission constraints.

In [5], we focused on the gathering and personalized broadcasting problem in grids. Moreover, we considered the non-buffering model: when a node receives a message at some step, it must transmit it during the next step. In this setting, though the problem of determining the complexity of computing the optimal makespan in a grid is still open, we presented linear (in the number of messages) algorithms that compute schedules for gathering with $d_I \in 0, 1, 2$. In particular, we presented an algorithm that achieves the optimal makespan up to an additive constant 2 when $d_I = 0$. If no messages are “close” to the axes (the base station being the origin),

our algorithms achieve the optimal makespan up to an additive constant 1 when $d_I = 0$, 4 when $d_I = 2$, and 3 when both $d_I = 1$ and the base station is in a corner.

7.3. Capillary Network Solutions

Participants: Patrice Raveneau, Trista Lin, Marco Fiore, Hervé Rivano, Razvan Stanica.

7.3.1. Connected Vehicles

Managing user mobility is historically one of the most critical issues in cellular radio access networks (RANs). That task will become an even greater challenge due to cellular users on-board vehicles and networked cars that autonomously access Internet-based services, whose number is expected to grow dramatically in the next few years. There is thus a need to characterize RAN access from/by vehicles in a similar way to what has been done for traditional pedestrian access. In [11], we proposed a first study of the macroscopic and microscopic features of pervasive vehicular access in a case-study large-scale urban environment, in presence of realistic datasets of the road traffic and RAN deployment. We found that pervasive vehicular access is characterized by unique temporal and spatial variability in the urban region, such that it may require a dedicated RAN capacity planning: the presence of stable vehicular access load patterns and mobility flows can help to that end. Also, we identified the theoretical distributions that best fit key metrics for RAN planning, i.e., the vehicular users' inter-arrival and residence times at cells, and discuss how their parameters vary over time and space.

Smart parking, allowing drivers to access parking information through their smart-phone, is another important service for vehicular users, which can be provided not only through cellular networks, but also by using metropolitan wireless networks, whose deployment strategy needs to be guided by efficiency and functionality. In [8], we introduced and studied a deployment strategy for wireless on-street parking sensor networks. We defined a multiple-objective problem in our analysis, and solved it with two real-world street parking maps. We presented the results on the tradeoff among minimum energy consumption, sensing information delay and the amount of deployed mesh routers and Internet gateways, i.e., the cost of city infrastructure. We also analyzed these tradeoffs to see how different urban layouts affect the optimal solutions. The overall smart parking architecture and services made the object of the PhD thesis of Trista Lin [1], where the analysis of the entire system can be found, including results on the wireless sensor networks used to collect data from parking places and the Publish-Subscribe service used to disseminate this information to users.

7.3.2. Offloading Cellular Networks

Offloading is a promising technique for alleviating the ever-growing traffic load from infrastructure-based networks such as the Internet. Offloading consists in using alternative methods of transmission as a cost-effective solution for network operators to extend their transport capacity. Wi-Fi offloading is one of the most effective approaches to relieve the cellular radio access from part of the burgeoning mobile demand. To date, Wi-Fi offloading has been mainly leveraged in limited contexts, such as home, office or campus environments. In [18], we investigated the scaling properties of Wi-Fi offloading, by studying how it would perform on a much larger scope than those considered today. To that end, we considered a real-world citywide scenario, built on data about actual infrastructure deployments and mobile traffic demand, and observed which amount of traffic could be accommodated by the existing pervasive Wi-Fi access infrastructure, were it opened to mobile users. We found that more than 80% of the mobile traffic demand in a large urban area may be easily served by Wi-Fi access points, under a wide range of system settings.

A new offloading technique was introduced in [20] and further detailed in [4], where we advocate the use of conventional vehicles equipped with storage devices as data carriers whilst being driven for daily routine journeys. The road network can be turned into a large-capacity transmission system to offload bulk transfers of delay-tolerant data from the Internet. The challenges we addressed include how to assign data to flows of vehicles and while coping with the complexity of the road network. We proposed an embedding algorithm that computes an offloading overlay where each logical link spans over multiple stretches of road from the underlying road infrastructure. We then formulated the data transfer assignment problem as a novel linear programming model we solve to determine the optimal logical paths matching the performance requirements

of a data transfer. We evaluated our road traffic allocation scheme using actual road traffic counts in France. The numerical results show that 20% of vehicles in circulation in France equipped with only one Terabyte of storage can offload Petabyte transfers in a week.