RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2014

Section New Results

Edition: 2015-03-24

ASAP Project-Team

6. New Results

6.1. Highlights of the Year

- Anne-Marie Kermarrec is the recipient of the **ACM/IFIP/USENIX/Middleware 10-Years Best Paper Award**, for her paper *The peer sampling service: Experimental evaluation of unstructured gossip-based implementations* (Middleware 2004), co-authored with Márk Jelasity, Rachid Guerraoui, and Maarten van Steen.
- Anne-Marie Kermarrec is the recipient of the **WISE 2014 Best Paper Award**, for her paper [18], co-authored with Alexandra Olteanu and Karl Aberer.
- Michel Raynal is the recipient of the **PODC 2014 Best Paper Award**, for his paper [34], co-authored with Achour Mostefaoui and Moumen Hamouna.
- The MEDIEGO recommendation engine was demonstrated at **Le Web 14** in partnership with FranceTV.

BEST PAPERS AWARDS :
[18] **15th International Conference on Web Information System Engineering (WISE 2014)**. O. ALEXANDRA, A.-M. KERMARREC, K. ABERER.
[34] **ACM PODC**. A. MOSTEFAOUI, M. HAMOUNA, M. RAYNAL.

6.2. Models and abstractions for distributed systems

6.2.1. Signature-free asynchronous Byzantine consensus

Participant: Michel Raynal.

In [34] we present a new round-based asynchronous consensus algorithm that copes with up to $t < n/3$ Byzantine processes, where n is the total number of processes. In addition of not using signature, not assuming a computationally-limited adversary, while being optimal with respect to the value of t, this algorithm has several noteworthy properties: the expected number of rounds to decide is four, each round is composed of two or three communication steps and involves $O(n^2)$ messages, and a message is composed of a round number plus a single bit. To attain this goal, the consensus algorithm relies on a common coin as defined by Rabin, and a new extremely simple and powerful broadcast abstraction suited to binary values. The main target when designing this algorithm was to obtain a cheap and simple algorithm. This was motivated by the fact that, among the first-class properties, simplicity –albeit sometimes under-estimated or even ignored– is a major one.

This is a joint work with Achour Mostéfaouin and Hamouma Moumen. It received the PODC 2014 Best Paper Award.

6.2.2. Randomized mutual exclusion with constant amortized RMR complexity on the DSM

Participant: George Giakkoupis.

In [30] we settle an open question by determining the remote memory reference (RMR) complexity of randomized mutual exclusion, on the distributed shared memory model (DSM) with atomic registers, in a weak but natural (and stronger than oblivious) adversary model. In particular, we present a mutual exclusion algorithm that has constant expected amortized RMR complexity and is deterministically deadlock free. Prior to this work, no randomized algorithm with $o(\log n/\log\log n)$ RMR complexity was known for the DSM model. Our algorithm is fairly simple, and compares favorably with one by Bender and Gilbert (FOCS 2011) for the CC model, which has expected amortized RMR complexity $O(\log^2 \log n)$ and provides only probabilistic deadlock freedom.

This is a joint work with Philipp Woelfel (Univ. of Calgary, Canada).

6.2.3. Reliable shared memory abstraction on top of asynchronous Byzantine message-passing systems

Participants: Michel Raynal, Julien Stainer.

This work is on the construction and the use of a shared memory abstraction on top of an asynchronous message-passing system in which up to t processes may commit Byzantine failures. This abstraction consists of arrays of n single-writer/multi-reader atomic registers, where n is the number of processes. Differently from usual atomic registers which record a single value, each of these atomic registers records the whole history of values written to it. A distributed algorithm building such a shared memory abstraction it first presented. This algorithm assumes $t < n/3$, which is shown to be a necessary and sufficient condition for such a construction. Hence, the algorithm is resilient-optimal. Then we present distributed algorithms built on top of this shared memory abstraction, which cope with up to t Byzantine processes. The simplicity of these algorithms constitutes a strong motivation for such a shared memory abstraction in the presence of Byzantine processes. For a lot of problems, algorithms are more difficult to design and prove correct in a message-passing system than in a shared memory system. Using a protocol stacking methodology, the aim of the proposed abstraction is to allow an easier design (and proof) of distributed algorithms, when the underlying system is an asynchronous message-passing system prone to Byzantine failures.

This work was done in collaboration with Damien Imbs and Sergio Rajsbaum. It has been published in SIRROCCO [32] and as a technical report [43].

6.2.4. Distributed Universality

Participants: Michel Raynal, Julien Stainer.

A notion of a universal construction suited to distributed computing has been introduced by M. Herlihy in his celebrated paper "Wait-free synchronization" (ACM TOPLAS, 1991). A universal construction is an algorithm that can be used to wait-free implement any object defined by a sequential specification. Herlihy's paper shows that the basic system model, which supports only atomic read/write registers, has to be enriched with consensus objects to allow the design of universal constructions. The generalized notion of a k-universal construction has been recently introduced by Gafni and Guerraoui (CONCUR 2011). A k-universal construction is an algorithm that can be used to simultaneously implement k objects (instead of just one object), with the guarantee that at least one of the k constructed objects progresses forever. While Herlihy's universal construction relies on atomic registers and consensus objects, a k-universal construction relies on atomic registers and k-simultaneous consensus objects (which are wait-free equivalent to k-set agreement objects in the read/write system model). This work significantly extends the universality results introduced by Herlihy and Gafni-Guerraoui. In particular, we present a k-universal construction which satisfies the following five desired properties, which are not satisfied by the previous k-universal construction: (1) among the k objects that are constructed, at least ℓ objects (and not just one) are guaranteed to progress forever; (2) the progress condition for processes is wait-freedom, which means that each correct process executes an infinite number of operations on each object that progresses forever; (3) if any of the k constructed objects stops progressing, all its copies (one at each process) stop in the same state; (4) the proposed construction is contention-aware, in the sense that it uses only read/write registers in the absence of contention; and (5) it is generous with respect to the obstruction-freedom progress condition, which means that each process is able to complete any one of its pending operations on the k objects if all the other processes hold still long enough. The proposed construction, which is based on new design principles, is called a (k,ℓ)-universal construction. It uses a natural extension of k-simultaneous consensus objects, called (k,ℓ)-simultaneous consensus objects ((k,ℓ)-SC). Together with atomic registers, (k,ℓ)-SC objects are shown to be necessary and sufficient for building a (k,ℓ)-universal construction, and, in that sense, (k,ℓ)-SC objects are (k,ℓ)-universal.

This work was done in collaboration with Gadi Taubenfeld. It has been published as a brief announcement in PODC [37] and the full version appeared in OPODIS [38]. A version has also been published as a technical report [45].

6.2.5. Computing in the presence of concurrent solo executions

Participants: Michel Raynal, Julien Stainer.

In a wait-free model any number of processes may crash. A process runs solo when it computes its local output without receiving any information from other processes, either because they crashed or they are too slow. While in wait-free shared-memory models at most one process may run solo in an execution, any number of processes may have to run solo in an asynchronous wait-free message-passing model. This work is on the computability power of models in which several processes may concurrently run solo. It first introduces a family of round-based wait-free models, called the d-solo models, $1 \leq d \leq n$, where up to d processes may run solo. We then give a characterization of the colorless tasks that can be solved in each d-solo model. We also introduce the (d,ϵ)-solo approximate agreement task, which generalizes ϵ-approximate agreement, and proves that (d,ϵ)-solo approximate agreement can be solved in the d-solo model, but cannot be solved in the $(d+1)$-solo model. We study also the relation linking d-set agreement and (d,ϵ)-solo approximate agreement in asynchronous wait-free message-passing systems. These results establish for the first time a hierarchy of wait-free models that, while weaker than the basic read/write model, are nevertheless strong enough to solve non-trivial tasks.

This work was done in collaboration with Maurice Herlihy and Sergio Rajsbaum. It has been published in LATIN [31].

6.2.6. A simple broadcast algorithm for recurrent dynamic systems

Participants: Michel Raynal, Julien Stainer.

This work presents a simple broadcast algorithm suited to dynamic systems where links can repeatedly appear and disappear. The algorithm is proved correct and a simple improvement is introduced, that reduces the number and the size of control messages. As it extends in a simple way a classical network traversal algorithm to the dynamic context, the proposed algorithm has also pedagogical flavor.

This work was done in collaboration with Jiannong Cao and Weigang Wu. It has been published in AINA [36].

6.2.7. Fisheye consistency: Keeping data in synch in a georeplicated world

Participants: Michel Raynal, François Taïani.

Over the last thirty years, numerous consistency conditions for replicated data have been proposed and implemented. Popular examples of such conditions include linearizability (or atomicity), sequential consistency, causal consistency, and eventual consistency. These consistency conditions are usually defined independently from the computing entities (nodes) that manipulate the replicated data; i.e., they do not take into account how computing entities might be linked to one another, or geographically distributed. To address this lack, as a first contribution, this work [41] introduces the notion of proximity graph between computing nodes. If two nodes are connected in this graph, their operations must satisfy a strong consistency condition, while the operations invoked by other nodes are allowed to satisfy a weaker condition. The second contribution is the use of such a graph to provide a generic approach to the hybridization of data consistency conditions into the same system. We illustrate this approach on sequential consistency and causal consistency, and present a model in which all data operations are causally consistent, while operations by neighboring processes in the proximity graph are sequentially consistent. The third contribution of this work is the design and the proof of a distributed algorithm based on this proximity graph, which combines sequential consistency and causal consistency (the resulting condition is called fisheye consistency). In doing so this work not only extends the domain of consistency conditions, but provides a generic provably correct solution of direct relevance to modern georeplicated systems.

This work was done in collaboration with Roy Friedman (The Technion, Haifa, Israel)

6.3. Large-scale and user-centric distributed systems

6.3.1. Archiving cold data in warehouses with clustered network coding

Participants: Fabien André, Anne-Marie Kermarrec.

Modern storage systems now typically combine plain replication and erasure codes to reliably store large amount of data in datacenters. Plain replication allows a fast access to popular data, while erasure codes, e.g., Reed-Solomon codes, provide a storage-efficient alternative for archiving less popular data. Although erasure codes are now increasingly employed in real systems, they experience high overhead during maintenance, i.e., upon failures, typically requiring files to be decoded before being encoded again to repair the encoded blocks stored at the faulty node.

In this work, we proposed a novel erasure code system, tailored for networked archival systems. The efficiency of our approach relies on the joint use of random codes and a clustered placement strategy. Our repair protocol leverages network coding techniques to reduce by 50% the amount of data transferred during maintenance, by repairing several cluster files simultaneously. We demonstrated both through an analysis and extensive experimental study conducted on a public testbed that our approach significantly decreases both the bandwidth overhead during the maintenance process and the time to repair lost data. We also showed that using a non-systematic code does not impact the throughput, and comes only at the price of a higher CPU usage. Based on these results, we evaluated the impact of this higher CPU consumption on different configurations of data coldness by determining whether the cluster's network bandwidth dedicated to repair or CPU dedicated to decoding saturates first.

This work has been conducted in collaboration with Erwan Le Merrer, Nicolas Le Scouarnec, Gilles Straub (Technicolor) and A. van Kempen (Univ. Nantes) and published in ACM Eurosys 2014 [19].

6.3.2. 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 a preliminary version of the library at Middleware 2014 [47].

6.3.3. Large-scale graph processing in datacenters with bandwidth guarantees

Participants: Nitin Chiluka, Anne-Marie Kermarrec.

Recent research has shown that the performance of data-intensive applications in multi-tenant datacenters can be severely impacted by each other's network usage. Starvation for network bandwidth in such datacenters typically results in significantly longer completion times for large-scale distributed applications. To address this concern, researchers propose bandwidth guarantees for all the virtual machines (VMs) initiated by each tenant in the datacenter in order to provide a predictable performance for their applications. In our work, we focus on large-scale graph processing in such datacenters. More specifically, given k VMs with their respective bandwidth constraints and a large graph, we perform a k-way partition on the graph such that the subsequent computation of various algorithms (e.g., PageRank, graph factorization) take minimal time.

6.3.4. Scaling KNN computation over large graphs on a PC

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

Frameworks such as GraphChi and X-Stream are increasingly gaining attention for their ability to perform scalable computation on large graphs by leveraging disk and memory on a single commodity PC. These frameworks rely on the graph structure to remain the same for the entire pe- riod of computation of various algorithms such as PageRank and triangle counting. As a consequence, these frameworks are not applicable to algorithms that require the graph struc- ture to change during their computation. In this work, we focus on one such algorithm – K-Nearest Neighbors (KNN) – which is widely used in recommender systems. Our approach aims to minimize random accesses to disk as well as the amount of data loaded/unloaded from/to disk so as to better utilize the computational power, thus improving the algorithmic efficiency. The preliminary design and results of our approach appeared in Middleware 2014 [23].

6.3.5. Privacy-preserving distributed collaborative filtering

Participants: Davide Frey, Arnaud Jégou, Anne-Marie Kermarrec.

In collaboration with Antoine Boutet from the Univ. St Etienne, and Rachid Guerraoui from EPFL, we proposed a new mechanism to preserve privacy while leveraging user profiles in distributed recommender systems. Our mechanism relies on two contributions: (*i*) an original obfuscation scheme, and (*ii*) a randomized dissemination protocol. We showed that our obfuscation scheme hides the exact profiles of users without significantly decreasing their utility for recommendation. In addition, we precisely characterized the conditions that make our randomized dissemination protocol differentially private.

We compared our mechanism with a non-private as well as with a fully private alternative. We considered a real dataset from a user survey and report on simulations as well as planetlab experiments. In short, our extensive evaluation showed that our twofold mechanism provides a good trade-off between privacy and accuracy, with little overhead and high resilience.

6.3.6. Behave: Behavioral cache for web content

Participants: Davide Frey, Anne-Marie Kermarrec.

In collaboration with Mathieu Goessens, a former intern of the team, we proposed Behave: a novel approach for peer-to-peer cache-oriented applications such as CDNs. Behave relies on the principle of Behavioral Locality inspired from collaborative filtering. Users that have visited similar websites in the past will have local caches that provide interesting content for one another.

Behave exploits epidemic protocols to build overlapping communities of peers with similar interests. Peers in the same one-hop community federate their cache indexes in a Behavioral cache. Extensive simulations on a real data trace show that Behave can provide zero-hop lookup latency for about 50% of the content available in a DHT-based CDN. The results of this work were published at DAIS 2014 [26].

6.3.7. HyRec: Leveraging browsers for scalable recommenders

Participants: Davide Frey, Anne-Marie Kermarrec.

The ever-growing amount of data available on the Internet calls for personalization. Yet, the most effective personalization schemes, such as those based on collaborative filtering (CF), are notoriously resource greedy. In this work, we proposed HyRec, an online cost-effective scalable system for user-based CF personalization. HyRec offloads recommendation tasks onto the web browsers of users, while a server orchestrates the process and manages the relationships between user profiles.

We fully implemented HyRec and we extensively evaluated it on several workloads from MovieLens and Digg. Our experiments conveyed the ability of HyRec to reduce the operation costs of content providers by nearly 50% and to provide a 100-fold improvement in scalability with respect to a centralized (or cloud-based recommender approach), while preserving the quality of personalization. HyRec is also virtually transparent to users and induces only 3% of the bandwidth consumption of a p2p solution. This work was done in collaboration with Antoine Boutet from the Univ. St Etienne, as well as with Rachid Guerraoui, and Rhicheek Patra from EPFL. It resulted in a publication at Middleware 2014 [22].

6.3.8. Landmark-based similarity for p2p collaborative filtering

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.

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 Jingjing Wang, and Rachid Guerraoui.

6.3.9. Adaptation for the masses: Towards decentralized adaptation in large-scale p2p recommenders

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

Decentralized recommenders have been proposed to deliver privacy-preserving, personalized and highly scalable on-line recommendation services. Current implementations tend, however, to rely on hard-wired, mechanisms that cannot adapt. Deciding beforehand which hard-wired mechanism to use can be difficult, as the optimal choice might depend on conditions that are unknown at design time. In [27], we have proposed a framework to develop dynamically adaptive decentralized recommendation systems. Our proposal supports a decentralized 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 services.

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

6.3.10. Tight bounds for rumor spreading with vertex expansion

Participant: George Giakkoupis.

In [28] we establish an upper bound for the classic PUSH-PULL rumor spreading protocol on general graphs, in terms of the vertex expansion of the graph. We show that $O(\log^2(n)/\alpha)$ rounds suffice with high probability to spread a rumor from any single node to all n nodes, in any graph with vertex expansion at least α. This bound matches a known lower bound, and settles the natural question on the relationship between rumor spreading and vertex expansion asked by Chierichetti, Lattanzi, and Panconesi (SODA 2010). Further, some of the arguments used in the proof may be of independent interest, as they give new insights, for example, on how to choose a small set of nodes in which to plant the rumor initially, to guarantee fast rumor spreading.

6.3.11. Greedy routing in small-world networks with power-law degrees

Participant: George Giakkoupis.

In [12] we study decentralized routing in small-world networks that combine a wide variation in node degrees with a notion of spatial embedding. Specifically, we consider a variant of J. Kleinberg's grid-based small-world model in which (1) the number of long-range edges of each node is not fixed, but is drawn from a power-law probability distribution with exponent parameter $\alpha \geq 0$ and constant mean, and (2) the long-range edges are considered to be bidirectional for the purposes of routing. This model is motivated by empirical observations indicating that several real networks have degrees that follow a power-law distribution. The measured power-law exponent α for these networks is often in the range between 2 and 3. For the small-world model we consider, we show that when $2 < \alpha < 3$ the standard greedy routing algorithm, in which a node forwards the message to its neighbor that is closest to the target in the grid, finishes in an expected number of $O(\log^{\alpha-1} n \cdot \log\log n)$ steps, for any source–target pair. This is asymptotically smaller than the $O(\log^2 n)$ steps needed in Kleinberg's original model with the same average degree, and approaches $O(\log n)$ as α approaches 2. Further, we show that when $0 \leq \alpha < 2$ or $\alpha \geq 3$ the expected number of steps is $O(\log^2 n)$, while for $\alpha = 2$ it is $O(\log^{4/3} n)$. We complement these results with lower bounds that match the upper bounds within at most a $\log\log n$ factor.

This is a joint work with Pierre Fraigniaud (Inria Paris-Rocquencourt and CNRS).

6.3.12. Randomized rumor spreading in dynamic graphs

Participant: George Giakkoupis.

In [29] we consider the well-studied rumor spreading model in which nodes contact a random neighbor in each round in order to push or pull the rumor. Unlike most previous works which focus on static topologies, we look at a dynamic graph model where an adversary is allowed to rewire the connections between vertices before each round, giving rise to a sequence of graphs, $G_1, G_2, \ldots$ Our first result is a bound on the rumor spreading time in terms of the conductance of those graphs. We show that if the degree of each node does not change much during the protocol (that is, by at most a constant factor), then the spread completes within t rounds for some t such that the sum of conductances of the graphs G_1 up to G_t is $O(\log n)$. This result holds even against an adaptive adversary whose decisions in a round may depend on the set of informed vertices before the round, and implies the known tight bound with conductance for static graphs. Next we show that for the alternative expansion measure of vertex expansion, the situation is different. An adaptive adversary can delay the spread of rumor significantly even if graphs are regular and have high expansion, unlike in the static graph case where high expansion is known to guarantee fast rumor spreading. However, if the adversary is oblivious, i.e., the graph sequence is decided before the protocol begins, then we show that a bound close to the one for the static case holds for any sequence of regular graphs.

This is a joint work with Thomas Sauerwald (Univ. of Cambridge, UK) and Alexandre Stauffer (Univ. of Bath, UK).

6.3.13. Privacy-preserving dissemination in social networks and microblogs

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

Online micro-blogging services and social networks, as exemplified by Twitter and Facebook, have emerged as an important means of disseminating information quickly and at large scale. A standard mechanism in micro-blogging that allows for interesting content to reach a wider audience is that of *reposting* (i.e., *retweeting* in Twitter, or *sharing* in Facebook) of content initially posted by another user. Motivated by recent events in which users were prosecuted merely for reposting anti-government information, we present in [42] Riposte, a randomized reposting scheme that provides privacy guarantees against such charges. The idea is that if the user likes a post, Riposte will repost it only with some (carefully chosen) probability; and if the user does not like it, Riposte may still repost it with a slightly smaller probability. These probabilities are computed for each user as a function of the number of connections of the user in the network, and the extent to which the post has already reached those connections. The choice of these probabilities is based on results for branching processes, and ensures that interesting posts (liked by a large fraction of users) are likely to disseminate widely, whereas uninteresting posts (or spam) do not spread. Riposte is executed locally at the user, thus the user's opinion on the post is not communicated to the micro-blogging server. In this work, we quantify Riposte's ability to protect users in terms of differential privacy and provide analytical bounds on the dissemination of posts. We also do extensive experiments based on topologies of real networks, including Twitter, Facebook, Renren, Google+ and LiveJournal.

This work has been carried out in collaboration with Rachid Guerraoui (EPFL).

6.3.14. Adaptive streaming

Participants: Ali Gouta, Anne-Marie Kermarrec.

HTTP Adaptive Streaming (HAS) is gradually being adopted by Over The Top (OTT) content providers. In HAS, a wide range of video bitrates of the same video content are made available over the internet so that clients' players pick the video bitrate that best fit their bandwidth. Yet, this affects the performance of some major components of the video delivery chain, namely CDNs or transparent caches since several versions of the same content compete to be cached. In this context, we investigated the benefits of a Cache Friendly HAS system (CF-DASH), which aims to improve the caching efficiency in mobile networks and to sustain the quality of experience of mobile clients. We conducted our work by presenting a set of observations we made

on a large number of clients requesting HAS contents. We introduced the CF-Dash system and our testbed implementation. Finally, we evaluated CF-dash based on trace-driven simulations and testbed experiments. Our validation results are promising. Simulations on real HAS traffic show that we achieve a significant gain in hit-ratio that ranges from 15% up to 50%. This work was done in collaboration with Zied Aouini, Yannick Le Louedec and Diallo Mamadou, and was published in NOSSDAV 2014 [39].

6.3.15. Predictive capabilities of social and interest affinity for recommendations

Participant: Anne-Marie Kermarrec.

The advent of online social networks created new prediction opportunities for recommender systems: instead of relying on past rating history through the use of collaborative filtering (CF), they can leverage the social relations among users as a predictor of user tastes similarity. Alas, little effort has been put into understanding when and why (e.g., for which users and what items) the social affinity (i.e., how well connected users are in the social network) is a better predictor of user preferences than the interest affinity among them as algorithmically determined by CF, and how to better evaluate recommendations depending on, for instance, what type of users a recommendation application targets. This overlook is explained in part by the lack of a systematic collection of datasets including both the explicit social network among users and the collaborative annotated items. In this work, we conducted an extensive empirical analysis on six real-world publicly available datasets, which dissects the impact of user and item attributes, such as the density of social ties or item rating patterns, on the performance of recommendation strategies relying on either the social ties or past rating similarity. Our findings represent practical guidelines that can assist in future deployments and mixing schemes. This work has been done in collaboration with Karl Aberer and Alexandra Olteanu (EPFL Swizerland). The paper received the Best Paper Award at the WISE International Conference [18].

6.3.16. Polystyrene: The decentralized data shape that never dies

Participants: Anne-Marie Kermarrec, François Taïani.

Decentralized topology construction protocols organize nodes along a predefined topology (e.g. a torus, ring, or hypercube). Such topologies have been used in many contexts ranging from routing and storage systems, to publish-subscribe and event dissemination. Since most topologies assume no correlation between the physical location of nodes and their positions in the topology, they do not handle catastrophic failures well, in which a whole region of the topology disappears. When this occurs, the overall shape of the system typically gets lost. This is highly problematic in applications in which overlay nodes are used to map a virtual data space, be it for routing, indexing or storage. In this work [20], we propose a novel decentralized approach that maintains the initial shape of the topology even if a large (consecutive) portion of the topology fails. Our approach relies on the dynamic decoupling between physical nodes and virtual ones enabling a fast reshaping. For instance, our results show that a 51,200-node torus converges back to a full torus in only 10 rounds after 50% of the nodes have crashed. Our protocol is both simple and flexible and provides a novel form of collective survivability that goes beyond the current state of the art.

This work has been done in collaboration with Simon Bouget (ENS Rennes) and Hoel Kervadec (INSA Rennes).

6.3.17. Link-prediction for very large scale graphs using distributed graph engines

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

In this project, 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 extremely large graphs that a standard implementation of link prediction on cannot handle within the same platform. More precisely, we show that our approach 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 stand-alone machine running a non-distributed state-of-the-art solution.

6.3.18. GOSSIPKIT: A unified component framework for gossip

Participant: François Taïani.

Although the principles of gossip protocols are relatively easy to grasp, their variety can make their design and evaluation highly time consuming. This problem is compounded by the lack of a unified programming framework for gossip, which means developers cannot easily reuse, compose, or adapt existing solutions to fit their needs, and have limited opportunities to share knowledge and ideas. In [17], we have considered how component frameworks, which have been widely applied to implement middleware solutions, can facilitate the development of gossip-based systems in a way that is both generic and simple. We show how such an approach can maximise code reuse, simplify the implementation of gossip protocols, and facilitate dynamic evolution and re-deployment.

This work was done in collaboration with Shen Lin (SAP Labs) and Gordon Blair (Univ. of Lancaster, UK).

6.3.19. Towards a new model for cyber foraging

Participant: François Taïani.

Cyber foraging seeks to expand the capabilities and battery life of mobile devices by offloading intensive computations to nearby computing nodes (the surrogates). Although promising, current approaches to cyber foraging tend to impose a strict separation between the application state maintained on the mobile device, and data processed on the surrogates. In [33], we argue that this separation limits the applicability of cyber foraging, and explore how state sharing could be implemented in practice.

This work was done in collaboration with Diogo Lima and Hugo Miranda (Univ. of Lisbon, Portugal).

ATLANMOD Project-Team

6. New Results

6.1. Model Quality

Our work aims to enhance the quality of the modeling activity in the context of software engineering and language engineering. This year, this has translated in the following results:

- A systematic review [16] of all formal verification approaches targeting the quality evaluation of software models to be used as the basis for future research on the topic and as a kind of reference comparison to compare new tools with existing ones.
- A complete description of our CSP-based approach for the verification of UML/OCL models (where both the uml constructs and OCL expressions are translated into a constraint satisfaction problem) [12]
- A new test data generation approach for Model Transformations that combines partitions and constraint analysis to try to mazimize the coverage of the generates tests [29]

6.2. Model Driven approach to mobile applications development

Cross-platform and multi-device design, implementation and deployment is a barrier for today's IT solution providers, especially SME providers, due to the high cost and technical complexity of targeting development to a wide spectrum of devices, which differ in format, interaction paradigm, and software architecture. Our work aims at exploiting the modern paradigm of Model-Driven Engineering and code generation to simplify multi-device development, reducing cost and development times, so as to increase the profit of SME solution providers and at the same time reduce the price and total cost of ownership for end-customers. In [22] we defined a Platform Independent Modeling language for mobile applications. The language has been defined as a mobile extension of an OMG standard called Interaction Flow Modeling Language (IFML). The research included also the development of an Eclipse-based modeling tool for mobile apps and the first prototypes of automatic code generators.

6.3. Security

Most companies information systems are composed by heterogeneous components responsible of hosting, creating or manipulating critical information for the day-to-day operation of the company. Securing this information is therefore one of their main concerns, more particularly specifying Access Control (AC) policies. However, the task of implementing an AC security policy (sometimes relying on several mechanisms) remains complex and error prone as it requires knowing low level and vendor-specific facilities. In this context, discovering and understanding which security policies are actually being enforced by the Information System (IS) becomes critical. Thus, the main challenge consists in bridging the gap between the vendor-dependent security features and a higher-level representation. This representation has to express the policies by abstracting from the specificities of the system components, allowing security experts to better understand the policy and to implement all related evolution, refactoring and manipulation operations in a reusable way.

In 2014, we have presented a Ph.D. thesis tackling the aforementioned problems. It proposes a model-driven automatic reverse engineering mechanism capable of analyzing deployed security aspects of components (e.g. concrete firewall configurations) to derive the abstract model (e.g. network security global policy) that is actually enforced. Once the model is obtained, it can be reconciled with the expected security directives, to check its compliance, can be queried to test consistency or used in a process of forward engineering to generate validated security configurations. This work also provides the first steps towards the integration of the diverse security policies extracted from the subsystems composing a complex Information System in a global security representation.

6.4. Model-Driven Document Engineering

As a result of a long-term collaboration of one of the AtlanMod team members with the ISSI research group at the Universitat Politècnica de València, we have participated in the publication of several works on the area of the Document Engineering. In this research line, we have applied the MDE methods and tools to the product-line-based generation of customized documents resulting in the so-called DPL methodology [0]. The Document Product Lines (DPL) approach, which we throughly describe in a journal publication [17], provides a framework for variable content document generation that follows an alternate path to the traditional variable document generation. DPL has been created with a twofold goal: first, to make creating variable content documents available to non-experts by including a domain engineering process previous to the document generation itself; and secondly, to enforce content reuse at domain level.

DPLFW is the main tool supporting the DPL methodology, and in the demonstrations track of the MODELS conference we showed all its capabilities. In addition to these contributions, we have published several works demonstrating the applicability of the DPL–DPLFW tandem in different domains, such as the development of executable emergency plans in crisis managent contexts [25], the development of learning objects in the e-learning field [32] and the generation of customized documents in e-Government solutions [33].

6.5. Reverse Engineering and Evolution

Model Driven Reverse Engineering (MDRE), and its applications such as software modernization, is a discipline in which model-driven development (MDD) techniques are used to treat legacy systems. During this year, Atlanmod has continued working actively on this research area. 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. At conceptual-level, the MoDisco Model Discovery + Model Understanding overall two-step approach [11] has been published and promoted as an important part of the ARTIST migration methodology and process [18]. At tooling-level, several (MoDisco-based) model discovery components from Java and SQL have been developed and made available as part of the official ARTIST OS Release [0]. Directly related to some of these components, a promising work has been initiated on studying deeper the discovery of behavioral aspects of software and dealing with their further understanding based on the OMG FUML standard combined with different modeling techniques (transformation, slicing, etc.). Complementary work has also been performed in the context of the TEAP FUI project finishing by the end of this year. It concerns the related problem of data federation from heterogeneous sources in the domain of Enterprise Architecture. This has notably resulted in a prototype called EMF Views that can be practically used in such reverse engineering scenarios [36] and also in other cases to be further explored (cf. the MoNoGe FUI project dealing with (meta)model extension).
- In a web context, in a previous work we shown how to discover the schema which is implicit in JSON data. This year we built on that contribution to study how schemas coming from different JSON-based web APIs can be composed [24]. Thus, we presented an approach able to identify composition links between schemas of different APIs. This composition information plus the API schemas are used to render a graph where paths represent API compositions and are used to easily identify how to compose the APIs. For instance, we illustrated one application based on generating sequence diagrams from graph paths, where the diagram includes the API calls (and their corresponding parameters) that web developers have to perform in order to compose one or more APIs.
- In the context of our work around DSLs, we have been working on facilitating the definition of DSLs from existing APIs. Sometimes library developers prefer to provide their users with a DSL, instead of (or in addition to) an API. APIs and DSLs can be seen as alternative methods to access the library

[0] http://dpl.dsic.upv.es, only in Spanish
[0] http://www.artist-project.eu/tools-of-toolbox/193

functionalities, and are characterized by specific advantages. We therefore proposed a method to automatically analyze an existing object-oriented API and generate a DSL out of it. Our approach leverages on model-driven techniques to analyze and represent APIs at high-level of abstraction (i.e., as metamodels) which are later used to automatically generate the DSL components and the corresponding tooling, including parser, compiler and development environment. Developers can influence the DSL generation by editing the model-based API representation and by specifying design choices about the structure of the DSL to generate. A proof-of-concept implementation of the method has been developed, called *DSLit*, that is able to analyze Java APIs and generate textual DSLs.

- On the evolution side, we have been working on an approach to automatically resynchronize code-generation artefacts (in particular, model-to-text transformations) after changes on the target platform [28]

6.6. 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 [21], we lead the first open-set benchmark gathered from real-world cases to stress scalability issues in model transformation and query engines. This benchmark suite has been made public with a twofold goal: (i) to provide a reference benchmark suite to both the industry and the research community that can be used to compare and evaluate different technologies that may fulfill their needs; and (ii) to motivate the MDE community to be part of its development by allowing them to extend and contribute with additional cases not covered by the initial set.

On the other hand, we introduce Neo4EMF [20], a NoSQL database persistence framework based on Neo4j [0]. Neo4EMF provides light-weight on-demand loading and storage facilities for handling very large models. Additionally, we also show that Neo4EMF can handle the creation of very-large models without performing periodical saves manually.

In this paper [31], we argue that fUML may be leveraged to address the well-known interoperability issue between tools from different modeling platforms. This is done by providing a common execution language and by abstracting modeling frameworks into generic actions that perform elementary operations on models. User models can not only benefit from a unified execution semantics, but also modeling tools can benefit too. As a proof of concept, we show [37] how it can be applied to model transformation engines, in particular ATL. To this end, an prototype compiler from ATL to fUML has been built.

In [19], we present a model-based approach to define a dynamic oracle for checking global properties on distributed software. Our objective is to abstract relevant aspects of such systems into models by gathering data from different nodes and building a global view of the system, where properties are validated. These models are updated at runtime, by monitoring the corresponding distributed system. This process requires a distributed test architecture and tools for representing and validating global properties. To evaluate the ability of our approach, a real-scale experimental validation has been conducted.

[0] http://www.neo4j.org

CIDRE Project-Team

6. New Results

6.1. Highlights of the Year

The supervision of distributed system relies heavily on correlation mechanisms that are responsible for collecting alerts coming from sensors and detecting complex scenarios in the flow of alerts. The problem is that it requires to write complex correlation rules. The work we have performed proposes a technique to generate semi-automatically such correlation rules. It describes a process that uses an attack tree and a representation of the system as inputs, and generate a correlation tree that can be translated in an alert correlation description language. This work received the best paper award of SAR-SSI 2014 [50].

One approach to protect the privacy of users in personalized recommendation systems is to publish a sanitized version of the profile of the user by relying a non-interactive mechanism compliant with the concept of differential privacy. In a joint work with Raghavendran Balu and Teddy Furon (LinkMedia Inria team), we have consider two existing schemes offering a differentially private representation of profiles: BLIP (BLoom-and-flIP) and JLT (Johnson-Lindenstrauss Transform). For assessing their security levels, we play the role of an adversary aiming at reconstructing a user profile. To realize this, we design two inference attacks named single and joint decoding. The first inference attack tests the presence of a single item in the profile, and is iterated independently for each possible item of the item set. In contrast, the second inference attack aims at deciding whether a particular subset of items is likely to be in the user profile. This attack is tested on all the possible subsets of items. Our contributions are a theoretical analysis and practical implementations of both attacks tested on datasets composed of real user profiles revealing that joint decoding is the most powerful attack. This also gives useful insights on the setting the differential privacy parameter ϵ. This work has received the best student paper award at the conference ESORICS 2014.
BEST PAPERS AWARDS :
[27] **European Symposium on Research in Computer Security**. R. BALU, T. FURON, S. GAMBS.

6.2. Intrusion Detection

6.2.1. Intrusion detection based on an analysis of information flow control

In 2014, Laurent Georget has started his PhD thesis in the team, working on a subject related to the analysis of information flow control at the kernel level. The goal of his PhD thesis is to propose a formal semantics of the system calls for a real operating systems (namely Linux). This semantics will provide insights about these system calls in terms of information flow. This work will help us to test in a more systematic and efficient way, our reference implementation of a information monitor at the kernel level (Blare).

Blare allows monitoring information flow and identifies the flows that do not conform to a security policy that has been previously defined. Please notice that any explicit flows between OS objects (sockets, files, etc.) are monitored and that in consequence hidden channel attacks cannot be detected by this approach.

We have already developed a dedicated test framework for this software. However, each test written by the developer must be accompanied with the possible results in terms of information flows. The framework simply compares the effective result with the set of expected results. A test passes when the effective result belongs to the set of expected results, and fails otherwise. However, this strategy has turned to be less intuitive than expected. Some system calls must be tested by using several processes operating concurrently. In these cases, the scheduling of processes can produce many different scenarios that will translate quite differently in terms of information flows. To be more confident in our implementation, we really need a stronger and more formal path. The PhD thesis of Laurent Georget is trying to bridge the gap between Blare implementation and the interpretation of the results obtained by running the information flow monitor.

6.2.2. Malware characterization through information flow monitoring

Monitoring information flows consists in observing how pieces of information are disseminated in a given environment. At system level, it consists in intercepting actions performed by an application to deduce how the application disseminates information within the entire operating system. We have propose a new approach to classify and later detect applications infected by malware based on the way they disseminate their own data within an operating system. For this purpose, we first introduce a data-structure named System Flow Graph [thèse Rado to ref.] that offers a compact representation of how pieces of data flow inside a system. A system flow graph describes the external behavior of an application during one execution. Its construction requires no knowledge about the inner working of the application. The graph is built using Blare as an information-flow monitor and more precisely its produced log. We have presented in [25] how these graphs reveal helpful to understand malware behavior and thus why it can help an expert to give a diagnosis in case of intrusion.

6.2.3. Terminating-insensitive non-interference verification based on information flow control

In 2010-2011, we started an informal collaboration with colleagues from CEA LIST laboratory. This collaboration has turned into a reality by the funding of a PhD student (Mounir Assaf). This PhD thesis is about the verification of security properties of programs written in an imperative language with pointer aliasing (a subset of C language) by techniques borrowed from the domain of static analysis. One of the property of interest for the security field is called terminating-insensitive non-interference. Briefly speaking, when verified by a program, this property ensures that the content of any secret variable can not leak into public ones (for any terminating execution). However, this property is too strict in the sense that a large number of programs although perfectly secure are rejected by classical analyzers. Finally in 2014, Mounir Assaf enhanced his previous work on static analysis by introducing a method permitting to quantify information leakage in a C program. This approach requires a theoretical definition of the quantification of information flow leakage and is very promising.

6.2.4. Visualization of security events

The first part of this year was dedicated to tune a working prototype of ELVIS [38] in order to perform field trials with our partner DGA-MI. The prototype was largely well accepted. We were invited by the DGA-MI to present a poster in the Forum DGA Innovation 2014. We will also present ELVIS during the FIC 2014 in Lille on the Pôle Cyber-Défense area.

However, ELVIS also exhibited some limitations of our approach in the way multiple datasets are handled together. We therefore went for a new cycle of research whose objective is to enhance ELVIS in two ways: first to handle multiple datasets at the same time, and second to improve interactions so as to better fit with the processes in forensics. The results of our research lead to CORGI (Combination, Organization and Reconstruction through Graphical Interactions) [39] which was presented at VizSec 2014 (part of Vis 2014). CORGI improves ELVIS by introducing the concepts of *values of interest* that consist in interesting values found by an analyst and that can be used later to search and filter in the other datasets. They are an intuitive and efficient way to link various datasets while the analyst performs its tasks. An early prototype has been developed.

6.2.5. Control flow integrity

In [40] we have studied physical attacks that could disturb the normal execution of an embedded program of a smartcard. Such attacks can be performed using laser beams, electromagnetic glitches and can corrupt the flow of information or change the control flow of the program. We have studied the particular case of the control flow and we have developed software countermeasures that increase the robustness of the control flow. These countermeasures do not require any additional software or hardware external components which is useful for devices like smartcards whose architecture cannot be modified. The developed countermeasures have been validated with the help of the VIS model checker in order to verify that they do not disturb the original execution of the code.

6.2.6. 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 [50], [47], [36], we focus on the transformation process that allows to translate the description of a complex attack scenario into correlation rules. 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.

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 (occurrence 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 current work aims at identifying the possible errors tolerated by the compacted automaton.

6.3. Privacy

6.3.1. Privacy in location-based services

With the advent of GPS-equipped devices, a massive amount of location data is being collected, raising the issue of the privacy risks incurred by the individuals whose movements are recorded. In [17], we focus on a specific inference attack called the de-anonymization attack, by which an adversary tries to infer the identity of a particular individual behind a set of mobility traces. More specifically, we propose an implementation of this attack based on a mobility model called Mobility Markov Chain (MMC). A MMC is built out from the mobility traces observed during the training phase and is used to perform the attack during the testing phase. We design several distance metrics quantifying the closeness between two MMCs and combine these distances to build de-anonymizers that can re-identify users in an anonymized geolocated dataset. Experiments conducted on real datasets demonstrate that the attack is both accurate and resilient to sanitization mechanisms such as downsampling.

One example of a location-based services is dynamic carpooling (also known as instant or ad-hoc ridesharing), which is a service that arranges one-time shared rides on very short notice. This type of carpooling generally makes use of three recent technological advances: (i) navigation devices to determine a route and arrange the shared ride; (ii) smartphones for a traveller to request a ride from wherever she happens to be; and (iii) social networks to establish trust between drivers and passengers. However, the mobiquitous environment in which dynamic carpooling is expected to operate raises several privacy issues. Among all the personal identifiable information, learning the location of an individual is one of the greatest threats against her privacy. For instance, the spatio-temporal data of an individual can be used to infer the location of her home and workplace, to trace her movements and habits, to learn information about her centre of interests or even to detect a change from her usual behavior. Therefore, preserving location privacy is a major issue to be able to leverage the possibilities offered by dynamic carpooling. In a joint work with researchers from LAAS-CNRS

[16], we have propose to follow the privacy-by-design approach by integrating the privacy aspect in the design of dynamic carpooling, henceforth increasing its public (and political) acceptability and trust.

A secure location-based service requires that a mobile user certifies his position before gaining access to a resource. Currently, most of the existing solutions addressing this issue assume a trusted third party that can vouch for the position claimed by a user. However, as computation and communication capacities become ubiquitous with the large scale adoption of smartphones by individuals, these resources can be leverage on to solve this issue in a collaborative and private manner. More precisely together with researchers from LAAS-CNRS, we introduce PROPS, for Privacy-Preserving lOcation Proof System, which allows users to generate proofs of location in a private and distributed way using neighboring nodes as witnesses [35]. PROPS provides security properties such as unforgeability and non-transferability of the proofs, as well as resistance to classical localization attacks.

One of the fundamental building block to construct a location proof system such as PROPS is a distance-bounding protocol. More precisely, in distance-bounding authentication protocols a verifier assesses that a prover is (1) legitimate and (2) in the verifier's proximity. Proximity checking is done by running time-critical exchanges between both parties. This enables the verifier to detect relay attacks (also called mafia fraud). While most distance-bounding protocols offer resistance to mafia, distance, and impersonation attacks, only few protect the privacy of the authenticating prover. One exception is the protocol due to Hermans, Peeters, and Onete, which offers prover untraceability with respect to a Man-in-the-Middle adversary. However in this protocol as well as in all other distance-bounding protocols, any legitimate verifier can identify, and thus track, the prover. In order to counter the threats of possible corruption or data leakage from verifiers, together with Jean-Marc Robert (ETS, Montréal) we propose a distance-bounding protocol providing strong prover privacy with respect to the verifier and deniability with respect to a centralized back-end server managing prover creation and revocation [33]. In particular, we first formalize the notion of prover anonymity, which guarantees that even verifiers cannot trace provers, and deniability, which allows provers to deny that they were authenticated by a verifier. Finally, we prove that our protocol achieves these strong guarantees.

A particular class of relay attacks against distance-bounding protocols is called terrorist fraud in which a distant malicious prover colludes with an attacker located in a verier's proximity when authenticating. Existing distance-bounding protocols resisting such attacks are designed to be lightweight and thus symmetric, relying on a secret shared by the prover and the verifier. Recently, several asymmetric distance-bounding protocols were proposed by Gambs, Onete and Robert as well as by Hermans, Peter and Onete, but they fail to thwart terrorist fraud. One earlier asymmetric protocol aiming to be terrorist-fraud resistant is the DBPK-Log protocol due to Bussard and Bagga, which was unfortunately recently proven to achieve neither distance- nor terrorist-fraud resistance. In this work, we build on some ideas of the DBPK-Log scheme and propose a novel distance-bounding protocol resistant to terrorist fraud that does not require the pre-existence of a shared secret between the prover and the verifier [32]. Our construction, denoted as VSSDB (for Verifiable Secret Sharing and Distance-Bounding Protocol) relies on a veriable secret sharing scheme and on the concept of modes, which we introduce as a novel element to complement fast-round challenges in order to improve security. We prove that VSSDB achieves terrorist-fraud resistance in a relaxed security model called KeyTF-security, which we also present in this paper.

6.3.2. Equity in privacy-enhanced social networks

In [46], we have examined a novel issue in the field of policy conflict resolution, and applied it to privacy policy management in distributed social networking systems. We accepted as a starting point that in a privacy-enhanced social network, when a user publishes a document (e.g., a picture), any user referenced in this document (e.g., people tagged in pictures) should be entitled to issue a privacy policy over this document. In this case, when a given user tries to access a given document, multiple users may issue multiple access control decisions (or rulings), possibly resulting in a normative conflict. Quite a number of strategies are available for the resolution of such conflicts, the most common one being the "deny strategy", allowing any ruling denying access to the resource to take precedence over others. This is usually considered a "secure" way of dealing with access control. However, with this strategy as with many others, it is possible for a user to design her policy in a way that systematically prevents other users from interacting in a normal way, while allowing herself to

potentially benefit from other people's more flexible policies. This may leads to unfair situations, in which some users take advantage of the systems while others' experience is damaged. This is particularly an issue in social networking applications, in which information sharing is a core feature and access restrictions, while necessary to protect intimacy, can sometimes be considered aggressive.

To address this particular trade-off between privacy and usability, we have introduced the notion of equity in such scenarios, a situation being equitable when all involved users have seen their policy enforced or violated in the same proportion over past interactions. We have designed a conflict resolution algorithm aimed at improving this equity in our social networking scenario, and evaluated its impact by measuring Gini coefficients (an indicator commonly used by economists to measure the distribution of wealth in a population) over the distribution of enforcement proportions in the population of users. With respect to this criterion, it actually proved more efficient than other strategies. Following these positive results, we have recently taken steps towards a formalization and generalization of this intuitive concept of equity and the design of systematic tools to evaluate and compare the impact of any conflict resolution strategy over various possible flavors of the notion.

6.3.3. Private mobile services

The development of NFC-enabled smartphones has paved the way to new applications such as mobile payment (m-payment) and mobile ticketing (m-ticketing). However, often the privacy of users of such services is either not taken into account or based on simple pseudonyms, which does not offer strong privacy properties such as the unlinkability of transactions and minimal information leakage. In [48], [15], we introduce a lightweight privacy-preserving contactless transport service that uses the SIM card as a secure element. Our implementation of this service uses a group signature protocol in which costly cryptographic operations are delegated to the mobile phone. We have also conducted an interdisciplinary study with researchers from social sciences to analyze the media coverage in the modern public space on the topic of privacy with respect to mobile technologies [29]. Despite the difficulties highlighted by these studies, we argue that research efforts should support the emergence of mobile services that respect users' privacy as well as the development of a digital culture of privacy.

6.3.4. Architectures for privacy

In the current architecture of the Internet, there is a strong asymmetry in terms of power between the entities that gather and process personal data (e.g., major Internet companies, telecom operators, cloud providers, ...) and the individuals from which this personal data is issued. In particular, individuals have no choice but to blindly trust that these entities will respect their privacy and protect their personal data. In a position paper [34] in a collaboration with researchers from the Université de Montréal and Aarhus University, we propose an utopian crypto-democracy model based on existing scientific achievements from the field of cryptography. More precisely, our main objective is to show that cryptographic primitives, including in particular secure multiparty computation, offer a practical solution to protect privacy while minimizing the trust assumptions. In the crypto-democracy envisioned, individuals do not have to trust a single physical entity with their personal data but rather their data is distributed among several institutions. Together these institutions form a virtual entity called the Trustworthy that is responsible for the storage of this data but which can also compute on it (provided first that all the institutions agree on this). Finally, we also propose a realistic proof-of-concept of the Trustworthy, in which the roles of institutions are played by universities. This proof-of-concept would have an important impact in demonstrating the possibilities offered by the crypto-democracy paradigm.

Active fingerprinting schemes were originally invented to deter malicious users from illegally releasing an item, such as a movie or an image. To achieve this, each time an item is released, a different fingerprint is embedded in it. If the fingerprint is created from an anti-collusion code, the fingerprinting scheme can trace colluding buyers who forge fake copies of the item using their own legitimate copies. Charpentier, Fontaine, Furon and Cox were the first to propose an asymmetric fingerprinting scheme based on Tardos codes, the most efficient anti-collusion codes known to this day. However, their work focuses on security but does not preserve the privacy of buyers. To address this issue, we introduce the first privacy-preserving asymmetric fingerprinting protocol based on Tardos codes [30]. This protocol is optimal with respect traitor tracing. We

also formally define the properties of correctness, anti-framing, traitor tracing, as well as buyer- and item-unlinkability. Finally, we prove that our protocol achieves these properties and give exact bounds for each of them.

6.3.5. Privacy and web services

We have proposed [61] a new model of security policy based for a first part on our previous works in information flow policy and for a second part on a model of Myers and Liskov. This new model of information flow serves web services security and allows a user to precisely define where its own sensitive pieces of data are allowed to flow through the definition of an information flow policy. A novel feature of such policy is that they can be dynamically updated, which is fundamental in the context of web services that allow the dynamic discovery of services. We have also presented an implementation of this model in a web services orchestration in BPEL (Business Process Execution Language).

6.3.6. Privacy-preserving ad-hoc routing

Last year, we have proposed NoName, a privacy-preserving ad-hoc routing protocol. Based on trapdoor, virtual switching and partially disjoint multipaths using Bloom filter, NoName ensures the anonymity of the source, of the destination and of intermediate nodes. It also ensures unlinkability between source and message and between destination and message. Since then, we have demonstrated that colluding attackers analyzing Bloom filters can locate the origin node of routes requests messages. Thus, Noname, like ARMR, another privacy-preserving ad-hoc routing protocol using also Bloom filter, do not prevent the localization of the source. We have developed a cryptographic primitive called fuzzy cryptographic Bloom filter that offers the same functions as Bloom filters (in our case, preventing routing loops) while preventing localization of the source of route request messages.

6.4. Trust

Digital reputation mechanisms have indeed emerged as a promising approach to cope with the specificities of large scale and dynamic systems. Similarly to real world reputation, a digital reputation mechanism expresses a collective opinion about a target user based on aggregated feedback about his past behavior. The resulting reputation score is usually a mathematical object (*e.g.* a number or a percentage). It is used to help entities in deciding whether an interaction with a target user should be considered. Digital reputation mechanisms are thus a powerful tool to incite users to behave trustworthily. Indeed, a user who behaves correctly improves his reputation score, encouraging more users to interact with him. In contrast, misbehaving users have lower reputation scores, which makes it harder for them to interact with other users. To be useful, a reputation mechanism must itself be accurate against adversarial behaviors. Indeed, a user may attack the mechanism to increase his own reputation score or to reduce the reputation of a competitor. A user may also free-ride the mechanism and estimate the reputation of other users without providing his own feedback. From what has been said, it should be clear that reputation is beneficial in order to reduce the potential risk of communicating with almost or completely unknown entities. Unfortunately, the user privacy may easily be jeopardized by reputation mechanisms, which is clearly a strong argument to compromise the use of such a mechanism. Indeed, by collecting and aggregating user feedback, or by simply interacting with someone, reputation systems can be easily manipulated in order to deduce user profiles. Thus preserving user privacy while computing robust reputation is a real and important issue that we address in our work [51]. Specifically, our proposal aims at enhancing signatures of reputation mechanism proposed by Bethencourt and his colleagues in 2010 by handling negative votes. Taking into account negative votes implies major modifications with respect to the implementation of the mechanism. Specifically, in the mechanism of Bethencourt and co-authors, service providers locally store votes cast at the end of their interaction with their clients, and compute their reputation score by aggregating the received votes. In particular, they can keep only a subset of them, which clearly makes negative votes useless. We propose to improve upon this solution by guaranteeing that negative votes are taken into account. This is achieved by making both reputation scores and votes of service providers publicly available in order to prevent anyone from modifying or hiding them. Our proposition accomplishes this without jeopardizing the privacy of clients.

6.5. Other topics related to security and distributed computing

6.5.1. Network monitoring and fault detection

Monitoring a system consists in collecting and analyzing relevant information provided by the monitored devices, so as to be continuously aware of the system state (situational awareness). However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. We propose in [23], [42] to address this issue by pushing the monitoring task at the edge of the network. We present a peer-to-peer based architecture, which enables nodes to adaptively and efficiently self-organize according to their "health" indicators. By exploiting both temporal and spatial correlations that exist between a device and its vicinity, our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator. We show that the end-to-end detection process, *i.e.*, from the local detection to the management operator reporting, requires a logarithmic number of messages in the size of the network.

6.5.2. Secure data deduplication scheme

Data grows at the impressive rate of 50% per year, and 75% of the digital world is a copy [0]. Although keeping multiple copies of data is necessary to guarantee their availability and long term durability, in many situations the amount of data redundancy is immoderate. By keeping a single copy of repeated data, data deduplication is considered as one of the most promising solutions to reduce the storage costs, and improve users experience by saving network bandwidth and reducing backup time. However, this solution must now solve many security issues to be completely satisfying. In this paper we target the attacks from malicious clients that are based on the manipulation of data identifiers and those based on backup time and network traffic observation. In [43], we have presented a deduplication scheme mixing an intra-and an inter-user deduplication in order to build a storage system that is secure against the aforementioned type of attacks by controlling the correspondence between files and their identifiers, and making the inter-user deduplication unnoticeable to clients using deduplication proxies. Our method provides global storage space savings, per-client bandwidth network savings between clients and deduplication proxies, and global network bandwidth savings between deduplication proxies and the storage server. The evaluation of our solution compared to a classic system shows that the overhead introduced by our scheme is mostly due to data encryption which is necessary to ensure confidentiality. This work relies on Mistore [44], [45], a distributed storage system aiming at guaranteeing data availability, durability, low access latency by leveraging the Digital Subscriber Line infrastructure of an ISP. Mistore uses the available storage resources of a large number of home gateways and points of presence for content storage and caching facilities reducing the role of the data center to a load balancer. Mistore also targets data consistency by providing multiple types of consistency criteria on content and a versioning system allowing users to get access to any prior versions of their contents.

6.5.3. Metrics estimation on very large data streams

In [12], we consider the setting of large scale distributed systems, in which each node needs to quickly process a huge amount of data received in the form of a stream that may have been tampered with by an adversary (*i.e.*, data items ordering can be manipulated by an oblivious adversary). In this situation, a fundamental problem is how to detect and quantify the amount of work performed by the adversary. To address this issue, we propose AnKLe (for Attack-tolerant eNhanced Kullback- Leibler divergence Estimator), a novel algorithm for estimating the KL divergence of an observed stream compared to the expected one. AnKLe combines sampling techniques and information-theoretic methods. It is very efficient, both in terms of space and time complexities, and requires only a single pass over the data stream. Experimental results show that the estimation provided by AnKLe remains accurate even for different adversarial settings for which the quality of other methods dramatically decreases. Considering n as the number of distinct data items in a stream, we show that AnKLe is an (ε, δ)-approximation algorithm with a space complexity sublinear in the size of the domain value from which data items are drawn and the maximal stream length.

[0] The digital universe decade. Are you ready? John Gantz and David Reinsel, IDC information, may 2010.

We go a step further by proposing in [22] a metric, called codeviation, that allows to evaluate the correlation between distributed streams. 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.

6.5.4. Robustness analysis of large scale distributed systems

In the continuation of [59] which proposed an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, we analyze in [13], the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. When the number of Markov chains goes to infinity, we analyze the asymptotic behavior of the system for an arbitrary probability mass function governing the competition. We give conditions for the existence of the asymptotic distribution and we show how these results apply to cluster-based distributed systems when the competition between the Markov chains is handled by using a geometric distribution.

6.5.5. 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. 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 [53].

6.5.6. Randomized message-passing test-and-set

In [56], 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 $p < n$ of processes in which 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.

6.5.7. Agreement problems in unreliable systems

In [18], we consider the problem of approximate consensus in mobile ad-hoc networks in the presence of Byzantine nodes. Each node begins to participate by providing a real number called its initial value. Eventually all correct nodes must obtain final values that are different from each other within a maximum value previously defined (convergence property) and must be in the range of initial values proposed by the correct nodes (validity property). Due to nodes' mobility, the topology is dynamic and unpredictable. We propose an approximate Byzantine consensus protocol which is based on the linear iteration method. Each node repeatedly executes rounds. During a round, a node moves to a new location, broadcasts its current value, gathers values from its neighbors, and possibly updates its value. In our protocol, nodes are allowed to collect information during several consecutive rounds: thus moving gives them the opportunity to gather progressively enough values. An integer parameter R_c is used to define the maximal number of rounds during which values can be gathered and stored while waiting to be used. A novel sufficient and necessary condition guarantees the final convergence of the consensus protocol. At each stage of the computation, a single correct node is concerned by the requirement expressed by this new condition (the condition is not universal as it is the case in all previous related works). Moreover the condition considers both the topology and the values proposed by correct nodes. If less than one third of the nodes are faulty, the condition can be satisfied. We are working on mobility scenarios (random trajectories, predefined trajectories, meeting points) to assert that the condition can be satisfied for reasonable values of R_c. In [41], we extend the above protocol to solve the problem of clock synchronization in mobile ad-hoc networks.

In [20], we investigate the use of agreement protocols to develop transactional mobile agents. Mobile devices are now equipped with multiple sensors and networking capabilities. They can gather information about their surrounding environment and interact both with nearby nodes, using a dynamic and self-configurable ad-hoc network, and with distant nodes via the Internet. While the concept of mobile agent is appropriate to explore the ad-hoc network and autonomously discover service providers, it is not suitable for the implementation of strong distributed synchronization mechanisms. Moreover, the termination of a task assigned to an agent may be compromised if the persistence of the agent itself is not ensured. In the case of a transactional mobile agent, we identify two services, Availability of the Sources and Atomic Commit, that can be supplied by more powerful entities located in a cloud. We propose a solution in which these two services are provided in a reliable and homogeneous way. To guarantee reliability, the proposed solution relies on a single agreement protocol that orders continuously all the new actions whatever the related transaction and service.

COAST Team

5. New Results

5.1. An authentication/authorization framework for federated environments

Participants: Ahmed Bouchami, Olivier Perrin.

Collaborative environments have put an enormous challenge on the security of information processing systems used to manage them. In the context of the Open PaaS project, we worked on a decentralised hybrid framework for managing access control designed for support of these environments. In our proposal, we manage thress dimensions: the authentication, the access control, and the governance of the security.

Our authentication framework supports an interoperable authentication, a combination of RBAC, XACML for decentralized multiple administration (authorization). Both identities and resources are federated: the former are controlled by PaaS Federated Security Modules, while the later are by a PaaS Federated Security Modules. This work has been presented in the I-ESA conference ([10]).

We have also proposed a formal cloud-based authorization framework. We have considered trust to be a dynamic attribute to facilitate authorization decisions and have proposed models to handle different qualitative, quantitative and periodicity based temporal constraints. Further, we have presented an architecture for policies evaluation in the cloud. We presented our model in the CollaborateCom conference [17]. The model relies on a formal event-calculus based approach. We have introduced an architecture that considers different levels at which authorization policies can be specified and decisions can be taken and combines user level policies with the enterprise policies, and it considers real-time and dynamic environment changes (context), supports timed delegation, and the computation and specification of attributes based on trust. An implementation has been integrated in the Open PaaS platform.

A third aspect deals with the governance of the security aspects (mainly authorization). In this part, we have proposed to audit the various accesses to the resources, and we have proposed a model which is able to lower/raise the trust level of a member of the federated community.

During this year, we have also implemented and integrated the framework in the Open PaaS prototype, and all the code is now accessible in the repository of the project. The integration is done, and the other components of the project are now using the authentication/authorization component.

5.2. Experimental user studies for collaborative editing

Participants: Mehdi Ahmed-Nacer, François Charoy, Claudia-Lavinia Ignat, Gérald Oster, Pascal Urso.

With several tools to support collaborative editing such as Google Drive and Etherpad, the practice of colaborative editing is increasingly common, e.g., group note taking during meetings and conferences, and brainstorming activities. While collaborative editing tools meet technical goals, the requirements for group performance are unclear. One system property of general interest is delay between a modification of a user is performed and this modification is visible to the other users. This delay can be caused by different reasons such as network delay due to physical communication technology, the complexity of various algorithms for ensuring consistency and the type of underlying architectures. No prior work questioned the maximum acceptable delay for real-time collaboration or the efficacy of compensatory strategies.

In [14] we studied the effect of delay on group performance on an artificial collaborative editing task where a group of four participants located the release dates for an alphabetized list of movies and re-sorted the list in chronological order. The experiment was performed with eighty users. We measured sorting accuracy based on the insertion sort algorithm, average time per entry, strategies (tightly coupled or loosely coupled task decomposition of the task) and chat behavior between users. We found out that delay slows down participants which decrements the outcome metric of sorting accuracy. Tightly coupled task decomposition enhances outcome at minimal delay, but participants slow down with higher delays. A loosely coupled task decomposition at the beginning leaves a poorly coordinated tightly coupled sorting at the end, requiring more coordination as delay increases.

In asynchronous collaborative editing, such as version control, the main feature to allow collaboration is the merge feature. However, software merging is a time-consuming and error-prone activity, and if a merge feature return results with too many conflicts and errors, this activity becomes even more difficult. To help developers, several algorithms have been proposed to improve the automation of merge tools. These algorithms aim at minimising conflict situations and therefore improving the productivity of the development team, however no general framework is proposed to evaluated and compare their result.

In [9] we propose a methodology to measure the effort required to use the result of a given merge tool. We employ the large number of publicly available open-source development histories to automatically compute this measure and evaluate the quality of the merging tools results. We use the simple idea that these histories contains both the concurrent modifications and their merge results as approved by the developers. Through a study of six open-source repositories totalling more than 2.5 millions lines of code, we show meaningful comparison results between merge algorithms and how to use the results to improve them.

5.3. Optimization and security of business processes in SaaS contexts

Participants: Claude Godart, Elio Goettelmann, Samir Youcef.

Globalization and the increase of competitive pressures created the need for agility in business processes, including the ability to outsource, offshore, to take opportunity of the cloud, or otherwise distribute its once-centralized business processes or parts thereof. While hampered thus far by limited infrastructure capabilities, the increase in bandwidth and connectivity and decrease in communication cost have removed these limits. This is even more true with the advent of cloud, particularly in its “Service as a software” dimension. To adapt to such a context, there is a growing need for the ability to fragment one’s business processes in an agile manner, and be able to distribute and wire these fragments so that their combined execution recreates the function of the original process. Our work is focused on solving some of the core challenges resulting from the need to dynamically restructure enterprise interactions. Restructuring such interactions corresponds to the fragmentation of intra- and inter-enterprise business process models. It describes how to identify, create, and execute process fragments without loosing the operational semantics of the original process models. In addition, this fragmentation is complicated by the constraints of quality of service, in particular the execution time and the cost, and of security, especially privacy. During the year, we consider this problem at two levels: the design of privacy-aware process models, and the optimization of process schedules. We developed a methodology to integrate privacy concerns in the design of a business process before distribution in the cloud [11]. Based on a risk analysis, the result of the design is a set of process (re)modeling actions, a set of constraints on process fragments assignments to clouds, and a set of constraints for cloud selection based on cloud properties [12].

CTRL-A Exploratory Action

6. New Results

6.1. Highlights of the Year

We have been invited to participate to the organization of events, which highlight our active presence in the scientific life in the two domains which we are bridging :

- autonomic computing: Eric Rutten is PC member, as well as workshops chair, of the 12th IEEE International Conference on Autonomic Computing, ICAC 2015 (http://icac2015.imag.fr/), and PC co-chair of the 3rd IEEE International Conference on Cloud and Autonomic Computing, CAC 2015 (http://autonomic-conference.org/), the two major conferences on the topic.
- control: Eric Rutten is organizer of a special session on discrete control for computing at the 12th IFAC - IEEE International Workshop on Discrete Event Systems, WODES 2014 (http://wodes2014.lurpa.ens-cachan.fr/), the main conference specialized in Discrete Event Systems, ; he is on the IFAC Technical Committee 1.3 on Discrete Event and Hybrid Systems, (http://tc.ifac-control.org/1/3/) and on the IEEE Control Systems Society Discrete Event Systems Technical Committee (http://discrete-event-systems.ieeecss.org).

6.2. Discrete control and reactive language support

Participants: Gwenaël Delaval, Eric Rutten, Stéphane Mocanu.

Concerning language support, we have designed and implemented BZR, a mixed imperative/declarative programming language: declarative contracts are enforced upon imperatively described behaviors (see 5.1). The semantics of the language uses the notion of Discrete Controller Synthesis (DCS) [5]. We target the application domain of adaptive and reconfigurable systems: our language can serve programming closed-loop adaptation controllers, enabling flexible execution of functionalities w.r.t. changing resource and environment conditions. DCS is integrated into a programming language compiler, which facilitates its use by users and programmers, performing executable code generation. The tool is concretely built upon the basis of a reactive programming language compiler, where the nodes describe behaviors that can be modeled in terms of transition systems. Our compiler integrates this with a DCS tool [3]. This work is done in close cooperation with the Inria team Sumo at Inria Rennes (H. Marchand). Ongoing work concerns aspects of compilation and debugging and logico-numeric extension of BZR based on the ReaX tool developed at Inria Rennes in the framework of the ANR Ctrl-Green project (see 8.2.1).

We are also currently working on on combining maximally permissive discrete control with runtime mechanisms for choosing between valid control values, involving e.g. a classical controller or stochastic aspects ; and on 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. In this research, we model adaptation as a supervisory control problem and synthesize a controller that guides the behavior of a software system during adaptation. The system during adaptation is modeled using a graph transition system and properties to be enforced are specified using an automaton. To ensure correctness, we then synthesize a controller that imposes constraints on the system during adaptation [14].

6.3. Design and programming

6.3.1. Component-based approaches

Participants: Frederico Alvares, Eric Rutten.

Component-based architectures have shown to be very suited for self-adaptation purposes, not only because of their intrinsic characteristics like reusability and modularity, but also as virtue of their dynamical reconfiguration capabilities. The issue, nevertheless, remains that adaptation behaviors are generally conceived by means of fine-grained reconfiguration actions from the very initial configurations. This way, besides the complexity in managing large-sized architectures, the space of reachable configurations is not know in advance, which prevents ensuring well-mastered adaptive behaviours. We address this problem by designing Ctrl-F, a domain-specific language which objective is to provide high-level support for describing adaptation behaviors and policies in component-based architectures. The proposed language lies on synchronous reactive programming, which means that it benefits of an entire environment and formal tooling allowing for the verification and control of reconfigurations. We show the applicability of Ctrl-F by first integrating it to FraSCAti, a Service Component Architecture middleware platform, and then by applying it to Znn.com, a well known self-adaptive case study.

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. On a related topic, we are also starting a cooperation on introducing reactive control in hierarchical autonomic architectures, with A. Diaconescu and E. Najm at TelecomParisTech.

6.3.2. Rule-based systems

Participants: Julio Cano, Adja Sylla, Gwenaël Delaval, Eric Rutten.

Event-Condition-Action (ECA) rules are a widely used language for the high level specification of controllers in adaptive systems, such as Cyber-Physical Systems and smart environments, where devices equipped with sensors and actuators are controlled according to a set of rules. The evaluation and execution of every ECA rule is considered to be independent from the others, but interactions of rule actions can cause the system behaviors to be unpredictable or unsafe. Typical problems are in redundancy of rules, inconsistencies, circularity, or application- dependent safety issues. Hence, there is a need for coordination of ECA rule-based systems in order to ensure safety objectives. We propose a tool-supported method for verifying and controlling the correct interactions of rules, relying on formal models related to reactive systems, and Discrete Controller Synthesis (DCS) to generate correct rule controllers [12].

We work on this topic in cooperation with CEA LETI/DACLE (L. Gurgen) and target the application and experimentation domain of smart environment in the Internet of Things [11].

Another complementary direction on which we are starting a cooperation with CEA LETI/DACLE is the topic of a high-level language for safe rule-based programming in the LINC platform: the PhD of Adja Sylla on this topic will be co-advised with F. Pacull and M. Louvel at CEA.

6.4. Infrastructure-level support

6.4.1. Autonomic Cloud and Big-Data systems

This activity continues work started several years ago in the Sardes Inria-team, before it split into Erods (at LIG) and Ctrl-A (at Inria).

6.4.1.1. Coordination in multiple-loop autonomic Cloud systems

Participants: Soguy Gueye, Gwenaël Delaval, Stéphane Mocanu, Bogdan Robu, 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. We address this problem in the context of the ANR project Ctrl-Green (see 8.2.1), in cooperation with LIG (N. de Palma) in the framework of the PhD of S. Gueye. We propose a method focusing on the

discrete control of the interactions of managers [7] [9]. We follow a component-based approach and explore modular discrete control, allowing to break down the combinatorial complexity inherent to the state-space exploration technique [13]. This improves scalability of the approach and allows constructing a hierarchical control. It also allows re-using complex managers in different contexts without modifying their control specifications. We build a component-based coordination of managers, with introspection, adaptivity and reconfiguration. We validate our method on a multiple-loop multi-tier system.

We are currently working on the distributed execution of modular controllers and on considering more control objectives, beyond purely discrete or logical ones, evaluating the new tool ReaX developed at Inria Rennes (Sumo) (see 6.2) and exploring continuous or stochastic control of servers provisioning.

6.4.1.2. Control for Big data

Participants: Bogdan Robu, Mihaly Berekmeri, Nicolas Marchand.

To deal with the issue of ensuring performance constraints while also minimizing costs in systems for Big Data analytics based on the parallel programming paradigm MapReduce, we propose a control theoretical approach, based on techniques that have already proved their usefulness for the control community. We develop an algorithm to create the first linear dynamic model for a Big Data MapReduce system, running a concurrent workload. Furthermore we identify two major performance constraint use cases: relaxed-minimal resource and strict performance constraints. For the first case we developed a feedback control mechanism and, to minimize the number of control actuations, an event-based feedback controller. For the second case we add a feedforward controller that efficiently suppresses the effects of large workload size variations. The work is validated in a simulated Matlab environment build at GIPSA-lab and online on a real 60 node MapReduce cluster (part of GRID 500), running a data intensive Business Intelligence workload. Our experiments demonstrate the success of the control strategies employed in assuring service time constraints [17], [18].

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

6.4.2. 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 work 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) [10]. We propose 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, and extending the use of control techniques by evaluating the new tool ReaX developed at Inria Rennes (Sumo) in the framework of the ANR Ctrl-Green project (see 6.2 and 8.2.1).

6.4.3. Autonomic memory management in HPC

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

Concurrent programs need to manage the time trade-off between synchronization and computing. A high concurrency level may decrease computing time but at the same time increase synchronization cost among threads. The traditional way to handle synchronization problems is through implementing locks. However locks suffer from the likelihood of deadlocks, vulnerability to failures, faults etc.. Software Transactional Memory (STM) has emerged as a promising technique to address synchronization issues through transactions. In STM, blocks of instructions accessing the shared data are wrapped into transactions. In STM each

transaction executes speculatively, and conflicts may be aroused when two transactions are trying to modify the same area simultaneously. A way to reduce conflicts is by adjusting concurrency levels. A suitable concurrency level can maximize program performance. However, there is no universal rule to decide the best concurrency level for a program from an offline view. Hence, it becomes necessary to adopt a dynamical tuning strategy to better manage a STM system, so that a program can achieve a better performance. In the context of the action-team HPES of the Labex Persyval-lab[0] (see 8.1), we explore the autonomic computing approach and control techniques to address these runtime tuning problems as a feedback control loop to automate the choices of concurrency levels, conflict management policies, and other parameters, with the objective of optimizing program execution time. This work is performed in cooperation with LIG (J.F. Méhaut) in the framework of the PhD of N. Zhou.

6.4.4. Control of smart environments

Participants: Julio Angel Cano Romero, Mengxuan Zhao, Eric Rutten, Hassane Alla [Gipsa-lab].

6.4.4.1. 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 are working 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 [15]. This work is performed in cooperation with Orange labs (G. Privat) in the framework of the Cifre PhD of M. Zhao.

6.4.4.2. Rule-based specification

In the Internet of things, Event - Condition - Action (ECA) 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 have worked on a case study for safe applications development in IoT and smart home environments [11].

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

MIMOVE Team

6. New Results

6.1. Introduction

MiMove's research activities in 2014 have focused on a set of areas directly related to the team's research topics. Hence, we have worked on Emergent Middleware (§ 6.3) and Service-oriented Computing in the Future Internet (§ 6.4), in relation to our research topic regarding Emergent Mobile Distributed Systems (§ 3.2). With respect to Large-scale Mobile Sensing & Actuation (§ 3.3), we have developed activities on Service-oriented Middleware for the Mobile Internet of Things (IoT) (§ 6.5), Composing Applications in the IoT (§ 6.6), and Lightweight Streaming Middleware for the IoT (§ 6.7). Last, our effort on Middleware for Mobile Social Networks (§ 6.8) is linked to our research on Mobile Social Crowd-sensing (§ 3.4).

Before presenting our new results in the areas mentioned above, we briefly discuss next the highlights of the year.

6.2. Highlights of the Year

This year has seen the following acknowledgments of the team's contributions:

- Valérie Issarny was distinguished as Chevalier de la Legion d'Honneur for her contributions to science and European scientific cooperation in research and education.
- One of the team's major publication by S. Ben Mokhtar, D. Preuveneers, N. Georgantas, V. Issarny, and Y. Berbers, titled "EASY: Efficient semAntic Service discoverY in pervasive computing environments with QoS and context support" [1], published in the Journal of Systems and Software (Volume 81, Issue 5), is one of the top ten (10) most cited papers among all the papers published by JSS in 2008.

6.3. Emergent Middleware

Participants: Emil Andriescu, Valérie Issarny, Thierry Martinez.

Our previous work on emergent middleware has focused on interconnecting functionally-compatible components, i.e., components that at some high level of abstraction require and provide compatible functionalities, but are unable to interact successfully due to mismatching interfaces and behaviors. To address these differences without changing the components, mediators that systematically enforce interoperability between functionally-compatible components by mapping their interfaces and coordinating their behaviors are required [18]. Our approach for the automated synthesis of mediators is performed through *interface matching*, which identifies the semantic correspondence between the actions required by one component and those provided by the other, followed by the *synthesis of correct-by-construction mediators*. To do so, we analyze the behaviors of components so as to generate the mediator that coordinates the matched actions in a way that guarantees that the two components progress and reach their final states without errors [2]. Our contribution primarily lies in handling interoperability from the application to the middleware layer in an integrated way. The mediators we synthesize act as: (i) translators by ensuring the meaningful exchange of information between components, (ii) controllers by coordinating the behaviors of the components to ensure the absence of errors in their interaction, and (iii) middleware by enabling the interaction of components across the network so that each component receives the data it expects at the right moment and in the right format.

In our latest work, we have particularly focused on item (iii) above. We recognize that modern distributed systems and Systems of Systems (SoS) are built as a composition of existing components and services. As a result, systems communicate (either internally, locally or over networks) using protocol stacks of ever-increasing complexity whose messages need to be translated (i.e., interpreted, generated, analyzed and transformed) by third-party systems. We are particularly interested in the application of message translation to achieve protocol interoperability via protocol mediators. We observe that current approaches are unable to provide an efficient solution towards reusing message translators associated with the message formats composed in protocol stacks. Instead, developers must write ad hoc "glue-code" whenever composing two or more message translators.

Ideally, message translators may be developed by separate parties, using various technologies, while developers should be able to compose them using an easy to use mechanism. However, parsers are monolithic and tightly constructed, which often makes it impossible to combine them, knowing that combining two unambiguous grammars (corresponding to two arbitrary parsers) may result in an ambiguous grammar, and that the ambiguity detection problem for context-free grammars is undecidable in the general case.

In addition to parser composition, the data structures of the parsing output must be manually defined, integrated and harmonized with the target systems (i.e., in this case, the Mediation Engine). As far as we know, the problem of inferring the output schema (or the data type) of an arbitrary tree transformation has not yet been solved, while it is known that, in general, a transformation might not be recognizable by a schema.

Following the challenges above, in [17], we make two major contributions to the issue of systematic message translation for modern distributed systems:

1. Starting from the premise that "off-the-shelf" message translators for individual protocols are readily available in at least an executable form, we propose a solution for the automated composition of message translators. The solution simply requires the specification of a composition rule that is expressed using a subset of the navigational core of the W3C XML query language XPath.
2. We provide a formal mechanism, using tree automata, which based on the aforementioned composition rule, generates an associated AST *data-schema* for the translator composition. This contribution enables the inference of correct data-schemas, relieving developers from the time-consuming task of defining them. On a more general note, the provided method solves the type inference problem for the *substitution* class of tree compositions in linear time on the size of the output. The provided inference algorithm can thus be adapted to a number of applications beyond the scope of this work, such as XML Schema inference for XSLT transformations.

The composition approach that we introduced functions as a purely "black-box" mechanism, thus allowing the use of third-party parsers and message serializers independently of the parsing algorithm they use internally, or the method by which they were implemented/generated. Our solution goes beyond the problem of translator composition by inferring AST data-schemas relative to translator compositions. This feature allows newly generated translators to be seamlessly (or even automatically) integrated with existing systems, and most notably our protocol mediation engine [2].

6.4. Service-oriented Computing in the Future Internet

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

With an increasing number of services and devices interacting in a decentralized manner, *choreographies* represent a scalable framework for the Future Internet. The service oriented architecture inherent to choreographies allows abstracting diverse systems as application components that interact via standard middleware protocols. However, the heterogeneous nature of such systems leads to choreographies that do not only include conventional services, but also sensor-actuator networks, databases and service feeds. We reason about the behavior of such systems by introducing abstract middleware connectors that follow base interaction paradigms, such as client-service (CS), publish-subscribe (PS) and tuple space (TS). These heterogeneous connectors are made interoperable through a service bus connector, the *eXtensible Service Bus* (XSB) [11].

In previous work, we identified and verified the behavioral semantics of the XSB connector derived from the interconnection of base connectors, and introduced a method for constructing protocol converters enabling this interconnection. We implemented our XSB solution into an extensible development and execution platform for application and middleware designers. We also provided a lightweight implementation of the XSB, the *Light Service Bus* (LSB), appropriate for resource-constrained environments and systems. Next, leveraging on the functional interoperability across interaction paradigms offered by the XSB, we initiated our study of end-to-end Quality of Service (QoS) properties of choreographies, where in particular we focus on the effect of middleware interactions on QoS.

Building on the above results, we refine our analysis of QoS on top of the identified interaction paradigms. We have introduced a motivating application scenario inspired from the *2014 D4D Challenge*[0]. More specifically, *Data for Development Senegal* is an innovation challenge on ICT Big Data for the purposes of societal development. Mobile network provider Sonatel (part of the Orange Group) has made anonymous data extracted from the mobile network in Senegal available to international research laboratories, encouraging research related to the development and welfare of the local population.

Our scenario targets the development of an application platform for citywide and countrywide transport information management relying on mobile social crowd-sensing. This takes into account the particular context and constraints in Senegal. More specifically, the local transportation system, although developing, still consists of many unplanned and informal settlements with unreliable services and infrastructure. Additionally, despite wide use of mobile phones in the country, mobile Internet access remains limited, making SMS the only alternative for data access for a large part of the population. Our proposition aims to complement the scarce authoritative transport information coming from structured information sources and compensate for the lack of such information. In particular, in our approach we intend to study and experiment with appropriate interaction paradigms (CS, PS, TS) on top of 3G/2G/SMS data connections, further depending on the specific application and data. We are especially interested in interaction adaptation depending on the network conditions (e.g., switching to SMS-based protocol when the 3G/2G network is unavailable).

We have taken a first step towards enabling such an application platform. This consists in evaluating the publish/subscribe interaction style in a large-scale setting where resources of mobile users are limited, which translates into limited and intermittent connectivity in the system. Additionally, such an application platform must guarantee that the sensing data is processed and delivered to the corresponding mobile users *on-time*, despite the intermittent connectivity of the latter. We have opted for the publish/subscribe paradigm, as it is deemed appropriate for loose spatio-temporal interaction between mobile entities.

In particular, we introduce a queueing network model for the end-to-end interaction within a large-scale mobile publish/subscribe system. We leverage the *D4D dataset* provided by Orange Labs to parametrize this model. We then develop a simulator named *MobileJINQS*[0] that implements our model and uses the dataset traces as realistic input load to the system model over the time span of a whole year. Prior to this, we extensively analyze the D4D dataset in order to identify the data that we are interested in and infer primary results [0]. Based on the results of our simulation-based experiments, we thoroughly evaluate the behavior of the publish/subscribe system and identify ways of tuning the system parameters in order to satisfy certain design requirements. More precisely, we provide results of simulations of our publish/subscribe system with varied incoming loads, service delays and event lifetime periods. We use connection data of various pairs of mobile network antennas to derive realistic traces for both incoming loads and service delays. System or application designers are able to tune the system by selecting appropriate lifetime periods. We demonstrate that varying incoming loads and service delays have a significant effect on response time. By properly setting event lifetime spans, designers can best deal with the tradeoff between freshness of information and information delivery success rates. Still, both of these properties are highly dependent on the dynamic correlation of the event input flow and delivery flow processes, which are intrinsically decoupled.

[0] http://www.d4d.orange.com/en/home
[0] http://xsb.inria.fr/d4d#mobilejinqs
[0] http://xsb.inria.fr/d4d

Our future work includes comparison of the publish/subscribe interaction paradigm with other interaction paradigms (client-server, tuple space), in relation with the network access capacity and the application requirements. Also, we intend to study the response time and success rate for the various combinations of antennas in more fine-grained scales (e.g., check what their evolution is over one day).

6.5. Service-oriented Middleware for the Mobile Internet of Things

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

The Internet of Things (IoT) is characterized by a wide penetration in the regular user's life through an increasing number of Things embedding sensing, actuating, processing, and communication capacities. A considerable portion of those Things will be mobile Things, which come with several advantages yet lead to unprecedented challenges. The most critical challenges, that are directly inherited from, yet amplify, today's Internet issues, lie in handling i) the large scale of users and mobile Things which lead to high communication and computation costs especially with the anticipated large volumes of data to exchange, ii) providing interoperability across the heterogeneous Things which host sensors and actuators providing services and producing data that follow different format/schema specifications, and iii) overcoming the unknown dynamic nature of the environment, due to the mobility of an ultra-large number of Things.

Service-Oriented Architecture (SOA) provides solid basis to address the above challenges as it allows the functionalities of sensors/actuators embedded in Things to be provided as services, while ensuring loose-coupling between those services and their hosts, thus abstracting their heterogeneous nature. In spite of its benefits, SOA has not been designed to address the ultra-large scale of the mobile IoT. Consequently, an alternative is provided within a novel Thing-based Service-Oriented Architecture, that revisits SOA interactions and functionalities, service discovery and composition in particular. Our work on the revisited Thing-based SOA is detailed in [9], [23], [15]. The novel architecture is concretized within MobIoT, a middleware solution that is specifically designed to manage and control the ultra-large number of mobile Things in partaking in IoT-related tasks.

In accordance with SOA, MobIoT comprises Discovery, Composition & Estimation, and Access components, yet modifies their internal functionalities. In more detail, the Discovery component enables Thing-based service registration (for Things to advertise hosted services) and look-up (for Things to retrieve remote services of interest). In order to handle the ultra large number of mobile Things and their services in the IoT, the component revisits the Service-Oriented discovery and introduces probabilistic protocols to provide, not all, but only a sufficient subset of services that can best approximate the result that is being sought after [23], [15] based on a predefined set of requirements such us sensing coverage of the area of interest and the location of the Things. By limiting the participation of Things, the communication costs and volumes of data to process are decreased without jeopardising the quality of the outcome.

Furthermore, the Composition & Estimation component (C&E) provides automatic composition of Thing-based services. This capacity is of interest in the case where no service can perform a required measurement/action task directly (based on its atomic functionalities). To that end, we model our composition specification as mathematical formulas defined semantically within a dedicated ontology. Thing-based service composition executes in three phases: i) expansion, where composition specifications are automatically identified; ii) mapping, where actual service instances (running services) are selected based on their functionalities and the physical attributes of their hosts; and iii) execution, where the services are accessed and the composition specifications are executed. Thing-based service composition revisits Service-Oriented composition by executing seamlessly with no involvement from developers or end users and relying on semantic technologies to identify the most appropriate services to compose.

Last but not least, the Access component provides an easy to use interface for developers to sample sensors/actuators while abstracting sensor/actuator hardware specifications. It revisits Service-Oriented access and leverage semantic technologies by executing access to services transparently and wrapping access functionalities internally. Thus, it alleviates that burden from users, initially in charge of this task. The Access component supports real-time query-based access to remote services and to locally hosted services.

To assess the validity of our proposed architecture, we provide a prototype implementation of MobIoT(§ 5.4) along with a set of extensive evaluations that demonstrate, not only the feasibility of our approach, but also the resulting quality of the discovery approach, along with its scalability, as compared to a regular SOA-based approach.

6.6. Composing Applications in the Internet of Things

Participants: Aness Bajia, Animesh Pathak, Françoise Sailhan.

Resilient computing is defined as the ability of a system to remain dependable when facing changes. To mitigate faults at runtime, dependable systems embark fault tolerance mechanisms such as replication techniques. These mechanisms have to be systematically and rigorously applied in order to guarantee the conformance between the application runtime behavior and its dependability requirements.

Given that devices and networks constituting the IoT are prone to failure and consequent loss of performance, it is natural that IoT applications are expected to encounter and tolerate several classes of faults - something that still largely remains within the purview of low-level-protocol designers. As part of our work on the MURPHY project (§ 7.1.1.1), we are addressing this issue by proposing: i) a set of abstractions that can be used during macroprogramming to express application-level fault tolerance requirements, as well as by developers of fault tolerance protocols to identify the abilities and requirements of their techniques; ii) a runtime system that employs adaptive fault tolerance (AFT) to provide fault tolerance to the networking sensing application; and iii) compilation techniques to instantiate and map tasks as needed to satisfy the requirements of the application for a given deployment. Through our work [26], we demonstrate that our approach provides this much-needed feature to networked sensing applications with negligible development- and minimal performance- overhead.

Complementary to the above, we have proposed task mapping algorithms to satisfy those requirements through a constraint programming approach [24]. Through evaluations on realistic application task graphs, we show that our constraint programming model can effectively capture the end-to-end requirements and efficiently solves the combinatorial problem introduced.

We have been continually incorporating our research results in the above areas into *Srijan* (§ 5.5), which provides an easy-to-use graphical front-end to the various steps involved in developing an application using the ATaG macroprogramming framework.

6.7. Lightweight Streaming Middleware for the Internet of Things

Participants: Benjamin Billet, Valérie Issarny.

The IoT raises many challenges related to its very large scale and high dynamicity, as well as the great heterogeneity of the data and systems involved (e.g., powerful versus resource-constrained devices, mobile versus fixed devices, continuously-powered versus battery-powered devices, etc.). These challenges require new systems and techniques for developing applications that are able to: (i) collect data from the numerous data sources of the IoT, and (ii) interact both with the environment using the actuators and with the users using dedicated GUIs. Given the huge volume of data continuously being produced by sensors (measurements and events), we must consider: (i) data streams as the reference data model for the IoT and, (ii) continuous processing as the reference computation model for processing these data streams. Moreover, knowing that privacy preservation and energy consumption are increasingly critical concerns, we claim that all the Things should be autonomous and work together in restricted areas as close as possible to the users rather than systematically shifting the computation logic into powerful servers or into the cloud.

Toward that goal, we have been developing Dioptase [3], a service-oriented middleware for the IoT, which aims to integrate the Things and their streams into today's Web by presenting sensors and actuators as Web services. The research work around the Dioptase middleware consists in designing new service-oriented architectures where services continuously process data streams instead of finite datasets. In this context, new composition mechanisms are investigated in order to provide a way to describe complex fully-distributed stream-based tasks and to deploy them dynamically, at any time, as task graphs, over available Things of the network, including

resource-constrained ones. To this end, Dioptase enables task graphs to be composed of Thing-specific tasks (directly implemented on the Thing) and dynamic tasks that communicate using data streams. Dynamic tasks are then described in a lightweight DSL, called *DiSPL*, which is directly interpreted by the middleware and provides specific primitives to manipulate data streams.

As part of the design of such composition mechanisms, we have been investigating the problem of task mapping and automated deployment, which basically consists of mapping a set of tasks onto a set of nodes. Given the specific challenges introduced by the IoT, we worked on a new formalization of the task mapping problem that captures the varying consumption of resources and various constraints (location, capabilities, QoS) in order to compute a mapping that guarantees the lifetime of the concurrent tasks inside the network and the fair allocation of tasks among the nodes (load balancing). This formalization, called *Task Graph to Concrete Actions (TGCA)* [19], results in a binary programming problem for which we provide an efficient heuristic that allows its resolution in polynomial time. Our experiments show that our heuristic: (i) gives solutions that are close to optimal, and (ii) can be implemented on reasonably powerful Things and performed directly within the network without requiring any centralized infrastructure.

6.8. Middleware for Mobile Social Networks

Participants: Animesh Pathak, George Rosca.

As recent trends show, online social networks (OSNs) are increasingly turning mobile and further calling for decentralized social data management. This trend is only going to increase in the near future, based on the increased activity, both by established players like Facebook and new players in the domain such as Google, Instagram, and Pinterest. Modern smart phones can thus be regarded as *social sensors*, collecting data not only passively using, e.g., Bluetooth neighborhoods, but actively in the form of, e.g., "check-in"s by users to locations. The resulting (mobile) social ecosystems are thus an emergent area of interest.

The recent years have seen three major trends in the world of online social networks: *i)* users have begun to care more about the privacy of their data stored by large OSNs such as Facebook, and have won the right (at least in the EU) to remove it completely from the OSN if they want to; *ii)* OSNs are making their presence felt beyond casual, personal interactions to corporate, professional ones as well, starting with LinkedIn, and most recently with the purchase by Microsoft of Yammer, the enterprise social networking startup, and the launch of Google Plus for enterprise customers; and *iii)* users are increasingly using the capabilities of their (multiple) mobile devices to enrich their social interactions, ranging from posting cellphone-camera photos on Instagram to "checking-in" to a GPS location using Foursquare.

In view of the above, we envision that in the near future, the use of ICT to enrich our social interactions will grow (including both personal and professional interactions), both in terms of size and complexity. However current OSNs act mostly like data silos, storing and analyzing their users' data, while locking in these users to their servers, with non-existent support for federation; this is reminiscent of the early days of email, where one could only email those who had accounts on the same Unix machine. The knee-jerk reaction to this has been to explore completely decentralized social networks, which give the user complete control over and responsibility of their social data, while resorting to peer-to-peer communication protocols to navigate their social networks. Unfortunately, there are few techniques available to reconcile with the fact that the same user might have multiple devices, or that it is extremely resource-consuming to perform complex analysis of social graphs on small mobile devices.

Our view lies somewhere in the middle of the two extremes, taking inspiration from the manner in which users currently use email. While their inboxes contain an immense amount of extremely personal data, most users are happy to entrust it to corporate or personal email providers (or store and manage it individually on their personal email servers) all the while being able to communicate with users on any other email server. The notion of *Federated Social Networks* (FSNs) —already gaining some traction— envisions a similar ecosystem where users are free to choose OSN providers which will provide storage and management of their social information, while allowing customers using different OSN providers to interact socially. Such a federation can be beneficial in three major ways, among others: *i)* it allows users to enjoy properties such as reliability,

availability, and computational power of the hosting infrastructure of their choice, while not being locked down in terms of whom they can communicate with; *ii)* much like spam filtering services provided by modern email providers, that are tuned by feedback from their users, FSN users can benefit from the behavior of others sharing the same OSN provider [0]; and *iii)* this fits perfectly with enterprise needs, where ad-hoc teams can be formed across corporate OSN providers of two organizations to work on a joint project.

In [30], we presented a set of requirements, followed by a survey of the state of the art in social networking solutions, with a special focus on their ability to support rich privacy and access control policies in federated settings. Through this extensive analysis we offer a broad vision on existing social networking platforms, protocols involved but also their privacy and access policies. By doing so, we identify the main components of a federated social platform together with presenting the current trends in standards and security paradigms underlying actual open source solutions which offers their implementation, and finally provides recommendations on constructing such systems. Our research is continually being incorporated into the Yarta middleware for mobile social networking(§ 5.7).

[0]This also gives an incentive to commercial OSN providers to provide value-added services.

MYRIADS Project-Team

5. New Results

5.1. Highlights of the Year

- The Contrail project coordinated by Christine Morin received the "Excellent" grade at its final review held on March 14th, 2014 in Brussels.
- Anne-Cécile Orgerie has been awarded the Young Researcher prize of the Lyon city in November 2014.
- Christine Morin has been awarded one of the 12 "Etoile de l'Europe 2014" prizes in December 2014 for the coordination of the Contrail European project.

BEST PAPERS AWARDS :
[18] **4th International Conference on Cloud Computing and Services Science**. H. FERNANDEZ, C. STRATAN, G. PIERRE.

5.2. Dependable Cloud Computing

Participants: Jiajun Cao, Stéphane Chevalier, Gene Cooperman, Teodor Crivat, Roberto-Gioacchino Cascella, Stefania Costache, Florian Dudouet, Filippo Gaudenzi, Anna Giannakou, Yvon Jégou, Ancuta Iordache, Christine Morin, Anne-Cécile Orgerie, Edouard Outin, Nikolaos Parlavantzas, Jean-Louis Pazat, Guillaume Pierre, Aboozar Rajabi, Louis Rilling, Matthieu Simonin, Arnab Sinha, Cédric Tedeschi.

5.2.1. Deployment of distributed applications in a multi-provider environment

Participants: Roberto-Gioacchino Cascella, Stefania Costache, Florian Dudouet, Filippo Gaudenzi, Yvon Jégou, Christine Morin, Arnab Sinha.

The move of users and organizations to Cloud computing will become possible when they are able to exploit their own applications, applications and services provided by cloud providers, as well as applications from third party providers in a trustful way on different cloud infrastructures. In the framework of the Contrail European project [2] [46], we have designed and implemented the Virtual Execution Platform (VEP) service in charge of managing the whole life cycle of OVF distributed applications under Service Level Agreement rules on different infrastructure providers [47]. In 2013, we designed the CIMI inspired REST-API for VEP 2.0 with support for Constrained Execution Environment (CEE), advance reservation and scheduling service, and support for SLAs [56], [55] [57]. We integrated support for delegated certificates and developed test scripts to integrate the Virtual Infrastructure Network (VIN) service. VEP 1.1 was slightly modified to integrate the usage control (Policy Enforcement Point (PEP)) solution developed by CNR. The CEE management interface was developed during 2013 and is available through the graphical API as well as through the RESTful API.

5.2.2. Checkpointing for multi-cloud environments

Participants: Jiajun Cao, Gene Cooperman, Christine Morin, Matthieu Simonin.

Most cloud platforms currently rely on each application to provide its own fault tolerance. A uniform mechanism within the cloud itself serves two purposes: (a) direct support for long-running jobs, which would otherwise require a custom fault-tolerant mechanism for each application; and (b) the administrative capability to manage an over-subscribed cloud by temporarily swapping out jobs when higher priority jobs arrive.

We propose ([31]) a novel *Checkpointing as a Service* approach, which enables application checkpointing and migration in heterogeneous cloud environments. Our approach is based on a non-invasive mechanism to add fault tolerance to an existing cloud platform *after the fact*, with little or no modification to the cloud platform itself. It achieves its cloud-agnostic property by using an external checkpointing package, independent of the target cloud platform. We implemented a prototype of the service on top of both OpenStack and Snooze IaaS clouds. We conducted a preliminary performance evaluation using the Grid'5000 experimentation platform.

5.2.3. Towards a distributed cloud inside the backbone

Participants: Anne-Cécile Orgerie, Cédric Tedeschi.

The DISCOVERY proposal currently in phase of construction and lead by Adrien Lèbre from the ASCOLA team, and currently on leave at Inria aims at designing a distributed cloud, leveraging the resources we can find in the network's backbone. [0]

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. The design, implementation, and experimentation of the overlay has been described in an article published in 2014 [22].

5.2.4. 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. A distributed Cloud is neither a usual Cloud built around a single data center, nor a Cloud Federation interconnecting different data centers owned and run by different administrative entities. Moreover, in the Cloud organization targeted here, the network capabilities can be dynamically configured in order to apply optimizations to guarantee QoS for streaming or negotiated bandwidth for example. 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 the Distributed Cloud in order to consider both optimizations for energy consumption and for users' QoS needs. The objectives of these optimizations 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 will rely 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) will be used for the implementation of the system. The technical developments are based on the Openstack framework.

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.

5.2.5. Multi-cloud application deployment in ConPaaS

Participants: Stéphane Chevalier, Teodor Crivat, Guillaume Pierre.

We extended ConPaaS to support the deployment of smartphone backend applications in mobile operators' base stations. The motivation is to reduce the latency compared to a traditional deployment where the backend is located in an external cloud. This requires building a lightweight infrastructure which allows one to easily create containers that can be seamlessly migrated (roaming). A publication on this topic will appear in 2015 [23].

5.2.6. Application Performance Modeling in Heterogeneous Cloud Environments

Participants: Ancuta Iordache, Guillaume Pierre.

Heterogeneous cloud platforms offer many possibilities for applications for make fine-grained choice 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 publication on this topic is currently under review.

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

5.2.7. Dynamic reconfiguration for multi-cloud applications

Participants: Nikolaos Parlavantzas, Aboozar Rajabi.

In the context of the PaaSage European project, we are working on model-based self-optimisation 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 2014, we have defined the architecture of the adaptation system and produced a first prototype[30].

5.2.8. Self-adaptable Monitoring for Security in the Cloud

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

We aim at designing a self-adaptable system for security monitoring in clouds. The considered system should cope with the dynamic nature of virtual infrastructures in clouds and have a minimal impact on performance. In 2014, we studied the state of the art in cloud security monitoring, which is composed of various approaches for intrusion detection systems (IDS), based on traditional IDS techniques such as signature-based detection and anomaly-based detection.

As a first step towards our goal of making self-adaptable a complete security monitoring architecture for cloud environments, we defined a simple initial monitoring scenario for identifying the impact of the dynamicity of a cloud architecture on the intrusion detection process. In this scenario, the security monitoring infrastructure is composed of two network IDS instances, which are used to monitor the virtual infrastructures network traffic of two cloud clients (one virtual infrastructure per client), and also eventually monitor the physical infrastructure (that is the operator's infrastructure). The virtual network traffic in each host machine is monitored by only one of the IDS instances, so that the IDS instances must be adapted to topology changes (such as migration of VMs) in the cloud environment. The adaptation process includes updates of the rules configured in the instance (deletion or creation).

In 2014, we built our testbed based on OpenStack technology for the underlying IaaS cloud platform and Snort for the network IDS. At this point the testbed consists of only five machines (on the Grid'5000 platform) but we aim to increase the number of host machines and deploy more VMs for having a more realistic representation of a production network. This will allow us to study performance issues and also more complex security monitoring setups. Our goal is also to enable monitoring of other elements, such as resource usage (both per host and per VM) on the cloud provider side.

5.2.9. 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 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 the service distribution.

We first focus on the formal expression of these requirements, by considering first the requirements provided by a Service Operator to the "Fog" Infrastructure (required computing resources, minimal quality of experience (QoE) level, etc.). The resource allocation strategies should also take into account the topology of the "Fog" Infrastructure, the heterogeneous capabilities of the equipments and of the underlying network. Based on this information, strategies and algorithms for resource allocation should be designed that will participate in the process of building an efficient platform for the service distribution. Evaluation of this efficiency will be an important process to justify the relevance of the strategies.

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

5.3. Heterogeneous Resource Management

Participants: Eliya Buyukkaya, Djawida Dib, Eugen Feller, Christine Morin, Nikolaos Parlavantzas, Guillaume Pierre.

5.3.1. Cross-resource scheduling in heterogeneous cloud environments

Participants: Eliya Buyukkaya, Guillaume Pierre.

Allocating resources to applications in a heterogeneous cloud environment is harder than in a homogeneous environment. In a heterogeneous cloud some rare resources are more precious than others, and should be treated carefully to maximize their utilization. Similarly, applications may request groups of resources that exhibit certain inter-resource properties such as the available bandwidth between the assigned resources. We are currently investigating scheduling algorithms for handling such scenarios.

5.3.2. Maximizing private cloud provider profit in cloud bursting scenarios

Participants: Djawida Dib, Christine Morin, Nikolaos Parlavantzas.

Current PaaS offerings either provide no support for SLA guarantees or provide limited support targeting a restricted set of application types. To overcome this limitation, we have developed an open, cloud-bursting PaaS system, called Meryn, designed to be easily extensible to host new application types. The system integrates a decentralized optimization policy that maximises the PaaS provider profit, taking into account the payment of penalties incurred when quality guarantees are unsatisfied. The system was implemented and evaluated on the Grid5000 testbed using batch and MapReduce workloads. The results demonstrated the effectiveness of the policy in increasing provider profit [16] This work was part of Djawida Dib's PhD thesis [10] defended in July 2014.

5.3.3. Data life-cycle management in clouds

Participants: Eugen Feller, Christine Morin.

Infrastructure as a Service (IaaS) clouds provide a flexible environment where users can choose and control various aspects of the machines of interest. However, the flexibility of IaaS clouds presents unique challenges for storage and data management in these environments. Users use manual and/or ad-hoc methods to manage storage and data in these environments. FRIEDA is a Flexible Robust Intelligent Elastic Data Management framework that employs a range of data management strategies approaches in elastic environments. This year, our work carried out in the context of the DALHIS associate team [0], was focused on the extended design and evaluation of the FRIEDA data management system. FRIEDA was tested to work on Amazon EC2 resources. In addition, we layered a commandline utility atop FRIEDA that allows users to plug-in applications to run in FRIEDA. These tools have been adopted by the LBL-ATLAS group to run their experiments on Amazon [29].

5.4. Energy-efficient Resource Infrastructures

Participants: Maria Del Mar Callau Zori, Alexandra Carpen-Amarie, Bogdan Florin Cornea, Ismael Cuadrado Cordero, Djawida Dib, Eugen Feller, Sabbir Hasan Rochi, Yunbo Li, Christine Morin, Anne-Cécile Orgerie, Jean-Louis Pazat, Guillaume Pierre, Lavinia Samoila.

5.4.1. Energy-efficient IaaS clouds

Participants: Alexandra Carpen-Amarie, Christine Morin, Anne-Cécile Orgerie.

[0] http://project.inria.fr/dalhis

Energy consumption has always been a major concern in the design and cost of data centers. The wide adoption of virtualization and cloud computing has added another layer of complexity to enabling an energy-efficient use of computing power in large-scale settings. Among the many aspects that influence the energy consumption of a cloud system, the hardware-component level is one of the most intensively studied. However, higher-level factors such as virtual machine properties, their placement policies or application workloads may play an essential role in defining the power consumption profile of a given cloud system. In this work, we explored the energy consumption patterns of Infrastructure-as-a-Service (IaaS) cloud environments under various synthetic and real application workloads. For each scenario, we investigated the power overhead triggered by different types of virtual machines, the impact of the virtual cluster size on the energy-efficiency of the hosting infrastructure and the tradeoff between performance and energy consumption of MapReduce virtual clusters through typical cloud applications [45].

5.4.2. Energy-aware IaaS-PaaS co-design

Participants: Maria Del Mar Callau Zori, Alexandra Carpen-Amarie, Djawida Dib, Anne-Cécile Orgerie, Guillaume Pierre, Lavinia Samoila.

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 argue that cross-layer cooperation in clouds is a key to achieving an optimized resource management, both performance and energy-wise. To this end, we claim there is a need for a cooperation API between IaaS and PaaS clouds, enabling each layer to share specific information and to trigger correlated decisions. We identified the drawbacks raised by such co-design objectives and discuss opportunities for energy usage optimizations. A position paper has been published on these aspects [15]. Ongoing work is currently conducted in order to quantify the actual possible gains both energy and performance-wise for this IaaS-PaaS co-design approach.

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

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

Cloud computing is increasingly becoming an essential component for Internet service provision, yet at the same time its energy consumption has become a key environmental and economic concern. It becomes urgent to improve the energy efficiency of such infrastructures. Our work aims at designing energy-efficient resource allocation for Cloud infrastructures. Yet, energy is not the only criterion to take into account at risk of losing users. A multi-criteria approach is required in this context to satisfy both users and Cloud providers.

The proposed resource allocation algorithms will take into account not only the computing resources but also the storage and networking resources. Indeed, the ever-growing appetite of new applications for network resources leads to an unprecedented electricity bill for network resources, and for these bandwidth-hungry applications, networks can become an significant bottleneck. This phenomenon is emphasized with the emergence of the big data paradigm. The designed algorithms would thus integrate the data locality dimension to optimize computing resource allocation while taking into account the fluctuating limits of network resources.

In 2014, several experiments were performed to understand and quantify networking energy consumption. These experiments include network protocol energy consumption in the devices, configuration energy consumption in switching/routing devices and associated energy consumption to real cloud computing applications (e.g. Google drive). These experiments have been performed over systems provided by Inria such as Grid'5000 and specific network devices (e.g. level 3 router for a private LAN). Based on this work, we developed an analytic model of networking energy consumption in a cloud computing environment. This analysis will serve as a basis for designing an energy-efficient architecture and related algorithms.

5.4.4. Simulating Energy Consumption of Wired Networks

Participants: Bogdan Florin Cornea, Anne-Cécile Orgerie.

Predicting the performance of applications, in terms of completion time and resource usage for instance, is critical to appropriately dimension 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 dimension 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 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. Through this tool, we have studied the energy consumption of data transfers in Clouds [19]. This work has been done in collaboration with the Avalon team from LIP in Lyon.

5.4.5. 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.

5.4.6. 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 propose an approach driven by Service Level Agreements (SLAs) for Cloud auto-scaling. A SLA defines a formal contract between a service provider and a service consumer on an expected QoS level. The main idea of this thesis is to exploit the SLA requirements to (i) avoid the interferences between the Cloud autonomic managers by a cross-layer coordination of SLA contracts; (ii) fine-tune the resources needs according to SLA by proposing both dynamic resources provisioning for optimizing the energy footprint and dynamic reconfiguration at the SaaS level to optimize the expected QoS. In particular, we propose to address renewable energy in the SLA contract. The objective is twofold. First, for ecological reasons, it allows Cloud users to express their preferences about the energy provider and the nature of the energy in the data center. Then, for economic reasons, it takes advantage of renewable energy costs (expressed in the SLA) to reconfigure resource allocation and energy usage. The integration of such SLAs in each layer of the Cloud stack and their management by an autonomic manager or by the coordination of autonomic managers still remain open issues.

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

5.4.7. Simulating the impact of DVFS within SimGrid

Participants: Alexandra Carpen-Amarie, Christine Morin, Anne-Cécile Orgerie.

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 experimentations 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.

5.5. Decentralised and Adaptive workflows

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

5.5.1. Template workflows

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

In the framework of the DALHIS associate team [0], we started to combine the high-level template workflow language TIGRES [0], developed by our partner team from Lawrence Berkeley National Lab (LBL) with the workflow management system developed in the team [5]. The design of this integration and its benefits have been presented in a workhsop article [24].

5.5.2. Adaptive Workflows with Chemical Computing

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

We are currently designing a complete programming model for the management of adaptive workflows, based on an extension of the HOCL language, in particular workflows that may evolve at run time in their shape. An article is under preparation.

5.5.3. Best-effort decentralised workflow execution

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

[0] http://project.inria.fr/dalhis
[0] http://tigres.lbl.gov/home

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 centralised orchestrator. Such a workflow engine relies on point-to-point cooperation between nodes supporting the execution.

5.6. Experimental Platforms

Participants: Maxence Dunnewind, Nicolas Lebreton, Julien Lefeuvre, David Margery, Eric Poupart.

5.6.1. Energy measurement

Participants: Maxence Dunnewind, Nicolas Lebreton, David Margery, Eric Poupart.

In the context of the ECO_2Clouds project, the BonFIRE infrastructure was updated. At the software layer, the complete monitoring stack was revisited so as to attribute power consumption values to all VMs running on the infrastructure and to expose this information to users. This was used by the project partners to confirm that using an eco-aware scheduler could significantly reduce eco-impact of running a distributed infrastructure.

5.6.2. BonFIRE

Participants: Maxence Dunnewind, Julien Lefeuvre, David Margery, Eric Poupart.

The project was reviewed in December 2013 during CloudCom 2013 in Bristol and rated Excellent. It has been kept in working state through our commitment to the BonFIRE foundation. The main acheivment on this topic was to evolve the cloud reservation system so as to support tracking usage using allocation blocks, as a fragment of the physical machines. Instance types can therefore have a different footprint in number of allocation blocks depending on the hardware they are scheduled on.

5.6.3. Fed4FIRE

Participants: Nicolas Lebreton, 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. Here, we contributed to a proposal to secure usage of OMF and to a design to allow using BonFIRE through SFA. In 2014, the main area of work has been maintenance of the infrastructure and initial prototyping of an SFA API to BonFIRE.

REGAL Project-Team

5. New Results

5.1. Highlights of the Year

- *Garbage collection for big data on large-memory NUMA machines*. We developed NumaGiC, a high-throughput garbage collector for big-data algorithms running on large-memory NUMA machines (see Section 4.1). This result, a collaboration with the Whisper team, will be presented at ASPLOS 2015 [29].
- *Explicit consistency*. We propose an alternative approach to the strong-vs.-weak consistency conundrum, *explicit consistency*. Static analysis identifies precisely what is the minimal amount of synchronisation that is necessary to maintain the invariants required by an application (see Section 5.3.11). This result will be presented at EuroSys 2015 [53].
- *Lower bounds and optimality for CRDTs*. This is the first paper to study the inherent lower bounds of replicated data types. The contribution includes derivation of lower bounds for several data types, improvement of some implementations, and proved optimality of others (see Section 5.3.10). This result was presented at POPL 2014 [25].

5.2. Distributed algorithms for dynamic networks

Participants: Luciana Bezerra Arantes [correspondent], Rudyar Cortes, Raluca Diaconu, Jonathan Lejeune, Olivier Marin, Sébastien Monnet, Franck Petit [correspondent], Karine Pires, Pierre Sens, Véronique Simon, 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.

Efficiency in such environments requires specialised protocols, providing features such as fault or heterogeneity tolerance, scalability, quality of service, and self-*. Our approach covers the whole spectrum from theory to experimentation. We design algorithms, prove them correct, implement them, and evaluate them in simulation, using OMNeT++ or PeerSim, and on large-scale real platforms such as Grid'5000. The theory ensures that our solutions are correct and whenever possible optimal; experimental evidence is necessary to show that they are relevant and practical.

Within this thread, we have considered a number of specific applications, including massively multi-player on-line games (MMOGs) and peer certification.

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: mutual exclusion, failure detection, data dissemination and data finding in large scale systems, self-stabilization and self-* services.

5.2.1. Self-Stabilization.

We have also approached fault tolerance through self-stabilization. Self-stabilization is a versatile technique to design distributed algorithms that withstand transient faults.

In [43], we proposed a silent self-stabilizing leader election algorithm (SSLE, for short) for bidirectional connected identified networks of arbitrary topology. Starting from any arbitrary configuration, SSLE converges to a terminal configuration, where all processes know the ID of the leader, this latter being the process of minimum ID. Moreover, as in most of the solutions from the literature, a distributed spanning tree rooted at the leader is defined in the terminal configuration. This algorithm is written in the locally shared memory model. It assumes the distributed unfair daemon, the most general scheduling hypothesis of the model. Our algorithm requires no global knowledge on the network (such as an upper bound on the diameter or the number of processes, for example). We showed that its stabilization time is in $\Theta(n^3)$ steps in the worst case, where n is the number of processes. Its memory requirement is asymptotically optimal, *i.e.*, $\Theta(\log n)$ bits per processes. Its round complexity is of the same order of magnitude — *i.e.*, $\Theta(n)$ rounds — as the best existing algorithm designed with similar settings. To the best of our knowledge, this was the first self-stabilizing leader election algorithm for arbitrary identified networks that is proven to achieve a stabilization time polynomial in steps. By contrast, we show that the previous best existing algorithm designed with similar settings stabilizes in a non polynomial number of steps in the worst case.

We have also implemented SSLE in a high-level simulator to empirically evaluate its average performances. Experimental results tend to show that its worst case in terms of rounds ($\Theta(3n + D)$ rounds) is rare.

5.2.2. Dynamic Distributed Systems

The first key challenge in understanding highly dynamic networks consists in developing appropriate models that are as close as possible to the phenomena that one wishes to capture. This requires the use of a formalism sufficiently expressive to formulate complex temporal properties. Recently, a vast collection of concepts, formalisms, and models has been unified in a framework called Time-Varying Graphs (TVG) [0], which are represented as time-ordered sequences of graphs defined over a fixed set of nodes. A hierarchy of classes over TVG has been described, mainly depending on properties related to connectivity and recurrence of dynamic. Such an hierarchy is an interesting tool for study computability issues. As an example, if one is able to prove an impossibility result in a class of the hierarchy with strong properties, then this impossibility result also holds in any class of the hierarchy with (strictly) weaker properties. In this context, we provide a generic framework to prove impossibility results in this model [45]. This framework helps to formally prove classical arguments about convergence of sequence of time-varying graphs used to build counter-examples. We apply this generic framework to the study of covering problems (such as minimal dominating set and maximal matching) in the context of time-varying graphs. We obtain a characterization of the weakest topology assumption that makes these problems computable. We also propose a general time complexity measure since time-varying graph model lacks so far of such a definition.

5.2.3. Swarm 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. Those robots must collaborate to solve a collective task, namely *exclusive perpetual exploration*, despite being limited with respect to input from the environment, asymmetry, memory, etc. The area to be explored is modeled as a graph and the exclusive perpetual exploration task requires every possible vertex to be visited infinitely often by every robot, with the additional constraint that no two robots may be present at the same node at the same time or may concurrently traverse the same edge of the graph.

In [28], we presented and implemented a generic method for obtaining all possible protocols for a swarm of mobile robots operating in a particular discrete space, namely an anonymous rings. Our method permits to discover new protocols that solve the problem, and to assess specific optimization criteria (such as individual coverage, visits frequency, etc.) that are met by those protocols. To our best knowledge, this was the first attempt to mechanize the discovery and fine-grained property testing of distributed mobile robot protocols.

[0] A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro, Time-varying graphs and dynamic networks, International Journal of Parallel, Emergent and Distributed Systems 27(5):387-408, 2012

5.3. Management of distributed data

Participants: Pierpaolo Cincilla, Raluca Diaconu, Jonathan Lejeune, Mesaac Makpangou, Olivier Marin, Sébastien Monnet, Karine Pires, Dastagiri Reddy Malikireddy, Masoud Saeida Ardekani, Pierre Sens, Marc Shapiro, Véronique Simon, Julien Sopena, Vinh Tao Thanh, Serdar Tasiran, Marek Zawirski.

Storing and sharing information is one of the major reasons for the use of large-scale distributed computer systems. Replicating data at multiple locations ensures that the information persists despite the occurrence of faults, and improves application performance by bringing data close to its point of use, enabling parallel reads, and balancing load. This raises numerous issues:

- Where to store or replicate the data, in order to ensure that it is available quickly and remains persistent despite failures and disconnections.
- How many copies, located where, are needed to face dynamically-changing demand (load) and offer (elasticity).
- How to parallelize writes and hence how to ensure consistency between replicas.
- Tradeoffs between synchronised, consistent but slow updates, and fast but weakly-consistent ones.
- When and how to move data to computation, or computation to data, in order to improve response time while minimizing storage or energy usage.
- How to apply our approaches towards addressing the above issues onto a challenging use case: achieving true scalability for online games.

5.3.1. Long term durability

To tolerate failures, distributed storage systems replicate data. However, despite the replication, pieces of data may be lost (i.e. all the copies are lost). We have previously proposed a mechanism, RelaxDHT, to make distributed hash tables (DHT) resilient to high churn rates.

We have observed that a given system with a given replication mechanism can store a certain amount of data above which the loss rate would be greater than an “acceptable”/fixed threshold. This amount of data can be used as a metric to compare replication strategies. We have studied the impact of the data distribution layout upon the loss rate. The way the replication mechanism distribute the data copies among the nodes has a great impact. If node contents are very correlated, the number of available sources to heal a failure is low. On the opposite, if the data copies are shuffled/scattered among the nodes, many source nodes may be available to heal the system, and thus, the system losses less pieces of data. In order to study data durability on a long term, we have designed a model, and implemented a discrete event based simulator that can simulate a 100 node system over years within several hours. Our model, SPLAD [49] (for scattering and placing data replicas to enhance long-term durability), allows us to vary the data scattering degree by tuning a selection range width. We are also studying the impact of the policy used while choosing a storing node within the selection range (e.g., randomly, the least loaded, or smarter policies like the power of two choices). This policy has an important impact on both the storage load distribution among nodes and the number of lost pieces of data.

5.3.2. Achieving scalability for online games

Massively Multiplayer Online Games (MMOGs) such as *World of Warcraft* constitute a great use case for the management of distributed data on a large scale. Commercial support systems for MMOGs rely almost exclusively on traditional client/server architectures that are centralized. These architectures do not scale properly, both in terms of the number of players and of the number objects used to model virtual universes that grow ever more complex. Most MMOGs avoid this problem by limiting the scale of the universe: the virtual environment is partitioned into several parallel and totally disconnected worlds, such as the *Realms* in *World of Warcraft*. Each partition, handled in a centralized way, limits the number of players it can host; avatars created on different partitions will never meet in the game.

From a systems point of view, achieving true scalability raises many challenging issues for MMOGs. For instance the system must be very reactive: if the update latency on a player node is too high, the game becomes unplayable. Since these games are meant to operate on a large scale, they induce a trade-off between availability and consistency of data. The consistency aspect is critical because MMOGs incur a high degree of cheating.

Designing and implementing a scalable service for Multiplayer Online Games requires an extensive knowledge of the habits, behaviors and expectations of the players. The first part of our work on MMOGs aimed at gathering and analyzing traces of real games offers to gain insight on these matters. We collected public data from a *League of Legends* server (information over more than 56 million game sessions): the resulting database is freely available online, and an ensuing publication [34] details the analysis and conclusions we draw from this data regarding the expected requirements for a scalable MMOG service.

We steered a second part of our work on MMOGs in 2014 towards designing a peer to peer refereeing system that remains highly efficient, even on a large scale, both in terms of performance and in terms of cheat prevention. Simulations show that such a system scales easily to more than 30,000 nodes while leaving less than 0.013% occurrences of cheating undetected on a mean total of 24,819,649 refereeing queries. This work got published in the Multimedia Systems Journal [21].

Finally, we also worked on the design of a scalable architecture for online games. The goal is to balance the load among nodes to allow the simulation of a whole, contiguous, virtual space.

5.3.3. 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 2014, we focused our research on data placement and on gathering traces from target applications in order to assess our future solutions.

With respect to placement, we worked on a scheme to store and access massive streams of data efficiently. We designed a solution that extends distributed prefix tree indexing structures for this purpose. Our new maintenance protocol anticipates every data insertion on provisional child nodes and thus significantly reduces overhead and improves query response time. This work has lead to the publication of an Inria research report (RR- 8637) [46].

With respect to application traces, we targeted sport tracker applications. Designing and implementing a big data service for sport tracker applications requires an extensive knowledge of both data distribution and input load. Gathering and analysing traces from a real world sports tracker service provides insight on these matters, but such services are very protective of their data due to competition as well as privacy issues. We avoided these issues by gathering public data from a popular sports tracker server called EndoMondo. The resulting database is freely available online, and allowed an in-depth analysis from a dynamic big data perspective. This study has lead to the publication of an Inria research report (RR- 8636) [47] .

5.3.4. Adaptative replication

Different pieces of data have different popularity: some data are stored but never accessed while other pieces are very “hot” and are requested concurrently by many clients. This implies that different pieces of data with different popularity should have a different number of copies to efficiently serve the requests without wasting resources. Furthermore, for a given piece of data, the popularity may vary drastically among time. It is thus important that the replication mechanism dynamically adapt the number of replicas to the demand. In the context of the ODISEA2 FUI project, we have studied the popularity distribution and evolution of live video streams [31], [36].

5.3.5. Keyword-based Indexing and Search Substract for Structured P2P Information System

Number of large scale information systems rely on a DHT-based storage infrastructure. To help users to find suitable information, one attractive solution is to maintain an index that maps keywords to suitable data. Maintaining and exploiting an index distributed towards a DHT is confronted to the performance issue. Mainly, the computation of the intersection of postings related to provided keywords could generate too large traffic over the network; also one is confronted to some unbalanced on peers' load due to the fact that certain world are too popular!

In 2014, we propose *FreeCore*, a DHT-based distributed indexing substract that can be used to build efficient keyword-based search facilities for large scale information systems. A *FreeCore* index, considers keyword sets, then summarizes each set with a Bloom Filter. To limit the probability of false positive, we anticipate that one will use large size filters together enough hash functions. Thanks to this representation, we transform the searching problem, to the one of bitmaps matching as each query is also coded by a Bloom Filter. To distribute resulting summaries towards peers, *FreeCore* considers each summary as a sequence of binary keywords. Each binary keyword is assigned a peer and all summaries containing this binary keyword are stored at its assigned peer. Finally, to reduce the traffic overhead as well as the the size of local indices, *FreeCore* fragments each filter such as to factorize sequence of bits that occur more than once. In [40], we report the performances of the initial implementation of *FreeCore*. Thought a number of improvements were not included within this initial evaluation, *FreeCore* offers better performances than existing state of the art. Current work focusses on developping applications that exploit *FreeCore*.

5.3.6. Large-Scale File Systems

Storage architectures for large enterprises are evolving towards a hybrid cloud model, mixing private storage (pure SSD solutions, virtualization-on-premise) with cloud-based service provider infrastructures. Users will be able to both share data through the common cloud space, and to retain replicas in local storage. In this context we need to design data structures suitable for storage, access, update and consistency of massive amounts of data at the object, block or file system level.

Current designs consider only data structures (e.g., trees or B+-Trees) that are strongly consistent and partition-tolerant (CP). However, this means that they are not available when there is a network problem, and that replicating a CP index across sites is painful. The traditional approaches include locking, journaling and replaying of logs, snapshots and Merkle trees. All of these are difficult to scale using generic approaches, although it is possible to scale them in some specific instances. For instance, synchronization in a single direction (the Active/Passive model) is relatively simple but very limited. A multi-master (Active/Active) model, where updates are allowed at multiple replicas and synchronization occurs in both directions, is difficult to achieve with the above techniques.

Our previous work has shown that many storage indexing operations commute; this enables a the highly-scalable CRDT approach. For those that do not, the explicit consistency approach (Section 5.3.11) appears promising.

This work is part of a CIFRE agreeement with Scality (see Section 6.2.1).

5.3.7. Strong consistency

When data is updated somewhere on the network, it may become inconsistent with data elsewhere, especially in the presence of concurrent updates, network failures, and hardware or software crashes. A primitive such as consensus (or equivalently, total-order broadcast) synchronises all the network nodes, ensuring that they all observe the same updates in the same order, thus ensuring strong consistency. However the latency of consensus is very large in wide-area networks, directly impacting the response time of every update. Our contributions consist mainly of leveraging application-specific knowledge to decrease the amount of synchronisation.

When a database is very large, it pays off to replicate only a subset at any given node; this is known as partial replication. This allows non-overlapping transactions to proceed in parallel at different locations and decreases the overall network traffic. However, this makes it much harder to maintain consistency. We designed and implemented two *genuine* consensus protocols for partial replication, i.e., ones in which only relevant replicas participate in the commit of a transaction.

Another research direction leverages isolation levels, particularly Snapshot Isolation (SI), in order to parallelize non-conflicting transactions on databases. We prove a novel impossibility result: under standard assumptions (data store accesses are not known in advance, and transactions may access arbitrary objects in the data store), it is impossible to have both SI and GPR. Our impossibility result is based on a novel decomposition of SI which proves that, like serializability, SI is expressible on plain histories.

We designed an efficient protocol that maintains side-steps this impossibility but maintains the most important features of SI:

1. (Genuine Partial Replication) only replicas updated by a transaction T make steps to execute T;
2. (Wait-Free Queries) a read-only transaction never waits for concurrent transactions and always commits;
3. (Minimal Commit Synchronization) two transactions synchronize with each other only if their writes conflict.

The protocol also ensures Forward Freshness, i.e., that a transaction may read object versions committed after it started.

Non-Monotonic Snapshot Isolation (NMSI) is the first strong consistency criterion to allow implementations with all four properties. We also present a practical implementation of NMSI called Jessy, which we compare experimentally against a number of well-known criteria. Our measurements show that the latency and throughput of NMSI are comparable to the weakest criterion, read-committed, and between two to fourteen times faster than well-known strong consistencies.

An interesting side-effect of this research is an apples-to-apples comparison of many strong-consistency protocols. This work was published at LADIS 2014 [41] and at Middleware 2014 [33].

This research is supported in part by ConcoRDanT ANR project (Section 7.1.7) and by the FP7 grant SyncFree (Section 7.2.1.1).

5.3.8. Distributed Transaction Scheduling

Parallel transactions in distributed DBs incur high overhead for concurrency control and aborts. Our Gargamel system proposes an alternative approach by pre-serializing possibly conflicting transactions, and parallelizing non-conflicting update transactions to different replicas. This system provides strong transactional guarantees. In effect, Gargamel partitions the database dynamically according to the update workload. Each database replica runs sequentially, at full bandwidth; mutual synchronisation between replicas remains minimal. Both our simulations and the experimental results obtained with our prototype show that Gargamel improves both response time and load by an order of magnitude when contention is high (highly loaded system with bounded resources), and that otherwise slow-down is negligible.

We have studied Gargamel's behavior while running over multiple geographically distant sites. One instance of Gargamel runs on each site, synchronizations among the different sites occur off the critical path [39]. Our experiments with the Amazon platform show that or solution can be used to support failures of whole sites.

5.3.9. Eventual consistency

Eventual Consistency (EC) aims to minimize synchronisation, by weakening the consistency model. The idea is to allow updates at different nodes to proceed without any synchronisation, and to propagate the updates asynchronously, in the hope that replicas converge once all nodes have received all updates. EC was invented for mobile/disconnected computing, where communication is impossible (or prohibitively costly). EC also appears very appealing in large-scale computing environments such as P2P and cloud computing. However, its apparent simplicity is deceptive; in particular, the general EC model exposes tentative values, conflict

resolution, and rollback to applications and users. Our research aims to better understand EC and to make it more accessible to developers.

We propose a new model, called *Strong Eventual Consistency* (SEC), which adds the guarantee that every update is durable and the application never observes a roll-back. SEC is ensured if all concurrent updates have a deterministic outcome. As a realization of SEC, we have also proposed the concept of a Conflict-free Replicated Data Type (CRDT). CRDTs represent a sweet spot in consistency design: they support concurrent updates, they ensure availability and fault tolerance, and they are scalable; yet they provide simple and understandable consistency guarantees.

This new model is suited to large-scale systems, such as P2P or cloud computing. For instance, we propose a "sequence" CRDT type called Treedoc that supports concurrent text editing at a large scale, e.g., for a wikipedia-style concurrent editing application. We designed a number of CRDTs such as counters (supporting concurrent increments and decrements), sets (adding and removing elements), graphs (adding and removing vertices and edges), and maps (adding, removing, and setting key-value pairs).

CRDTs are the main topic of the ConcoRDanT ANR project (Section 7.1.7) and the FP7 grant SyncFree (Section 7.2.1.1). After developing the SwiftCloud extreme-scale CRDT platform (see Section 4.3), we are currently developing a flexible cloud database called Antidote (see Section 4.4).

5.3.10. Lower bounds and optimality of CRDTs

CRDTs raise challenging research issues: What is the power of CRDTs? Are the sufficient conditions necessary? How to engineer interesting data types to be CRDTs? How to garbage collect obsolete state without synchronisation, and without violating the monotonic semi-lattice requirement? What are the upper and lower bounds of CRDTs?

We co-authored an innovative approach to these questions, published at Principles of Programming Languages (POPL) 2014 [25]. Geographically distributed systems often rely on replicated eventually consistent data stores to achieve availability and performance. To resolve conflicting updates at different replicas, researchers and practitioners have proposed specialized consistency protocols, called replicated data types, that implement objects such as registers, counters, sets or lists. Reasoning about replicated data types has however not been on par with comparable work on abstract data types and concurrent data types, lacking specifications, correctness proofs, and optimality results. To fill in this gap, we propose a framework for specifying replicated data types using relations over events and verifying their implementations using replication-aware simulations. We apply it to seven existing implementations of 4 data types with nontrivial conflictresolution strategies and optimizations (last-writer-wins register, counter, multi-value register and observed-remove set). We also present a novel technique for obtaining lower bounds on the worst-case space overhead of data type implementations and use it to prove optimality of four implementations. Finally, we show how to specify consistency of replicated stores with multiple objects axiomatically, in analogy to prior work on weak memory models. Overall, our work provides foundational reasoning tools to support research on replicated eventually consistent stores.

5.3.11. Explicit Consistency: Strengthening Eventual Consistency to support application invariants

The designers of the replication protocols for geo-replicated storage systems have to choose between either supporting low latency, eventually consistent operations, or supporting strong consistency for ensuring application correctness. We propose an alternative consistency model, *explicit consistency*, that strengthens eventual consistency with a guarantee to preserve specific invariants defined by the applications. Given these application-specific invariants, a system that supports explicit consistency must identify which operations are unsafe under concurrent execution, and help programmers to select either violation-avoidance or invariant-repair techniques. We show how to achieve the former while allowing most of operations to complete locally, by relying on a reservation system that moves replica coordination off the critical path of operation execution. The latter, in turn, allow operations to execute without restriction, and restore invariants by applying a repair operation to the database state. We designed and evaluated Indigo, a middleware that provides Explicit

Consistency on top of a causally-consistent data store. Indigo guarantees strong application invariants while providing latency similar to an eventually consistent system.

This work was presented at W-PSDS 2014 [24] and LADIS 2014 [38]. It was selected for presentation at EuroSys 2015 [23]. This research is supported in part by the FP7 grant SyncFree (Section 7.2.1.1).

5.4. Memory management for big data

Participants: Antoine Blin, Lokesh Gidra, Sébastien Monnet, Marc Shapiro, Julien Sopena [correspondent], Gaël Thomas.

5.4.1. Garbage collection for big data on large-memory NUMA machines

On contemporary cache-coherent Non-Uniform Memory Access (ccNUMA) architectures, applications with a large memory footprint suffer from the cost of the garbage collector (GC), because, as the GC scans the reference graph, it makes many remote memory accesses, saturating the interconnect between memory nodes. We address this problem with NumaGiC, a GC with a mostly-distributed design. In order to maximise memory access locality during collection, a GC thread avoids accessing a different memory node, instead notifying a remote GC thread with a message; nonetheless, NumaGiC avoids the drawbacks of a pure distributed design, which tends to decrease parallelism. We compared NumaGiC with Parallel Scavenge and NAPS on two different ccNUMA architectures running on the Hotspot Java Virtual Machine of OpenJDK 7. On Spark and Neo4j, two industry-strength analytics applications, with heap sizes ranging from 160 GB to 350 GB, and on SPECjbb2013 and SPECjbb2005, NumaGiC improves overall performance by up to 45% over NAPS (up to 94% over Parallel Scavenge), and increases the performance of the collector itself by up to 3.6× over NAPS (up to 5.4× over Parallel Scavenge).

This research is accepted for presentation at the ASPLOS 2015 conference [29].

5.4.2. 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. We propose Puma [35] (for Pooling Unused Memory in Virtual Machines) which allows I/O intensive applications running on top of VMs to benefit of large caches.

This is 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. Puma is a kernel level remote caching mechanism that is: (*i*) block device, file system and hypervisor agnostic; and (*ii*) efficient both locally and remotely. It can increase applications performance up to 3 times without impacting potential activity peaks.

SCALE Team

6. New Results

6.1. Programming Languages for Distributed Systems

One of the objectives of the Scale team is to design programming models easing the development and safe execution of distributed systems. This section describes our results in this direction.

6.1.1. Multi-active Objects

Participants: Ludovic Henrio, Fabrice Huet, 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 published the extension of multi-active objects to support scheduling and thread limitation [12].
- We developed a 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 [25], and is under submission to a conference.
- 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).
- Extensive use of multiactive objects in our CAN P2P network and implementation of usecases [2].
- We formalised in Isabelle/HOL a first version of the semantics of multiactive objects. This work was done in collaboration with Florian Kammuller

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

6.1.2. Autonomic Monitoring and Management of Components

Participants: Françoise Baude, Ludovic Henrio.

We have completed the design of a framework for autonomic monitoring and management of component-based applications. We have provided an implementation using GCM/ProActive taking advantage of the possibility of adding components in the membrane. The framework for autonomic computing allows the designer to describe in a separate way each phase of the MAPE autonomic control loop (Monitoring, Analysis, Planning, and Execution), and to plug them or unplug them dynamically.

- This year, we published a journal paper summarising our approach in the GCM/ProActive framework and our contribution on componentised membranes for autonomic computing [3].
- We also improved, in the context of the SCADA associate team and during the internship of Matias Ibañez, the support for autonomic components, providing all the architecture and API so that the programmer of autonomic aspect can do them in a DSL reconfiguration language, called GCMScript. This was implemented and experimented, a publication is under submission on this work.

6.1.3. Algorithmic skeletons

Participant: Ludovic Henrio.

In the context of the SCADA associated team, we worked on the algorithmic skeleton programming model. The structured parallelism approach (skeletons) takes advantage of common patterns used in parallel and distributed applications. The skeleton paradigm separates concerns: the distribution aspect can be considered separately from the functional aspect of an application. In the previous year we designed the possibility for a skeleton to output events, which increases the control and monitoring capabilities. This year we published our previous results in [14] and realised additional steps:

- Study of different ways to predict the execution time for a skeleton, inspired from simple statistic functions. This improvement together with the distributed execution of skeletons should allow us to publish a journal paper on this subject in 2015

6.1.4. Optimization of data transfer in event-based programming models

Participants: Iyad Alshabani, Françoise Baude, Laurent Pellegrino.

In [6], we extended a previous work with conceptual and experimental performance evaluations. This previous and collaborative work [1] developed an innovative approach of "lazy copy and transfer" of the data parts of event objects exchanged by peers in the context of event-driven architecture applications.

While event notifications are routed in a conventional manner through an event service, data parts of the events are directly and transparently transferred from publishers to subscribers. The theoretical analysis shows that we can reduce the average event delivery time by half, compared to a conventional approach requiring the full mediation of the event service. The experimental analysis confirms that the proposed approach outperforms the conventional one (both for throughput and delivery time) even though the middleware overhead, introduced by the specific adopted model, slightly reduces the expected benefits.

6.1.5. Behavioural Semantics

Participants: Ludovic Henrio, Eric Madelaine, Min Zhang.

We have studied Parameterised Networks of Automata (pNets) from a theoretical perspective. We started 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. Our framework can model full systems, using (closed) hierarchies of pNets; we can also build (open) pNet systems expressing composition operators. Concerning more fundamental aspects, we defined a strong bisimulation theory specifically for the pNet model, proved its properties, and illustrated it on some examples. One of the original aspects of the approach is to relate the compositional nature of pNets with the notion of bisimulation; this was exemplified by studying the properties of a flattening operator for pNets. This work has been accepted for publication at PDP'2015 ([24]).

6.1.6. A Time-sensitive Heterogeneous Behavioural Model

Participants: Eric Madelaine, Yanwen Chen.

This work concludes the PhD research of Yanwen Chen, targetting a timed-sensitive extension of the pNets model with logical clocks inherited from the CCSL language. The main results of this year are: 1) a new notion of Time Specification (TS), used to handle the abstract properties of each level of processes in a pNet structure, 2) algorithms to compute such TSs for basic parameterized and timed processes, and from composition of timed-pNets, 3) conditions for checking the compatibility of composition, 4) a use-case from the area of intelligent transportation systems, illustrating the whole chain of modeling, upto a symbolic simulation of the full composed system, with the TimeSquare tool. This work was published as [4], [23], and in the PhD thesis of Y. Chen, defended on 2014, Nov. 30th.

6.1.7. Structure and structural correctness for GCM components

Participants: Ludovic Henrio, Oleksandra Kulankhina, Eric Madelaine.

We have defined a set of rules characterizing the well-formed composition of components in order to guarantee their safe deployment and execution. This work focuses on the structural aspects of component composition; it puts together most of the concepts common to many component models, but never formalized as a whole. Our formalization characterizes correct component architectures made of functional and non-functional aspects, both structured as component assemblies. So-called 'Interceptor chains' can be used for a safe and controlled interaction between the two aspects. Our well-formed components guarantee a set of properties ensuring that the deployed component system has a correct architecture and can run safely. Those definitions constitute the formal basis for VerCors tool. This work was done in the context of O. Kulankhina phd research, and in collaboration with Dongqian Liu (ECNU Shanghai), as part of the Associated Team DAESD.

6.2. Run-time/middle-ware level

6.2.1. Scalable and robust Middleware for distributed event based computing

Participants: Françoise Baude, Fabrice Huet, Laurent Pellegrino, Maeva Antoine.

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. This has been the core content of Laurent Pellegrino PhD thesis [2], and the corresponding software deposit at the APP for this middleware.

As a distributed system, this middleware can suffer from failures. To resist to such situations, we have added a capability of checkpointing. In [18] we present how to design an adaptation of the famous Chandy and Lamport algorithm for distributed snapshot taking, to the case of the Event Cloud. Indeed, as the Event Cloud peers are multi-active objects, we need to take care when and how to serve the chekpointing request and so, when to apply the Chandy Lamport protocol operations. Consequently, we have make sure that the obtained distributed snapshot is indeed consistent. As publication of events are triggered from the outside of the Event Cloud, we however are not able to recover them from the last saved snapshot in case of peer crash and subsequent whole Event Cloud recovery. However, we ensure any event injected through a peer, before this peer was participating in the last global checkpoint taking is safely part of it.

As a distributed system handling huge amount of information, this middleware can suffer from data imbalances. In [22], [8], we have reviewed the litterature of structured peer to peer systems regarding the way they handle load imbalance. 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.

Storing highly skewed data in a distributed system has become a very frequent issue, in particular with the emergence of semantic web and big data. This often leads to biased data dissemination among nodes. Addressing load imbalance is necessary, especially to minimize response time and avoid workload being handled by only one or few nodes. We have proposed a protocol which allows a peer to change its hash function at runtime, without a priori knowledge regarding data distribution. This provides a simple but efficient adaptive load balancing mechanism. Moreover, we have shown that a structured overlay can still be consistent event when all peers do not apply the same hash function on data [7].

6.2.2. Virtual Machines Placement Algorithms

Participants: Fabien Hermenier, Vincent Kherbache.

In [21], [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. In [5] we rely on BtrPlace to achieve energy efficiency. To maintain an energy footprint as low as possible, data centres manage their VMs according to conventional and established rules. Each data centre is however made unique due to its hardware and workload specificities. This prevents the *ad-hoc* design of current VM schedulers from taking these particularities into account to provide additional energy savings. In this paper, we present Plug4Green, an application that relies on BtrPlace to customize an energy-aware VM scheduler. This flexibility is validated through the implementation of 23 SLA constraints and 2 objectives aiming at reducing either the power

consumption or the greenhouse gas emissions. On a heterogeneous test bed, Plug4Green specialization to fit the hardware and the workload specificities allowed to reduce the energy consumption and the gas emission by up to 33% and 34%, respectively. Finally, simulations showed that Plug4Green is capable of computing an improved placement for 7,500 VMs running on 1,500 servers within a minute.

Finally, we started to investigate on easing the jobs of data centre operators using BtrPlace. For example, server maintenance is a common but still critical operation. A prerequisite is indeed to relocate elsewhere the VMs running on the production servers to prepare them for the maintenance. When the maintenance focuses several servers, this may lead to a costly relocation of several VMs so the migration plan must be chose wisely. This however implies to master numerous human, technical, and economical aspects that play a role in the design of a quality migration plan. In[13], we study migration plans that can be decided by an operator to prepare for an hardware upgrade or a server refresh on multiple servers. We exhibit performance bottleneck and pitfalls that reduce the plan efficiency. We then discuss and validate possible improvements deduced from the knowledge of the environment peculiarities.

6.3. Application level

6.3.1. Simulation Software Architecture

Participant: Olivier Dalle.

In general purpose software engineering (as opposed to simulation software engineering), the motivations for reuse have long been advocated and demonstrated: lower risks of defects, collective support of potentially larger user community, lower development costs, and so on. In simulation software architectures, we can also cite business-specific motivations, such as providing a better reproducibility of simulation experiments, or avoiding a complex validation process. In [20], we show that although it is rarely discussed, reuse is a problem that may be considered in two opposite directions: reusing and being reused.

6.3.2. 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 [16] we propose a new data type for discretized time representation in DEVS, based on rational numbers. Indeed, we show that rational numbers offer good stability properties for the arithmetics used in DEVS, with a limited impact on the simulation execution performance.

6.3.3. GPU-based High Performance Cloud Computing

Participants: Michael Benguigui, Françoise Baude, Fabrice Huet.

To address HPC, GPU devices are now considered as unavoidable cheap, energy efficient, and very efficient alternative computing units. Our long term goal is to devise some generic solutions in order to incorporate GPU-specific code whenever relevant into a parallel and distributed computation.

As a challenging example, we have pursued our work on pricing American multi-dimensional (so very computation intensive) options in finance. From our previous work that achieved pricing a 40-assets based american option within 8 hours of computation on a single GPU, the work in [9] allows us to reach approximatively one hour of computation time. For this, we run using active objects coupled with OpenCL codes, on 18 GPU nodes acquired from the Grid'5000 platform (the maximum amount of available GPU on Grid'5000 that we could book at once).

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. This parallel and distributed pricing approach is also extended in the forthcoming PhD thesis of Michael Benguigui: it successfully tackles the Value At Risk computation of a portofolio composed of such complex financial products.

6.3.4. 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 [17], 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 Team

6. New Results

6.1. Highlights of the Year

In 2014, we are proud to have organized the 17th ACM SIGSOFT International Conference on Component-Based Software Engineering and Software Architecture (CompArch) that has been held in Lille from 30 June to 3 July 2014.

CompArch is the main conference of the ACM SIGSOFT group on software architectures and software components. The conference is held alternatively in North America and in Europe. The 17th edition has been held this year in France for the first time. The conference brings together about 100 researchers from the academia and the industry.

6.2. Distributed Context Monitoring

In 2014, we obtained some new results in the area of distributed context monitoring solutions to support the development of self-optimising software systems. Context monitoring has emerged as a key capability in various domains to connect software systems to the underlying hardware platform or to the physical world (in the case of ubiquitous systems). In particular, we have investigated to the capability of inferring high-level contextual situations from a large volume of raw data collected from a single device or in the wild. Both hardware (*e.g.*, accelerometer) or software (*e.g.*, performance counters) sensors tend to continuously produce raw data that a context monitoring solution has to quickly filter, process, and convert it into information that can be used by an application or understood by a user.

As a result of the PhD thesis of Adel Noureddine [14], defended in March 2014, we have developed a middleware toolkit to support *in-depth context monitoring* in the domain of green computing. In particular, we introduce a software library, named POWERAPI, that can estimate the power consumption in real-time at various granularities of software: from system processes to code methods (see Section 5.3). This non-invasive solution provides accurate insights on energy hotspots of software and can be used to derive the energy profile of any software library, thus guiding the developers in optimising the energy consumption of their developments.

As a result of the PhD thesis of Nicolas Haderer [12], defended in November 2014, we have contributed to the development of a middleware platform to support *in-breadth context monitoring* in the area of mobile computing. In particular, we promote the distributed middleware solution APISENSE® as an efficient approach to deploy mobile crowd-sensing tasks across a large population of volunteer participants (see Section 5.1). In particular, APISENSE® includes a task orchestration algorithm that preserves the privacy and the battery of sensing devices, while maintaining specific sensing coverage objectives (including time and space dimensions). The server-side infrastructure of APISENSE® is generated from a dedicated software product line, while the implementation is based on the FRASCATI platform (see Section 5.2).

6.3. Design and Runtime Support for Cloud Computing

In 2014, we obtained some new results in the domain both of the design and the runtime support of distributed applications for multi-cloud systems. The purpose is to deal with applications that span across several different cloud systems. Several reasons justify such a goal. For example, in order to avoid the so-called vendor lock-in syndrome, cloud application stakeholders need to be able to migrate as easily as possible their assets from one cloud system to another one. Other examples include the possibility of introducing diversify and fault-tolerance by deploying applications on different cloud systems, or hot migrating applications where computing resources are less expensive.

For the design of multi-cloud systems, we proposed a solution based on software product lines (SPL) [90] and ontologies. In order to specify the variability of such environments, we extended SPL with attributes, cardinalities, and constraints. In order to enable the evolution of these environments, we provided an automated support for maintaining the consistency based on constraint programming. Finally, we proposed an ontology based approach to bridge the gap between the concepts and artefacts defined by different cloud systems. This global solution is the result of the PhD thesis of Clément Quinton [16] that was defended in October 2014, and has been partially supported by the FP7 PaaSage project (see Section 8.3).

For the runtime support of multi cloud systems, we proposed the SOCLOUD platform. This solution enables to deploy, execute and manage an application that spans on several different cloud systems. SOCLOUD tackles the challenges of portability, provisioning, elasticity, and high availability. SOCLOUD defines a component-based and service-oriented architecture that provides an unified view of a set of cloud systems. SOCLOUD is the result of the PhD thesis of Fawaz Paraiso [15] that was defended in June 2014. SOCLOUD is implemented on top of the FRASCATI platform (see Section 5.2).

6.4. Extraction and Analysis of Knowledge for Automatic Software Repair

Automated software repair aims at assisting developers in order to improve the quality of software systems, for example by recommending some repair actions to fix bugs. Matias Martinez has presented in his PhD thesis [13] that was defended in June 2014, new results in this domain. These results aim at reducing the search space when repairing a software system. The solution relies on two techniques. The first one consists in building change models learnt from repairs performed by other developers. These repairs are mined from existing software repositories of open source projects, and analysed based on their types and frequencies. The second proposed technique is based on the inherent redundancy of code patterns. The assumption is that the probability that the repair code for a particular kind of defect is already present in the software system under study is high. We then take advantage of this inherent redundancy to reduce the search space when looking for repair actions.

WHISPER Team

6. New Results

6.1. Highlights of the Year

The paper "Faults in Linux 2.6" was published in the ACM journal Transactions on Computer Systems in June 2014 . It has been downloaded from the ACM digital library almost 300 times since then. The paper was reviewed in the Linux Weekly News, in the German professional IT website golem.de, and was the subject of an invited presentation at a joint session of the Linux Kernel Summit and LinuxCon North America.

Julia Lawall was invited to the 2014 Linux Kernel Summit, an invitation-only meeting of core Linux developers. She was subsequently invited to participate in the plenary Linux Kernel Developer Panel at LinuxCon Europe, with 2000 attendees.

Julia Lawall was invited to give a keynote at the conference Modularity (formerly AOSD) on her work on Coccinelle [16].

BEST PAPERS AWARDS :

[] **ACM Transactions on Computer Systems**. N. PALIX, G. THOMAS, S. SAHA, C. CALVÈS, G. MULLER, J. L. LAWALL.

6.2. Lock profiling in Java servers

Today, Java is regularly used to implement large multi-threaded server-class applications that use locks to protect access to shared data. However, understanding the impact of locks on the performance of a system is complex, and thus the use of locks can impede the progress of threads on configurations that were not anticipated by the developer, during specific phases of the execution. In our paper, "Continuously Measuring Critical Section Pressure with the Free-Lunch Profiler" [25], presented at OOPSLA 2014, we propose Free Lunch, a new lock profiler for Java application servers, specifically designed to identify, *in-vivo*, phases where the progress of the threads is impeded by a lock. Free Lunch is designed around a new metric, *critical section pressure* (CSP), which directly correlates the progress of the threads to each of the locks. Using Free Lunch, we have identified phases of high CSP, which were hidden with other lock profilers, in the distributed Cassandra NoSQL database and in several applications from the DaCapo 9.12, the SPECjvm2008 and the SPECjbb2005 benchmark suites. Our evaluation of Free Lunch shows that its overhead is never greater than 6%, making it suitable for *in-vivo* use.

6.3. Software engineering for infrastructure software

A kernel oops is an error report that logs the status of the Linux kernel at the time of a crash. Such a report can provide valuable first-hand information for a Linux kernel maintainer to conduct postmortem debugging. Recently, a repository has been created that systematically collects kernel oopses from Linux users. However, debugging based on only the information in a kernel oops is difficult. In a paper published at MSR [18], we consider the initial problem of finding the offending line, i.e., the line of source code that incurs the crash. For this, we propose a novel algorithm based on approximate sequence matching, as used in bioinformatics, to automatically pinpoint the offending line based on information about nearby machine-code instructions, as found in a kernel oops. Our algorithm achieves 92% accuracy compared to 26% for the traditional approach of using only the oops instruction pointer.

2014 was the second year of a two-year cooperation between Julia Lawall and David Lo of Singapore Management University, as part of the Merlion cooperation grant program of the Insitut Français. This cooperation resulted in four papers: two on word similarity [21], [26], one on bug localization [23], and one on an empirical study of testing practices in open source software [19]. As an offshoot of this work, Julia Lawall worked with the PhD student Ripon Saha of UT Austin and his advisors on the topic of assessing the effectiveness of a state-of-the-art bug localization technique on C programs as compared to Java programs [20]. This work built on the C parser developed for Coccinelle.

Finally, with colleagues from Aalborg University and with Nicolas Palix of Grenoble, Julia Lawall published an article in Science of Computer Programming assessing the applicability of Coccinelle to checking the coding style guidelines of the CERT C Secure Coding Standard [14].

6.4. Bugs in Linux 2.6

In August 2011, Linux entered its third decade. Ten years before, Chou et al. published a study of faults found by applying a static analyzer to Linux versions 1.0 through 2.4.1. A major result of their work was that the drivers directory contained up to 7 times more of certain kinds of faults than other directories. This result inspired numerous efforts on improving the reliability of driver code. Today, Linux is used in a wider range of environments, provides a wider range of services, and has adopted a new development and release model. What has been the impact of these changes on code quality? To answer this question, in an article published in ACM TOCS, we have transported Chou et al.'s experiments to all versions of Linux 2.6; released between 2003 and 2011. We find that Linux has more than doubled in size during this period, but the number of faults per line of code has been decreasing. Moreover, the fault rate of drivers is now below that of other directories, such as arch. These results can guide further development and research efforts for the decade to come. To allow updating these results as Linux evolves, we define our experimental protocol and make our checkers available.

6.5. Memory Monitoring in Smart Home gateways

Smart Home market players aim to deploy component-based and service-oriented applications from untrusted third party providers on a single OSGi execution environment. This creates the risk of resource abuse by buggy and malicious applications, which raises the need for resource monitoring mechanisms. Existing resource monitoring solutions either are too intrusive or fail to identify the relevant resource consumer in numerous multi-tenant situations. In our paper "Memory Monitoring in a Multi-tenant OSGi Execution Environment" [15], presented at CBSE 2014, we propose a system to monitor the memory consumed by each tenant, while allowing them to continue communicating directly to render services. We propose a solution based on a list of configurable resource accounting rules between tenants, which is far less intrusive than existing OSGi monitoring systems. We modified an experimental Java Virtual Machine in order to provide the memory monitoring features for the multi-tenant OSGi environment. Our evaluation of the memory monitoring mechanism on the DaCapo benchmarks shows an overhead below 46%. This work has been done as part of the PhD of Koutheir Attouchi [10] who was supported by a CIFRE grant with Orange Labs.

ALGORILLE Project-Team

6. New Results

6.1. Structuring applications for scalability

6.1.1. Combining locking and data management interfaces

Participants: Jens Gustedt, Mariem Saied.

Handling data consistency in parallel and distributed settings is a challenging task, in particular if we want to allow for an easy to handle asynchronism between tasks. Our publication [4] shows how to produce deadlock-free iterative programs that implement strong overlapping between communication, IO and computation.

A new implementation (ORWL) of our ideas of combining control and data management in C has been undertaken, see 5.2.1 . In 2014, work has demonstrated its efficiency for a large variety of platforms, see [20]. By using the example of dense matrix multiplication, we show that ORWL permits to reuse existing code for the target architecture, namely open source library ATLAS, Intel's compiler specific MKL library or NVidia's CUBLAS library for GPUs. ORWL assembles local calls into these libraries into efficient functional code, that combines computation on distributed nodes with efficient multi-core and accelerator parallelism.

Our next efforts will concentrate on the continuation of an implementation of a complete application (an American Option Pricer) that was chosen because it presents a non-trivial data transfer and control between different compute nodes and their GPU. ORWL is able to handle such an application seamlessly and efficiently, a real alternative to home made interactions between MPI and CUDA.

6.2. Experimental methodologies for the evaluation of distributed systems

6.2.1. Simulation and dynamic verification

6.2.1.1. SimGrid framework improvement

Participants: Paul Bédaride, Martin Quinson, Gabriel Corona.

On the technical side, we kept up with our regular releases of the SimGrid framework, integrating the work of our partners in the SONGS ANR project. This year, we reimplemented the simulation kernel in C++. This modularity improvement will ease the addition of performance models by external contributors. This work thus contributes to our overall goal of constituting a user community focused on this first-class tool.

[11] is a long awaited paper describing the current state of the project and its future roadmap. This constitutes the new reference paper on the SimGrid project (the previous article, a short paper from 2008, was cited over 350 times since its publication). We show that despite the common beliefs, the tool specialization is not necessarily a warrant for performance and correctness.

We also continued our animation of our scientific community, for example through our participation to the Joint Laboratory for Petascale Computing (Inria/ANL/UIUC/BSC). We co-organized a summer school on Performance Metrics, Modeling and Simulation of Large HPC Systems in June, to push our tools toward PhD students that need to assess their HPC applications.

6.2.1.2. Dynamic verification and SimGrid

Participants: Marion Guthmuller, Martin Quinson, Gabriel Corona.

This year, the PhD thesis of M. Guthmuller went into its third year. The proposed methodology maturated into a usable tool: we can now verify small-size real HPC applications using MPI in C/C++/Fortran. This relies on a heuristic exploration of the applicative state at the system level that was presented in [21], [22].

Also, we finally added the ability to dynamically verify some CTL properties over MPI implementations. SimGrid was one of the rare framework able to verify LTL liveness properties over real implementations. To the best of our knowledge, it becomes the very first tool verifying CTL properties on real C/C++/Fortran applications. The targeted properties quantify the stability of the applicative communication pattern. The applications that respect these properties can benefit from specific, more efficient, fault tolerance algorithms. Verifying these properties is thus of a major practical interest. A publication is in preparation, as well as the PhD manuscript of M. Guthmuller who will defend by 2015 Q1.

6.2.2. Experimentation on testbeds and production facilities, emulation

6.2.2.1. Evaluating load balancing and fault tolerance strategies on Distem

Participants: Joseph Emeras, Emmanuel Jeanvoine, Lucas Nussbaum.

(For context, see sections 3.3 and 5.4 .)

We extended our work [27] to enable the study of load balancing and fault tolerance strategies on Distem. Distem now supports the introduction of changing heterogeneity and imbalance among virtual nodes, as well as the introduction of failures. Two HPC runtimes targeting Exascale (Charm++ and OpenMPI) were used as target applications. This work was presented at the Joint Laboratory for Extreme-Scale Computing in June, and at the Grid'5000 Spring School. However, those results still have to be properly published.

6.2.2.2. Distem improvements: VXLAN, release and tutorial

Participants: Emmanuel Jeanvoine, Tomasz Buchert, Lucas Nussbaum.

(For context, see sections 3.3 and 5.4 .)

The scalability of Distem's networking layer was improved by adding support for VXLAN networks. This enabled experiments with up to 40,000 virtual nodes, presented at the CCGrid'2014 SCALE challenge (where we were selected as finalist) [17]. Version 1.0 of Distem was also released in March 2014, and featured in a tutorial at the Grid'5000 Spring School.

6.2.2.3. Kadeploy improvements: REST API, new image broadcast mechanism

Participants: Luc Sarzyniec, Stéphane Martin, Emmanuel Jeanvoine, Lucas Nussbaum.

(For context, see sections 3.3 and 5.4 .)

Kadeploy 3.2 was released in March 2014. Among many other changes, that release included a new REST API to interact with Kadeploy, replacing the old Ruby-specific RPC mechanism, and easing the automation of experiments by providing a way to call Kadeploy from scripts.

Kadeploy 3.3 was released in November 2014. This release is mostly a bug-fix release, with many bug fixes in the internal cache system, the shell runner, and others.

We also implemented an improved mechanism to broadcast machine images to nodes. The new tool, called Kascade, is fault tolerant, and its performance has been thoroughly tested. It was described in a publication accepted at HPDIC'2014 [24], included in Kadeploy 3.2, and used as the default method for environment broadcast since Kadeploy 3.3.

6.2.2.4. XPFlow

Participants: Tomasz Buchert, Stéphane Martin, Emmanuel Jeanvoine, Lucas Nussbaum, Jens Gustedt.

(For context, see sections 3.3 and 5.7 .)

A publication focusing on XPFlow was accepted at CCGrid'2014 [18], and XPFlow was also featured in a tutorial at Grid'5000 Spring School. Our ongoing work focuses on improved support for collecting provenance in XPFlow.

6.2.2.5. Survey of Experiment Management tools

Participants: Tomasz Buchert, Cristian Ruiz, Lucas Nussbaum.

We produced a survey of Experiment Management tools for distributed systems, published in Future Generation Computer Systems [10]. This survey provides an extensive list of features offered by general-purpose experiment management tools dedicated to distributed systems research on real platforms. It then uses it to assess existing solutions and compare them, outlining possible future paths for improvements.

6.2.2.6. Grid'5000

Participants: Émile Morel, Luc Sarzyniec, Lucas Nussbaum.

(For context, see sections 3.3 and 5.8 .)

The work on resources description, selection, reservation and verification was wrapped-up in a TridentCom'2014 paper [23].

As a member of the Grid'5000 architects committee, Lucas Nussbaum was involved in the submission (and acceptance) of ADT Laplace.

Lucas Nussbaum also presented a talk [12] on Reproducible Research and Grid'5000 at the Grid'5000 evaluation by the Scientific Committee, during the Spring School.

6.2.3. Convergence and co-design of experimental methodologies

6.2.3.1. Realis'2014

Participant: Lucas Nussbaum.

Lucas Nussbaum organized (with Olivier Richard) the second edition of the Realis event [14]. Associated to the Compas'14 conference, this workshop aimed at providing a place to discuss the reproducibility of the experiments underlying the publications submitted to the main conference. We hope that this kind of venue will motivate the researchers to further detail their experimental methodology, ultimately allowing others to reproduce their experiments.

6.2.3.2. Reproducible Research working group at Inria Nancy – Grand Est

Participant: Lucas Nussbaum.

Lucas Nussbaum is organizing a working group on Reproducible Research at Inria Nancy – Grand Est since May 2014. Meetings involve a dozen of members from many different teams, and discussion topics have so far covered online platforms to test algorithms and applications, and evaluation contests organized together with conferences and workshops.

Lucas Nussbaum has also been invited to participate in the Inria national initiative on reproducible research.

6.2.3.3. Organization of Reppar

Participant: Lucas Nussbaum.

Lucas Nussbaum co-organized the first edition of the Reppar workshop, held during Europar'2014, with a focus on experimental practices in parallel computing research.

6.3. Algorithmic schemes for efficient use of parallel devices in clusters

Participants: Sylvain Contassot-Vivier, Stéphane Vialle [External collaborator, SUPELEC].

During the year 2014, we have continued our studies about the design and implementation of efficient algorithmic schemes to fully exploit all the available computational resources inside a parallel system. In particular, we have proposed general schemes that optimize the use of GPUs in clusters [26]. This is achieved by performing two kinds of overlappings. The former corresponds to computation/communication overlappings, either for the communications between machines but also for the data transfers between central RAM and GPUs inside each machine. The latter is the computation/computation overlapping that consists in executing computations on the GPUs in parallel of some computations on the central CPUs. Moreover, in this work we have paid a particular attention to some important aspects of software engineering that are the development and maintenance costs. Those aspects are essential as they directly determine the practical usability of the schemes, especially in the industry where there is a permanent vigilance to minimize the associated costs.

6.4. Parallel schemes for the resolution of the RTE with finite volumes method

Participant: Sylvain Contassot-Vivier.

In the context of our collaboration with the Lemta laboratory (Fatmir Asllanaj), about the design and implementation of an efficient and high accuracy algorithm for solving the Radiative Transfer Equation (RTE), we have reached our second objective that consisted in the realization of a multi-threaded parallel version of the software. That new version is based on the optimized sequential version produced as a first objective. It makes use of the OpenMP library to exploit all the cores inside one machine. The results are very satisfying as our algorithm obtains very good speed up and efficiency (around 90% and above) in realistic contexts. Moreover, besides this work over performance, we focus also on the high quality (accuracy) of the results of our software by making a permanent effort to track any possible enhancement of our numerical scheme. Then, the actual implementation of each of these possible enhancements is considered according to its potential costs, either in performance degradation as well as in additional resource consumptions (CPUs, GPUs and RAM). Confrontations to other existing computational schemes to solve the RTE are regularly realized to corroborate the validity preservation of our software [9], [15].

6.5. Study of binary multiplication and dynamical approaches to the integer factorization

Participants: Sylvain Contassot-Vivier, Nazim Fatès.

In the context of a collaboration with Nazim Fatès over dynamical systems we have co-supervised the internship of Raphaël Rieu-Helft (student at the ENS Paris), during June and July 2014. The goal of this internship was to study the relevance of the dynamical systems formalism as an efficient way to express and solve two specific problems. The former one was the queens problem on chessboards of arbitrary size. This goal was to express a solving algorithm of the queens problem under the form of a cellular automaton. The second step was to extend the results obtained for the queens problem to a more complex and computationally expensive problem that is the integer factorization. Two dynamical systems (cellular automata) have been obtained for both problems and their respective efficiencies, either in terms of convergence speed or speed of solution reaching, have been experimentally evaluated.

ALPINES Project-Team

6. New Results

6.1. Highlights of the Year

We have released a version of FreeFem++ (v 3.33) which introduces new and important features related to high performance computing:

- Interface with PETSc library
- Interface with HPDDM (see above)
- improved interface with the parallel direct solver MUMPS

This release enables, for the first time, end-users to run the very same code on computers ranging from laptops to clusters and even large scale computers with thousands of computing nodes

6.2. 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 that reduce communication on multilevel hierarhical platforms.

In [17] we focus on the QR factorization. The Tall-Skinny QR (TSQR) algorithm is more communication efficient than the standard Householder algorithm for QR decomposition of matrices with many more rows than columns. 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. As a result, our final parallel QR algorithm outperforms ScaLAPACK and Elemental implementations of Householder QR and our implementation of CAQR on the Hopper Cray XE6 NERSC system.

In [18] we focus on performance predictions of multilevel communication optimal LU and QR factorizations on hierarchical platforms. This study focuses on the performance of two classical dense linear algebra algorithms, the LU and the QR factorizations, on multilevel hierarchical platforms. We first introduce a new model called Hierarchical Cluster Platform (HCP), encapsulating the characteristics of such platforms. The focus is set on reducing the communication requirements of studied algorithms at each level of the hierarchy. Lower bounds on communications are therefore extended with respect to the HCP model. We then introduce multilevel LU and QR algorithms tailored for those platforms, and provide a detailed performance analysis. We also provide a set of numerical experiments and performance predictions demonstrating the need for such algorithms on large platforms.

6.3. Enlarged Krylov methods

Krylov subspace methods are among the most practical and popular iterative methods today. They are polynomial iterative methods that aim to solve systems of linear equations ($Ax = b$) by finding a sequence of vectors $x_1, x_2, x_3, x_4, ..., x_k$ that minimizes some measure of error over the corresponding spaces $x_0 + \mathcal{K}_i(A, r_0), \quad i = 1, ..., k$ where $\mathcal{K}_i(A, r_0) = span\{r_0, Ar_0, A^2r_0, ..., A^{i-1}r_0\}$ is the Krylov subspace of dimension i, x_0 is the initial iterate, and r_0 is the initial residual. These methods are governed by Blas1 and

Blas2 operations as dot products and sparse matrix vector multiplications. Parallelizing dot products is constrained by communication since the performed computation is negligible. If the dot products are performed by one processor, then there is a need for a communication before and after the computation. In both cases, communication is a bottleneck. In [21] we introduce a new approach for reducing communication in Krylov subspace methods that consists of enlarging the Krylov subspace by a maximum of t vectors per iteration, based on the domain decomposition of the graph of A. The obtained enlarged Krylov subspace $\mathcal{K}_{t,k}(A, r_0)$ is a superset of the Krylov subspace $\mathcal{K}_k(A, r_0)$, $\mathcal{K}_k(A, r_0) \subset \mathcal{K}_{t,k+1}(A, r_0)$. Thus it is possible to search for the solution of the system $Ax = b$ in $\mathcal{K}_{t,k}(A, r_0)$ instead of $\mathcal{K}_k(A, r_0)$. Moreover, we show that the enlarged Krylov projection subspace methods lead to faster convergence in terms of iterations and parallelizable algorithms with less communication, with respect to Krylov methods.

6.4. Algebraic preconditioners

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

In [8] we introduce the block filtering decomposition, a new preconditioning technique that is suitable for matrices arising from the discretization of a system of PDEs on unstructured grids. The preconditioner satisfies a so-called filtering property, which ensures that the input matrix is identical with the preconditioner on a given filtering vector. This vector is chosen to alleviate the effect of low frequency modes on convergence and so decrease or eliminate the plateau which is often observed in the convergence of iterative methods. In particular, the paper presents a general approach that allows to ensure that the filtering condition is satisfied in a matrix decomposition. The input matrix can have an arbitrary sparse structure. Hence, it can be reordered using nested dissection, to allow a parallel computation of the preconditioner and of the iterative process. We present experimental results that demonstrate the efficiency of the proposed preconditioner on a set of matrices arising from the discretization of partial differential equations on two-dimensional and three-dimensional grids. We also show that the numerical efficiency of the preconditioner does not suffer from the reordering of the unknowns for the matrices in our test set, which can have highly heterogeneous and anisotropic coefficients.

In [9] we discuss the usage of overlapping techniques for improving the convergence of preconditioners based on incomplete factorizations. To enable parallelism, these preconditioners are usually applied after the input matrix is permuted into nested bordered block diagonal form. We use k-way partitioning with vertex separator (KPVS) to recursively partitions the corresponding graph of the input matrix into k subgraphs using a subset of its vertices called separators. In the case where $k = 2$, it is called nested dissection. The overlapping technique is then based on algebraically extending the associated subdomains of these subgraphs and their corresponding separators obtained from KPVS by their direct neighbours. This approach is known to accelerate the convergence of domain decomposition methods, where the input matrix is partitioned into a number of independent subdomains using k-way graph partitioning, a different graph decomposition technique. We discuss the effect of the overlapping technique on the convergence of two classes of preconditioners, based on nested factorization and block incomplete LDU factorization.

In [22] we introduce LORASC, a robust algebraic preconditioner for solving sparse linear systems of equations involving symmetric and positive definite matrices. The graph of the input matrix is partitioned by using k-way partitioning with vertex separators into N disjoint domains and a separator formed by the vertices connecting the N domains. The obtained permuted matrix has a block arrow structure. The preconditioner relies on the Cholesky factorization of the first N diagonal blocks and on approximating the Schur complement corresponding to the separator block. The approximation of the Schur complement involves the factorization of the last diagonal block and a low rank correction obtained by solving a generalized eigenvalue problem or a randomized algorithm. The preconditioner can be build and applied in parallel. Numerical results on a set of matrices arising from the discretization by the finite element method of linear elasticity models illustrate the robusteness and the efficiency of our preconditioner.

The Helmholtz equation governing wave propagation and scattering phenomena is difficult to solve numerically. Its discretization with piecewise linear finite elements results in typically large linear systems of equations. The inherently parallel domain decomposition methods constitute hence a promising class of precondi-

tioners. An essential element of these methods is a good coarse space. Here, the Helmholtz equation presents a particular challenge, as even slight deviations from the optimal choice can be devastating.

In [5], we present a coarse space that is based on local eigenproblems involving the Dirichlet-to-Neumann operator. Our construction is completely automatic, ensuring good convergence rates without the need for parameter tuning. Moreover, it naturally respects local variations in the wave number and is hence suited also for heterogeneous Helmholtz problems. The resulting method is parallel by design and its efficiency is demonstrated on 2D homogeneous and heterogeneous numerical examples.

Coarse spaces are instrumental in obtaining scalability for domain decomposition methods for partial differential equations (PDEs). However, it is known that most popular choices of coarse spaces perform rather weakly in the presence of heterogeneities in the PDE coefficients, especially for systems of PDEs. In [12], we introduce in a variational setting a new coarse space that is robust even when there are such heterogeneities. We achieve this by solving local generalized eigenvalue problems in the overlaps of subdomains that isolate the terms responsible for slow convergence. We prove a general theoretical result that rigorously establishes the robustness of the new coarse space and give some numerical examples on two and three dimensional heterogeneous PDEs and systems of PDEs that confirm this property.

Multiphase, compositional porous media flow models lead to the solution of highly heterogeneous systems of Partial Differential Equations (PDEs). In [7] , we focus on overlapping Schwarz type methods on parallel computers and on multiscale methods. We recall a coarse space that is robust even when there are such heterogeneities. The two-level domain decomposition approach is compared to multiscale methods.

In [16], we investigate two-level preconditioners on the extended linear system arising from the domain decomposition method. The additive Schwarz method is used as a smoother, and the coarse grid space is constructed by using the Ritz vectors obtained in the Arnoldi process. The coarse grid space can be improved adaptively as the Ritz vectors become a better approximation of the eigenvectors. Numerical tests on the model problem demonstrate the efficiency.

6.5. New results related to FreeFem++

In [10], we propose an efficient algorithm for the numerical approximation of metrics, used for anisotropic mesh adaptation on triangular meshes with finite element computations. We derive the metrics from interpolation error estimates expressed in terms of higher order derivatives, for the $P-k$-Lagrange finite element, $k>1$. Numerical examples of mesh adaptation done using metrics computed with our Algorithm, and derived from higher order derivatives as error estimates, show that we obtain the right directions of anisotropy.

In [2], we consider a system of two reaction-dispersion equations with non constant parameters. Both equations are coupled through the boundary conditions. We propose a mixed variational formulation that leads to a non symmetric saddle-point problem. We prove its well-posedness. Then, we develop a stabilized mixed finite element discretization of this problem and establish optimal a priori error estimates.

In [15], 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.

6.6. Auto adaptative algorithms

In [29], we develop an adaptive version of the inexact Uzawa algorithm applied to finite element discretizations of the linear Stokes problem. We base our developments on an equilibrated flux a posteriori error estimate distinguishing the different error components, namely the discretization error component, the inner algebraic solver error component, and the outer Uzawa iteration error component. On each outer Uzawa and inner linear algebraic solver iteration, we prove that our estimate gives a guaranteed upper bound on the total error, as well as a polynomial-degree-robust local efficiency. Our adaptive inexact algorithm stops the outer Uzawa iteration

and the inner linear algebraic solver iteration when the Uzawa error component, respectively the algebraic solver error component, do not have a significant influence on the total error. The developed framework covers all standard conforming and conforming stabilized finite element methods. The implementation into the FreeFem++ programming language is invoked and two numerical examples showcase the performance of our adaptive strategy.

6.7. Spectrum for a small inclusion of negative material

We studied a spectral problem $(\mathcal{P}^\delta)$ for a diffusion like equation in a 3D domain Ω. The main originality here lies in the presence of a parameter σ^δ, whose sign changes on Ω, in the principal part of the operator we consider. More precisely, σ^δ is positive on Ω except in a small inclusion of size $\delta > 0$. Because of the sign-change of σ^δ, for all $\delta > 0$ the spectrum of $(\mathcal{P}^\delta)$ consists of two sequences converging to $+\infty$ and $-\infty$. However, at the limit $\delta = 0$, the small inclusion vanishes so that there should only remain positive spectrum for $(\mathcal{P}^\delta)$. What happens to the negative spectrum? In this paper, we prove that the positive spectrum of $(\mathcal{P}^\delta)$ tends to the spectrum of the problem without the small inclusion. On the other hand, we establish that each negative eigenvalue of $(\mathcal{P}^\delta)$ behaves like $\delta^{-2}\mu$ for some constant $\mu < 0$. We also show that the eigenvectors associated with the negative eigenvalues are localized around the small inclusion. We end the article providing 2D numerical experiments illustrating these results.

6.8. Stability of electromagnetic cavities perturbed by small perfectly conducting inclusions

We consider an electromagnetic wave propagation problem in harmonic regime in a bounded cavity, in the case where the medium of propagation contains small perfectly conducting inclusions. We prove that the solution to this problem depends continuously on the data in a uniform manner with respect to the size of the inclusions.

6.9. Integral equations for acoustic scattering by partially impenetrable composite objects

We study direct first-kind boundary integral equations arising from transmission problems for the Helmholtz equation with piecewise constant coefficients and Dirichlet boundary conditions imposed on a closed surface. We identify necessary and sufficient conditions for the occurrence of so-called spurious resonances, that is, the failure of the boundary integral equations to possess unique solutions.

Following [A. Buffa and R. Hiptmair, Numer Math, 100, 1–19 (2005)] we propose a modified version of the boundary integral equations that is immune to spurious resonances. Via a gap construction it will serve as the basis for a universally well-posed stabilized global multi-trace formulation that generalizes the method of [X. Claeys and R. Hiptmair, Commun Pure and Appl Math, 66, 1163–1201 (2013)] to situations with Dirichlet boundary conditions.

6.10. Application domain: data analysis in astrophysics

One of the application domain on which our algorithms are validated is data analysis in astrophysics. Estimation of the sky signal from sequences of time order data is one of the key steps in the Cosmic Microwave Background (CMB) data analysis, commonly referred to as the map-making problem. Some of the most popular and general methods proposed for this problem involve solving generalised least squares (GLS) equations with non-diagonal noise weights given by a block-diagonal matrix with Toeplitz blocks. In [14] we study new map-making solvers potentially suitable for applications to the largest, anticipated data sets. They are based on iterative conjugate gradient (CG) approaches enhanced with novel, parallel, two-level preconditioners (2lvl-PCG). We apply the proposed solvers to examples of simulated, non-polarised and polarised CMB observations and a set of idealised scanning strategies with a sky coverage ranging from nearly a full sky down to small sky patches. We discuss in detail their implementation for massively parallel

computational platforms and their performance for a broad range of parameters characterising the simulated data sets. We find that our best new solver can outperform carefully optimised, standard solvers as used today, by as much as a factor of 5 in terms of the convergence rate and a factor of 4 in terms of the time to solution, and does so without increasing significantly the memory consumption or the volume of inter-processor communication. The performance of the new algorithms is also found to be more stable, robust and less dependent on specific characteristics of the analysed data set.We therefore conclude that the proposed approaches are well suited to address successfully challenges posed by new and forthcoming CMB data sets.

Spherical Harmonic Transforms (SHT) are at the heart of many scientific and practical applications ranging from climate modelling to cosmological observations. In many of these areas new, cutting-edge science goals have been recently proposed requiring simulations and analyses of experimental or observational data at very high resolutions and of unprecedented volumes. Both these aspects pose formidable challenge for the currently existing implementations of the transforms.

In [13] we describe parallel algorithms for computing SHT with two variants of intra-node parallelism appropriate for novel supercomputer architectures, multi-core processors and Graphic Processing Units (GPU). It also discusses their performance, alone and embedded within a top-level, MPI-based parallelisation layer ported from the S^2HAT library, in terms of their accuracy, overall efficiency and scalability. We show that our inverse SHT run on GeForce 400 Series GPUs equipped with latest CUDA architecture ("Fermi") outperforms the state of the art implementation for a multi-core processor executed on a current Intel Core i7-2600K. Furthermore, we show that an MPI/CUDA version of the inverse transform run on a cluster of 128 Nvidia Tesla S1070 is as much as 3 times faster than the hybrid MPI/OpenMP version executed on the same number of quad-core processors Intel Nehalem for problem sizes motivated by our target applications. Performance of the direct transforms is however found to be at the best comparable in these cases. We discuss in detail the algorithmic solutions devised for the major steps involved in the transforms calculation, emphasising those with a major impact on their overall performance, and elucidates the sources of the dichotomy between the direct and the inverse operations.

AVALON Project-Team

6. New Results

6.1. Energy efficiency of large scale distributed systems

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

6.1.1. Improving Energy Efficiency of Large Scale Systems without a priori Knowledge of Applications and Services

Unlike their hardware counterpart, software solutions to the energy reduction problem in large scale and distributed infrastructures hardly result in real deployments. At the one hand, this can be justified by the fact that they are application oriented. At the other hand, their failure can be attributed to their complex nature which often requires vast technical knowledge behind proposed solutions and/or thorough understanding of applications at hand. This restricts their use to a limited number of experts, because users usually lack adequate skills. In addition, although subsystems including the memory and the storage are becoming more and more power hungry, current software energy reduction techniques fail to take them into account. We propose a methodology for reducing the energy consumption of large scale and distributed infrastructures. Broken into three steps known as (i) phase identification, (ii) phase characterization, and (iii) phase identification and system reconfiguration; our methodology abstracts away from any individual applications as it focuses on the infrastructure, which it analyses the runtime behaviour and takes reconfiguration decisions accordingly.

The proposed methodology is implemented and evaluated in high performance computing (HPC) clusters of varied sizes through a Multi-Resource Energy Efficient Framework (MREEF). MREEF implements the proposed energy reduction methodology so as to leave users with the choice of implementing their own system reconfiguration decisions depending on their needs. Experimental results show that our methodology reduces the energy consumption of the overall infrastructure of up to 24% with less than 7% performance degradation. By taking into account all subsystems, our experiments demonstrate that the energy reduction problem in large scale and distributed infrastructures can benefit from more than “the traditional” processor frequency scaling. Experiments in clusters of varied sizes demonstrate that MREEF and therefore our methodology can easily be extended to a large number of energy aware clusters. The extension of MREEF to virtualized environments like cloud shows that the proposed methodology goes beyond HPC systems and can be used in many other computing environments.

6.1.2. Reservation based Usage for Energy Efficient Clouds: the Climate/Blazar Architecture

The FSN XLcloud project (cf Section 8.1) strives to establish the demonstration of a High Performance Cloud Computing (HPCC) platform based on OpenStask, that is designed to run a representative set of compute intensive workloads, including more specifically interactive games, interactive simulations and 3D graphics. XLcloud is based on OpenStack, and Avalon is contributing to the energy efficiency part of this project. We have proposed and brought our contribution to Climate, a new resource reservation framework for OpenStack, developed in collaboration with Bull, Mirantis and other OpenStack contributors. Climate allows the reservation of both physical and virtual resources, in order to provide a mono-tenancy environment suitable for HPC applications. Climate chooses the most efficient hosts (flop/W). This metric is computed from the CPU / GPU informations, mixed with real power consumption measurements provided by the Kwapi framework. The user requirements may be loose, allowing Climate to choose the best time slot to place the reservation. Climate has been improved with standby mode features, to shut down automatically the unused hosts. The first release of Climate was done in January 2014. Through the OpenStack process, Climate is now named Blazar.

6.1.3. Clustered Virtual Home Gateway (vHGW)

This result is a joint work between Avalon team (J.P. Gelas, L. Lefevre) and Addis Abeba University (M. Tsibie and T. Assefa). The customer premises equipment (CPE), which provides the interworking functions between the access network and the home network, consumes more than 80% of the total power in a wireline access network. In the GreenTouch initiative (cf Section 8.3), we aim at a drastic reduction of the power consumption by means of a passive or quasi-passive CPE. Such approach requires that typical home gateway functions, such as routing, security, and home network management, are moved to a virtual home gateway (vHGW) server in the network. In our first prototype virtual home gateways of the subscribers were put in LXC containers on a unique GNU/Linux server. The container approach is more scalable than separating subscribers by virtual machines. We demonstrated a sharing factor of 500 to 1000 virtual home gateways on one server, which consumes about 150 W, or 150 to 300 mW per subscriber. Comparing this power consumption with the power of about 2 W for the processor in a thick client home gateway, we achieved an efficiency gain of 5-10x. The prototype was integrated and demonstrated at TIA 2012 in Dallas. In our current work, we propose the Clustered vHGWs Data center architecture to yield optimal energy conservation through virtual machine's migration among physical nodes based on the current subscriber's service access state, while ensuring SLA respective subscribers. Thus, optimized energy utilization of the data center is assured without compromising the availability of service connectivity and QoS preferences of respective subscribers. The last prototype including those new features was integrated and demonstrated recently to the GreenTouch consortium members at Melbourne University.

6.1.4. Energy proportionality with heterogeneous computing resources

This work [16] focuses on improving energy proportionality of large scale virtualized environments. The main problem of such infrastructures is their high static costs due to high idle power consumption of idle servers. Our goal is to reach an infrastructure able to adapt its energy consumption to the current working load. Therefore we propose an original infrastructure composed of heterogeneous computing resources. We consider the heterogeneity at the level of the architecture, and we gather in our platform low power ARM processors together with powerful x86 servers. Around this infrastructure, we are developing a decisional framework to schedule applications on the architecture, or combination of architectures, most suitable to their current needs. The framework reacts dynamically to the resource needs evolutions by migrating the applications to the chosen destinations, and switching off unused nodes to save energy. We validate our scheduling policies by building a simulator based on a set of experimental inputs about power and performance hardware profiles and applications load profiles. This work is jointly done with IRIT Lab. (Toulouse) under the support of Inria Large Scale Initiative Hemera.

6.1.5. Energy efficient Core Networks

This work [11] 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 change the status of router ports and transponders on the two extremities of a link. The status of these components is set to sleep mode whenever a link is not required to transfer data, and brought back to operational state when needed. We have analyzed the implementation issues of an energy-efficient SDN-based traffic engineering in core networks. We propose 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 our proposed algorithms in the OMNET++ packet-based discrete event simulator. Experiments considering real network topologies (Germany50 and Ge´ant) and real dynamic traffic matrices allowed us to quantify the trade-off between energy saving and impact of our solution on network performance. As mean to reroute the traffic we use a promising new protocol, SPRING. This comes in contrast with other works, which use classical IP link weights changes or MPLS+RSVP-TE for this purpose. SPRING proved itself well suited for dynamic reconfiguration of the network. Experimental results show that the consumption of 44% of links can be reduced while preserving good quality of service.

6.1.6. Energy aware scheduling for multi data centers clouds

Our work tackles the challenge of improving the energy efficiency of server provisioning and workload management [17]. It introduces a metric allowing infrastructure administrators to specify their preferences between performance and energy savings. We describe a framework for resource management which provides control for informed and automated provisioning at the scheduler level while providing developers (administrator or end-user) with an abstract layer to implement aggregation and resource ranking based on contextual information such as infrastructure status, users' preferences and energy- related external events occurring over time. We integrate our solution in DIET which allows for managing heterogeneous nodes at the middleware layer. The evaluation is performed by means of simulations and real-life experiments on the GRID'5000 testbed. Results show improvements in energy efficiency with minimal impact on application and system performance. Implementation has been used within the industrial project Nu@ge in the context of a federation of modular datacenters.

6.2. Simulation of Large Scale Distributed Systems

Participants: Frédéric Desprez, Jonathan Rouzaud-Cornabas, Frédéric Suter.

6.2.1. Versatile, Scalable, and Accurate Simulation of Distributed Applications and Platforms

The study of parallel and distributed applications and platforms, whether in the cluster, grid, peer-to-peer, volunteer, or cloud computing domain, often mandates empirical evaluation of proposed algorithmic and system solutions via *simulation*. Unlike direct experimentation via an application deployment on a real-world testbed, simulation enables fully repeatable and configurable experiments for arbitrary hypothetical scenarios. Two key concerns are accuracy (so that simulation results are scientifically sound) and scalability (so that simulation experiments can be fast and memory-efficient). While the scalability of a simulator is easily measured, the accuracy of many state-of-the-art simulators is largely unknown because they have not been sufficiently validated. In this work we describe recent accuracy and scalability advances made in the context of the SIMGRID simulation framework. A design goal of SIMGRID is that it should be versatile, i.e., applicable across all aforementioned domains. We present quantitative results that show that SIMGRID compares favorably to state-of-the-art domain-specific simulators in terms of scalability, accuracy, or the trade-off between the two. An important implication is that, contrary to popular wisdom, striving for versatility in a simulator is not an impediment but instead is conducive to improving both accuracy and scalability.

6.2.2. Simulation of MPI Applications with Time-Independent Traces

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. In this work we present 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.

6.2.3. Adding Storage Simulation Capacities to the SimGrid Toolkit

For each kind of distributed computing infrastructures, i.e., clusters, grids, clouds, data centers or supercomputers, storage is 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 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 this work we extend SimGrid, a versatile toolkit for the simulation of large-scale distributed computing systems, with storage simulation capacities. We define the required abstractions and propose a new API to handle storage components and their contents in SimGrid-based simulators. Then we characterize the performance of the fundamental storage component that are disks and derive models of these resources. Finally we list several concrete use cases of storage simulations in clusters, grids, clouds, and data centers for which the proposed extension would be beneficial.

6.3. MapReduce Computations on Hybrid Distributed Computations Infrastructures

Participants: Gilles Fedak, Julio Anjos, Asma Ben Cheikh Ahmed.

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.

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

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. This study introduces 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 describe 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.

6.3.2. Parallel Data Processing in Dynamic Hybrid Computing Environment Using MapReduce

In this work, we propose a novel MapReduce computation model in hybrid computing environment called HybridMR is proposed. 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 show 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.

6.3.3. Ensuring Privacy for MapReduce on Hybrid Clouds Using Information Dispersal Algorithm

MapReduce is a powerful model for parallel data processing. The motivation of this work is to allow running map-reduce jobs partially on untrusted infrastructures, such as public Clouds and Desktop Grid, while using a trusted infrastructure, such as private cloud, to ensure that no outsider could get the 'entire' information. Our idea is to break data into meaningless chunks and spread them on a combination of public and private clouds so that the compromise would not allow the attacker to reconstruct the whole data-set. To realize this, we use the Information Dispersion Algorithms (IDA), which allows to split a file into pieces so that, by carefully

dispersing the pieces, there is no method for a single node to reconstruct the data if it cannot collaborate with other nodes. We propose a protocol that allows MapReduce computing nodes to exchange the data and perform IDA-aware MapReduce computation. We conduct experiments on the Grid'5000 testbed and report on performance evaluation of the prototype.

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

Participants: Gilles Fedak, Anthony Simonet.

Large scientific experiments drive scientists to use many storage and computing platforms as well as different applications, 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 difficulty managing and manipulating such large and highly distributed datasets also limits automated sharing and collaboration. In this collaboration with Kyle Chard and Ian Foster from Argonne National Lab and University of Chicago, 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 demon- strate that using a specialized data life cycle and programming model—Active Data—we can easily implement global progress monitoring, sharing and recovery from unexpected events in heterogeneous environments and automate human tasks.

6.5. HPC Component Model

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

6.5.1. 3D FFT and L^{2}C

We have studied the relevance of dealing with 3D FFT parallel algorithms with the software component model L^{2}C [31]. We have implemented several existing 3D FFT algorithms, and we have evaluated their performance, their scalability, and their reuse rate. Experiments made on clusters of Grid'5000 and on the Curie supercomputer up to 8192 cores show that L2C based 3D assemblies are scalable and have the same kind of performance than existing 3D libraries such as FFTW or 2DECOMP. This work confirms than components can be used for optimized HPC applications

6.5.2. Stencil Skeletons in L^{2}C

Mesh-based scientific simulation is an important class of scientific application which could benefit from component models. Therefore, we have studied and designed a first adaptation of the SIPSim model [33] (Structured Implicit Parallelism for scientific Simulations) to handle HPC component models. The heat equation application has been implemented on top of L^{2}C following this adapted SIPSim model. First experiments on clusters of Grid'5000 and on the Curie supercomputer show promising results, of which a complete analysis is still ongoing. This work is a first step toward a complete implicit parallelism stencil skeleton using L^{2}C.

6.5.3. Reconfigurable HPC component model

High-performance applications whose structure changes dynamically during execution are extremely complex and costly 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.

We aim at addressing performance, scalability and programmability by separating locking and synchronization concerns from reconfiguration code. To this end, we have defined *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* based on L^{2}C.

6.6. Security for Virtualization and Clouds

Participants: Eddy Caron, Arnaud Lefray, Jonathan Rouzaud-Cornabas.

Our framework Security Aware Models for Clouds has two purposes. The first one is, for a client, to model an IaaS application composed of virtual machines, applications, datas and communications and specify the associated security requirements. The whole modelization is contained into a XML file. The second one is the scheduling. It takes as inputs application models (XML) and the infrastructure of the cloud (currently in XML) i.e. a hierarchical set of physical machines. The scheduler encapsulates applications into virtual machines when needed and then maps virtual machines onto physical machines. The result of this scheduling is a file with the mapping i.e. a list of (VM, PM) couples.

The scheduler, as a standalone engine, can be used as simulator. But it can be interfaced with a Cloud stack (e.g. OpenStack, OpenNebula) to act as a production scheduler. This inferfacing is achieved by dynamically inferring the infrastructure model from the Cloud database and applying the decision i.e the output mapping list. Furthermore, the security policies (as input) are splitted for local security enforcement on each physical machine.

Sam4C (Security-Aware Models For Clouds) is a twofold framework, namely Sam4C-Modeler and Sam4C-Scheduler. The first is dedicated to modeling an application with the tenant's virtual machines and network interconnection. The second is is a security-aware scheduler, meaning it overrides the basic default scheduler with mainly the following enhanced capabilities

We have designed a scheduling module called SPS. This module is designed to support all the operations concerning the Cloud. It is based on the OpenStack and extends OpenStack with security aspects to fulfil the requirements of Seed4C.

6.7. Locality-aware Cooperation for VM Scheduling in Distributed Clouds

Participant: Frédéric Desprez.

In collaboration with the Ascola team (A. Lèbre, J. Pastor), ASAP team (Marin Bertier), and the Myriads team (C. Tedeschi), we worked on the design of a distributed Cloud Computing infrastructure [23]. The promotion of such infrastructures as the next platform to deliver the Utility Computing paradigm, leads to new virtual machines (VMs) scheduling algorithms leveraging peer-to-peer approaches. Although these proposals considerably improve the scalability, leading to the management of hundreds of thousands of VMs over thousands of physical machines (PMs), they do not consider the network overhead introduced by multi-site infrastructures. This overhead can have a dramatic impact on the performance if there is no mechanism favoring intra-site versus inter-site manipulations.

In 2014, we designed a new building block designed on top of a network with Vivaldi coordinates maximizing the locality criterion (*i.e.,* efficient collaborations between PMs) [12]. We combined such a mechanism with DVMS, a large-scale virtual machine scheduler and showed its benefit by discussing several experiments performed on four distinct sites of the Grid'5000 testbed. With our proposal and without changing the scheduling decision algorithm, the number of inter-site operations has been reduced by 72%. This result provides a glimpse of the promising future of using locality properties to improve the performance of massive distributed Cloud platforms.

HIEPACS Project-Team

6. New Results

6.1. Highlights of the Year

In the context of HPC-PME initiative, we started a collaboration with ALGO'TECH INFORMATIQUE and we have organised one of the first PhD-consultant action implemented by Xavier Lacoste led by Pierre Ramet. ALGO'TECH is one of the most innovative SMEs (small and medium sized enterprises) in the field of cabling embedded systems, and more broadly, automatic devices. The main target of the project is to validate the possibility to use the sparse linear solvers of our team in the area of electromagnetic simulation tools developped by ALGO'TECH. This collaboration will be developed next year in the context of the European project FORSTISSIMO. The principal objective of FORTISSIMO is to enable European manufacturing, particularly SMEs, to benefit from the efficiency and competitive advantage inherent in the use of simulation.

As a conclusion of the OPTIDIS project we organized the first International Workshop on Dislocation Dynamics Simulations that was devoted to the latest developments realized worldwide in the field of Discrete Dislocation Dynamics simulations. This international event held in December 10th to the 12th at "Maison de la Simulation" in Saclay, France and attracted 55 participants from many different countries including England, Germany, France, USA, ... The workshop gathered most of the active researchers working on dislocation dynamics from numerical simulations to experimentatios. Thanks to the success of this workshop, a second one will be scheduled in England during 2016.

6.2. High-performance computing on next generation architectures

6.2.1. Composing multiple StarPU applications over heterogeneous machines: a supervised approach

Enabling HPC applications to perform efficiently when invoking multiple parallel libraries simultaneously is a great challenge. Even if a uniform runtime system is used underneath, scheduling tasks or threads coming from different libraries over the same set of hardware resources introduces many issues, such as resource oversubscription, undesirable cache flushes or memory bus contention.

This work presents an extension of `StarPU`, a runtime system specifically designed for heterogeneous architectures, that allows multiple parallel codes to run concurrently with minimal interference. Such parallel codes run within *scheduling contexts* that provide confined execution environments which can be used to partition computing resources. Scheduling contexts can be dynamically resized to optimize the allocation of computing resources among concurrently running libraries. We introduce a *hypervisor* that automatically expands or shrinks contexts using feedback from the runtime system (e.g. resource utilization). We demonstrate the relevance of our approach using benchmarks invoking multiple high performance linear algebra kernels simultaneously on top of heterogeneous multicore machines. We show that our mechanism can dramatically improve the overall application run time (-34%), most notably by reducing the average cache miss ratio (-50%).
This work is developed in the framework of Andra Hugo's PhD. These contributions have been published in the international journal of High Performance Computing Applications [21].

6.2.2. A task-based $\mathcal{H}$-Matrix solver for acoustic and electromagnetic problems on multicore architectures

$\mathcal{H}$-Matrix is a hierarchical, data-sparse approximate representation of matrices that allows the fast approximate computation of matrix products, LU and LDL^T decompositions, inversion and more. This representation is suitable for the direct solution of large dense linear systems arising from the Boundary Element Method in $O(N \log_2^{\alpha}(N))$ operations. This kind of formulation is widely used in the industry for the numerical simulation of acoustics and electromagnetism scattering by large objetcs. Applications of this approach include

aircraft noise reduction and antenna sitting at Airbus Group. The recursive and irregular nature of these $\mathcal{H}$-Matrix algorithms makes an efficient parallel implementation very challenging, especially when relying on a "Bulk Synchronous Parallel" paradigm. We have considered an alternative parallelization for multicore architectures using a task-based approach on top of a runtime system, namely StarPU. We have showed that our method leads to a highly efficient, fully pipelined computation on large real-world industrial test cases provided by Airbus Group.
This research activity has been conduced in the framework of the EADS-ASTRIUM, Inria, Conseil Régional initiative in collaboration with the RUNTIME Inria project, and is part of Benoit Lize's PhD.

6.2.3. A task-based 3D geophysics application

Reverse Time Migration (RTM) technique produces underground images using wave propagation. A discretization based on the Discontinuous Galerkin (DG) method unleashes a massively parallel elastodynamics simulation, an interesting feature for current and future architectures. We have designed a task-based version of this scheme in order to enable the use of manycore architectures. At this stage, we have demonstrated the efficiency of the approach on homogeneous and cache coherent Non Uniform Memory Access (ccNUMA) multicore platforms (up to 160 cores) and designed a prototype version of a distributed memory version that can exploit multiple instances of such architectures. This work has been conducted in the context of the DIP Inria-Total strategic action in collaboration with the MAGIQUE3D Inria project and thanks to the long-term visit of George Bosilca funded by TOTAL. Geroge's expertise ensured an optimum usage of the PaRSEC runtime system onto which our task-based scheme has been ported.
This work was presented during HPCC conference [27] as well as during a TOTAL scientific event [26].

6.2.4. Resiliency in numerical simulations

For the solution of systems of linear equations, various recovery-restart strategies have been investigated in the framework of Krylov subspace methods to address the situations of core failures. The basic underlying idea is to recover fault entries of the iterate via interpolation from existing values available on neighbor cores. In that resilience framework, we have extended the recovey-restart ideas to the solution of linear eigenvalue problems. Contrary to the linear system case, not only the current iterate can be interpolated but also part of the subspace where candidate eigenpairs are searched.
This work is developed in the framework of Mawussi Zounon's PhD funded by the ANR RESCUE. These contributions have been presented in particuler at the international SIAM workshop on Exascale Applied Mathematics Challenges and Opportunities [40] in Chicago and the Householder symposium [41] in Spa. Notice that theses activities are also part of our contribution to the G8 ESC (Enabling Climate Simulation at extreme scale).

6.2.5. 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 [36] in Hyderabad.

6.3. High performance solvers for large linear algebra problems

6.3.1. Parallel sparse direct solver on runtime systems

The ongoing hardware evolution exhibits an escalation in the number, as well as in the heterogeneity, of the computing resources. The pressure to maintain reasonable levels of performance and portability, forces the application developers to leave the traditional programming paradigms and explore alternative solutions. PaStiX is a parallel sparse direct solver, based on a dynamic scheduler for modern hierarchical architectures. In this paper, we study the replacement of the highly specialized internal scheduler in PaStiX by two generic runtime frameworks: PaRSEC and StarPU. The tasks graph of the factorization step is made available to the two runtimes, providing them with the opportunity to optimize it in order to maximize the algorithm efficiency for a predefined execution environment. A comparative study of the performance of the PaStiX solver with the three schedulers - native PaStiX, StarPU and PaRSEC schedulers - on different execution contexts is performed. The analysis highlights the similarities from a performance point of view between the different execution supports. These results demonstrate that these generic DAG-based runtimes provide a uniform and portable programming interface across heterogeneous environments, and are, therefore, a sustainable solution for hybrid environments.

This work has been developed in the framework of Xavier Lacoste's PhD funded by the ANR ANEMOS. These contributions have been presented at the Heterogeneous Computing Workshop held jointly with the international conference IPDPS 2014 [32]. Xavier Lacoste will defend his PhD in February 2015.

6.3.2. Hybrid parallel implementation of hybrid solvers

In the framework of the hybrid direct/iterative MaPHyS solver, we have designed and implemented an hybrid MPI-thread variant. More precisely, the implementation relies on the multi-threaded MKL library for all the dense linear algebra calculations and the multi-threaded version of PaStiX. Among the technical difficulties, one was to make sure that the two multi-threaded libraries do not interfere with each other. The resulting software prototype is currently experimented to study its new capability to get flexibility and trade-off between the parallel and numerical efficiency. Parallel experiments have been conducted on the Plafrim plateform as well as on a large scale machine located at the USA DOE NERSC, which has a large number of CPU cores per socket.
This work is developed in the framework of the PhD thesis of Stojce Nakov funded by TOTAL.

6.3.3. Designing LU-QR hybrid solvers for performance and stability

New hybrid LU-QR algorithms for solving dense linear systems of the form $Ax = b$ have been introduced. Throughout a matrix factorization, these algorithms dynamically alternate LU with local pivoting and QR elimination steps, based upon some robustness criterion. LU elimination steps can be very efficiently parallelized, and are twice as cheap in terms of flops, as QR steps. However, LU steps are not necessarily stable, while QR steps are always stable. The hybrid algorithms execute a QR step when a robustness criterion detects some risk for instability, and they execute an LU step otherwise. Ideally, the choice between LU and QR steps must have a small computational overhead and must provide a satisfactory level of stability with as few QR steps as possible. In this work, we introduce several robustness criteria and we establish upper bounds on the growth factor of the norm of the updated matrix incurred by each of these criteria. In addition, we describe the implementation of the hybrid algorithms through an extension of the PaRSEC software to allow for dynamic choices during execution. Finally, we analyze both stability and performance results compared to state-of-the-art linear solvers on parallel distributed multicore platforms.
These contributions have been presented at the international conference IPDPS 2014 [30] in Phoenix. An extended version has been submitted to JPDC journal.

6.3.4. 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 will be presented at the international conference IPDPS 2015 [34] in Hyderabad.

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

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. implementation of some new features in the FMM library ScalFMM: adaptive variants of the Chebyshev and Lagrange interpolation based FMM kernels, multiple right-hand sides, generic tensorial nearfield...
2. The parallelization and the FMM core parts rely on ScalFMM (OpenMP/MPI) which has been updated all year round. Finally, ScalFMM offers two new shared memory parallelization strategies using OpenMP 4 and StarPU.

6.4.1. Low rank approximations of matrices

New fast algorithms for the computation of low rank approximations of matrices were implemented in a -soon to be- open-source C++ library. These algorithms are based on randomized techniques combined with standard matrix decompositions (such as QR, Cholesky and SVD). The main contribution of this work is that we make use of ScalFMM parallel library in order to power the large amount of matrix to vector products involved in the algorithms. Applications to the fast generation of Gaussian random fields were adressed. Our methods compare good with the existing ones based on Cholesky or FFT and potentially outpass their performances for specific distributions. We are currently in the process of writing a paper on that topic. Extensions to fast Kalman filtering is now considered. This work is done in collaboration with Eric Darve (Stanford, Mechanical Engineering) in the context of the associate team FastLA.

6.4.2. 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. We present a novel approach based on the re-ordering of the interaction matrices in slices. We end up with a custom multi-vectors/vector product operation and compute it using SIMD intrinsic functions. We take advantage of the new order of the computation to parallelize in shared and distributed memory. We demonstrate the performance of our system by studying the sequential Flop-rate and the parallel scalability, and provide results based on an industrial test-case with up to 32 nodes [43], [28]. From the middle of year 2014, we started working on the TD FMM for the BEM problem. A non optimized version is able to solve the TD BEM with the FMM on parallel distributed nodes. 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.

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

6.5.1. 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. More precisely, we propose to enrich the classic graph model with *interedges*, that represent the coupled code interactions. We describe two new algorithms, called AWARE and PROJREPART, and compare them to the currently used approach (called NAIVE). In recent experimental results, we notice that both AWARE and PROJREPART algorithms succeed to balance the computational load in the coupling phase and in some cases they succeed to reduce the coupling communications costs. Surprisingly we notice that our algorithms do not degrade the global graph edgecut, despite the additional constraints that they impose. In future work, we aim at validating our results on real-life cases in the field of aeronautic propulsion. In order to achieve that, we plan to integrate our algorithms within the Scotch framework. Finally, our algorithms should be implemented in parallel and should be extended in order to manage more complex applications with more than two interacting models.

6.5.2. Graph partitioning for hybrid solvers

Nested Dissection has been introduced by A. George and is a very popular heuristic for sparse matrix ordering before numerical factorization. It allows to maximize the number of parallel tasks, while reducing the fill-in and the operation count. The basic standard idea is to build a "small separator" S of the graph associated with the matrix in order to split the remaining vertices in two parts P_0 and P_1 of "almost equal size". The vertices of the separator S are ordered with the largest indices, and then the same method is applied recursively on the two sub-graphs induced by P_0 and P_1. At the end, if k levels of recursion are done, we get 2^k sets of independent vertices separated from each other by $2^k - 1$ separators. However, if we examine precisely the complexity analysis for the estimation of asymptotic bounds for fill-in or operation count when using Nested Dissection ordering, we can notice that the size of the halo of the separated sub-graphs (set of external vertices belonging to an old separator and previously ordered) plays a crucial role in the asymptotic behavior achieved. In the perfect case, we need halo vertices to be balanced among parts. Considering now hybrid methods mixing both direct and iterative solvers such as HIPS, MaPHyS, obtaining a domain decomposition leading to a good balancing of both the size of domain interiors and the Scalable numerical schemes for scientific applications size of interfaces is a key point for load balancing and efficiency in a parallel context. This leads to the same issue: balancing the halo vertices to get balanced interfaces. For this purpose, we revisit the algorithm introduced by Lipton, Rose and Tarjan which performed the recursion of nested dissection in a different manner: at each level, we apply recursively the method to the sub-graphs But, for each sub-graph, we keep track of halo vertices. We have implemented that in the Scotch framework, and have studied its main algorithm to build a separator, called greedy graph growing.
This work is developed in the framework of Astrid Casadei's PhD. These contributions have been presented at the international conference HIPC 2014 [29] in Goa.

6.6. Application Domains

6.6.1. Dislocation dynamics simulations in material physics

6.6.1.1. Long range interaction

Various optimizations have been performed in the Dislocation Dynamics code OptiDis for the long-ranged isotropic elastic force and energy models using a Fast Fourier based Fast Multipole Method (also known as

Uniform FMM). Furthermore the anisotropic elastic force model was implemented using spherical harmonics expansions of angular functions known as Stroh matrices. Optimizations with respect to the crystallographic symmetries were also considered. Once the corresponding semi-analytic formulae for the force field are derived this method should compare well with existing approaches based on expanding the anisotropic elastic Green's function.

6.6.1.2. Parallel dislocation dynamics simulation

This year we have focused on the improvements of our hybrid MPI-OpenMP parallelism of the OptiDis code. More precisely, we have continued the development of the cache-conscious data structure to manage 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.

These contributions have been presented in minisymposia at the 11th World Congress on Computational Mechanics [47], 7th MMM International Conference on Multiscale Materials Modeling [25], [61] and at the International Workshop on DD simulations [62].
This work is developed in the framework of the ANR OPTIDIS.

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

6.6.2.1. 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.

6.6.2.2. 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 is carried on in the framework of Fabien Rozar's PhD in collaboration with CEA Cadarache.

6.6.2.3. 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.
Preliminary results have been presented at the international HPCC workshop on HPC-CFD in Energy/Transport Domains [50] in Paris. The main contribution will be presented at the international conference IPDPS 2015 [33] in Hyderabad.

6.6.2.4. 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 will be presented as a poster during the SIAM CSE'15 conference and we are in the process to submit a paper in the Parallel CFD'15 conference.

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 Les Mureaux.

KerData Project-Team

6. New Results

6.1. Highlights of the Year

IEEE Cluster 2014. The KerData Team had a leading role the organization of the IEEE Cluster 2014 conference, held in Madrid (22–26 September 2014): Gabriel Antoniu as PC Chair, Luc Bougé as Student Mentoring Program Chair, Alexandru Costan as Submissions Chair.

6.2. Data Management for Geographically Distributed Workflows

6.2.1. OverFlow: a multi-site-aware framework for Big Data management

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 coincides with 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, as shown in earlier evaluations [21]. Existing solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemes. In turn, workflow engines need to find ad-hoc substitutes, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability.

We tackle these problems by trying to understand to what extent the intra- and inter-datacenter transfers can impact the total makespan of cloud workflows. We advocate storing data on the compute nodes and transferring files between them directly, in order to exploit data locality and to avoid the overhead of interacting with a shared file system. Under these circumstances, we propose a file management service that enables high throughput through self-adaptive selection among multiple transfer strategies (e.g. FTP-based, BitTorrent-based, etc.). Next, we focus on the more general case of large-scale data dissemination across geographically distributed sites. The key idea is to predict I/O and transfer performance accurately and robustly in a dynamic cloud environment in order to decide judiciously how to perform transfer optimizations over federated datacenters: predict the best combination of protocol and transfer parameters (e.g., multi-routes, flow count, multicast enhancement, replication degree) to maximize throughput or minimize costs, according to users policies. We have implemented these principles in OverFlow, as part of the Azure Cloud so that applications could use it using a Software-as-a-Service (SaaS) approach.

OverFlow [20] was validated on the Microsoft cloud across the 6 EU and US sites. 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.

6.2.2. Metadata management for geographically distributed workflows

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

Scientific workflow data can reach sizes that exceed single-site capabilities. It is needed to support fine-grain data stripping to handle either very large files or very large sets of small files across data centers. Therefore, metadata becomes a critical issue. Moreover, workflow metadata provides crucial information to optimize data management, particularly in the context of geographically distributed data centers. Many present-day distributed file systems, such as GoogleFS and HDFS, include a potential bottleneck as the number of files grows, because they use a centralized metadata management scheme. Thus, we argue for a new, *cloud-based, distributed metadata management* scheme.

We have designed four different approaches to a geographically distributed metadata registry, namely: a) baseline centralized version; b) distributed on each data center with centralized replication agent; c) decentralized non-replicated; and d) decentralized replicated with hierarchical access. A comparative analysis showed that the later strategy performs best in terms of metadata operations per time unit. We then evaluate each of our approaches against various workflow benchmarks, with the purpose of dynamically adapt the metadata handling scheme according to the underlying application and cloud contexts. In the next phase, we will provide a uniform metadata handling tool for scientific workflow engines across cloud datacenters, as well as derive a cost model to offer users the best trade-off (performance vs. cost) driven by their constraints.

6.2.3. Transfer-as-a-Service: a cost-effective model for multi-site cloud data management

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Existing cloud data management solutions are limited to cloud-provided storage, which offers low performance based on rigid cost schemas. Users are therefore forced to design and deploy custom solutions, achieving performance at the cost of complex system configurations, maintenance overheads, reduced reliability and reusability. In [19] we have proposed a dedicated cloud data-transfer service that supports largescale data dissemination across geographically distributed sites, advocating for a Transfer-as-a-Service (TaaS) paradigm. The system aggregates the available bandwidth by enabling multi-route transfers across cloud sites, based on the approach previously described.

We argue that the adoption of such a TaaS approach brings several benefits for both users and the cloud providers who propose it. For users of multi-site or federated clouds, our proposal is able to decrease the variability of transfers and increase the throughput up to three times compared to baseline user options, while benefiting from the well-known high availability of cloud-provided services. For cloud providers, such a service can decrease the energy consumption within a datacenter down to half compared to user-based transfers. Finally, we propose a dynamic cost model schema for the service usage, which enables the cloud providers to regulate and encourage data exchanges via a data transfer market.

6.3. Optimizing Map-Reduce processing

6.3.1. Optimizing Map-Reduce in virtualized environments

Participant: Shadi Ibrahim.

As data-intensive applications become popular in the cloud, their performance on the virtualized platform calls for empirical evaluations and technical innovations. Virtualization has become a prominent tool in data centers and is extensively leveraged in cloud environments: it enables multiple virtual machines (VMs) — with multiple operating systems and applications — to run within a physical server. However, virtualization introduces the challenging issue of providing effective QoS to VMs and preserving the high disk utilization (i.e., reducing the seek delay and rotation overhead) when allocating disk resources to VMs.

In [32], we developed a novel disk I/O scheduling framework, named *Pregather*, to improve disk I/O efficiency through exposure and exploitation of the spatial locality in the virtualized environment (regional and sub-regional spatial locality corresponds to the virtual disk space and applications' access patterns, respectively). In [14], we extend *Pregather* to improve disk I/O utilization further while reducing the disk resource contention and ensuring the I/O performance of VMs with different degrees of spatial locality. To do so, we developed an adaptive time-slice allocation scheme based on the spatial locality of VMs, to adjust the lengths of I/O time slices of VMs dynamically. We evaluated *Pregather* through extensive experiments that involve multiple simultaneous applications of both synthetic benchmarks and a Map-Reduce application (e.g., distributed sort) on Xen-based platforms.

Our evaluations use synthetic benchmarks, a Map-Reduce application (distributed sort) and database workloads. They demonstrate that *Pregather* achieves high disk spatial locality, yields a significant improvement in disk throughput, ensures the performance guarantees of VMs, and enables improved Hadoop performance. This work was done in collaboration with Hai Jin, Song Wu and Xiao Ling from Huazhong University of Science and Technology (HUST).

6.3.2. A simulation approach to evaluate Map-Reduce performance under failure

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

Map-Reduce is emerging as a prominent tool for large-scale data analysis. It is often advocated as an easier-to-use, efficient and reliable replacement for the traditional programming model of moving the data to the computation. The popular open source implementation of Map-Reduce, Hadoop, is now widely used by major companies, including Facebook, Amazon, Last.fm, and the New York Times. Fault tolerance is one of the key features of the Map-Reduce system. Map-Reduce is designed to handle various kind of failures including stop-fail and time failures: Map-Reduce re-executes failed tasks and re-launches another copy of slow tasks. Although many studies have been dedicated to investigate and improve the performance of Map-Reduce, comparatively little attention has been devoted on investigating the performance of Map-Reduce under failures.

In this ongoing work, we investigate how Map-Reduce (i.e., Hadoop) behaves under failures. To do so, we developed *iHadoop*, a Hadoop simulator developed in Java on top of SimGrid. Experimental results demonstrated that *iHadoop* accurately simulates the behavior of Hadoop and therefore can accurately predict the performance of Hadoop when running on large-scale system using the Grid'5000 testbed. In particular, iHadoop can accurately predict the percentage of Map tasks locality, the number of speculative tasks and, more importantly, the overall execution time of Map-Reduce applications under failures.

6.3.3. Waste-Free Preemption Strategy for Hadoop

Participants: Orçun Yildiz, Shadi Ibrahim, Gabriel Antoniu.

Hadoop is widely used in the computer industry because of its scalability, reliability, ease of use, and low cost of implementation. Hadoop hides the complexity of discovery and handling failures from the schedulers, but the burden of failure recovery relies entirely on users, regardless of root causes. We systematically assess this burden through a set of experiments, and argue that more effort to reduce this cost to users is desirable. We also analyze the drawback of current Hadoop 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 slots might not be free up as quickly as expected; and 2) the slots might belong to machines with no data on it, introducing extra cost for data transfer through network, which is normally the most scare resource in nowadays data centers.

In this ongoing work, we introduce a new algorithmic approach called the waste-free preemption. The waste-free preemption saves Hadoop scheduler from solely choosing between kill, which instantly releases the slots but is wasteful, and wait, which does not waste any previous effort but fails for the two above-mentioned reasons. With this new strategy, a preemptive version of Hadoop's default schedulers (FIFO and Fair) has been implemented. The evaluation demonstrates the effectiveness of the new feature by comparing its performance with the traditional Hadoop mechanism.

6.3.4. Optimizing incremental Map-Reduce computations for on-demand data upload

Participants: Stefan Ene, Alexandru Costan, Gabriel Antoniu.

Research on cloud-based Big Data analytics has focused so far on optimizing the performance and cost-effectiveness of the computations, while largely neglecting an important aspect: users need to upload massive datasets on clouds for their computations. In this context, we study the problem of running Map-Reduce applications by considering the simultaneous optimization of performance and cost of both the data upload and its corresponding computation taken together. We analyze the feasibility of incremental Map-Reduce approaches to let the computation progress as much as possible during the data upload by using already transferred data to compute intermediate results.

Current approaches that are either optimized for different purposes, or address the computational problem independent of the data upload. In contrast, to our best knowledge, this is the first approach which simultaneously focuses on both data upload and processing. In this context, we show in [17] that it is not always efficient to attempt to overlap the transfer time with as many incremental computations as possible: a better solution is to wait long enough to fill the computational capacity of the Map-Reduce cluster. Based on this idea, we

developed and evaluated a preliminary prototype. To demonstrate the viability of our prototype in real-life, we run extensive experiments in a distributed setting that involves a 11-node large incremental Map-Reduce deployment based on Hourglass. The results show significant benefits for our approach compared with a simple incremental strategy that starts the next incremental job immediately after the previous has finished: the time-to-solution is improved by 1%, the compute time after the data transfer is finished is reduced by up to 40% and the cost is reduced 10 %-44 %. Compared with a serialized strategy that starts the computation only after all data is transferred, the time-to-solution is improved by up to 30 %, the compute time after the upload finished is reduced by up to 60 % and the cost is reduced between 4 % and 23 %.

6.4. Energy-Aware Data Management in the Cloud and Exascale HPC Systems

6.4.1. Energy-efficiency in Hadoop

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

With increasingly inexpensive cloud storage and increasingly powerful cloud processing, the cloud has rapidly become the environment to store and analyze data. Most of the large-scale data computations in the cloud heavily rely on the Map-Reduce paradigm and its Hadoop implementation. Nevertheless, this exponential growth in popularity has significantly impacted power consumption in cloud infrastructures.

In [18], we focus on Map-Reduce and we investigate the impact of dynamically scaling the frequency of compute nodes on the performance and energy consumption of a Hadoop cluster. To this end, a series of experiments are conducted to explore the implications of Dynamic Voltage Frequency scaling (DVFS) settings on power consumption in Hadoop-clusters. By adapting existing DVFS governors (i.e., *performance*, *power-save*, *on-demand*, *conservative* and *user-space*) in the Hadoop cluster, we observe significant variation in performance and power consumption of the cluster with different applications when applying these governors: the different DVFS settings are only sub-optimal for different Map-Reduce applications. Furthermore, our results reveal that the current CPU governors do not exactly reflect their design goal and may even become ineffective to manage power consumption in Hadoop clusters.

More recently, we extended our work to further illustrate the behavior of different governors, which influence the energy consumption in Hadoop Map-Reduce. We extend our experimental platform from 15 to 40 nodes and we employ two additional benchmarks: K-means and wordcount. Moreover, we investigate preliminary DVFS models that adjust to the various stages of Hadoop applications. We also demonstrate that achieving better energy efficiency in Hadoop cannot be done by tuning the governors parameters, nor through a naive coarse-grained tuning of the CPU frequencies or the governors according the running phase (i.e., map phase or reduce phase). In addition, we provide an extensive discussion of the sensitivity for different parameters employed in *ondemand* and *conservative* governors.

6.4.2. Exploring the impact of dedicated resources on energy consumption in Exascale systems

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

The advent of fast, unprecedentedly scalable, yet energy-hungry Exascale supercomputers poses a major challenge consisting in sustaining a high performance-per-Watt ratio. While much recent work has explored new approaches to I/O management, aiming to reduce the I/O performance bottleneck exhibited by HPC applications (and hence to improve application performance), there is comparatively little work investigating the impact of I/O management approaches on energy consumption.

In [23], we explore how much energy a supercomputer consumes while running scientific simulations when adopting various I/O management approaches. We closely examine three radically different I/O schemes including time partitioning, dedicated cores, and dedicated nodes. We implement the three approaches within the Damaris I/O middleware and perform extensive experiments with one of the target HPC applications of the Blue Waters sustained-Petaflops supercomputer project: the CM1 atmospheric model. The experimental results obtained on the French Grid'5000 platform highlight the differences between these three approaches and illustrate in which way various configurations of the application and of the system can impact performance and energy consumption.

Based on those experimental results, we are working on building a new energy model which can estimate the energy consumptions of various I/O management approaches and help users in selecting the optimal I/O approach to run their application.

6.4.3. Energy impact of data consistency management in the HBase distributed cloud data store

Participants: Álvaro García Recuero, Shadi Ibrahim, Gabriel Antoniu.

Cloud Computing has recently emerged as a key technology providing individuals and companies with access to remote computing and storage infrastructures. In order to achieve high-availability and fault-tolerance, cloud data storage relies on replication. That comes with the issue of consistency among distant replicas so one can always get the most up-to-date values from any of them (*e.g.*, fresh data).

In that context, being able to provide data consistency and continuous availability in the Cloud is yet a non-trivial problem, mainly due to the ever-increasing volume, variety and velocity of data in storage systems. Big data processing engines (e.g., Hadoop, Spark, etc.) as well as modern NoSQL storage back-ends (HBase, Cassandra) have to therefore deal with these high volumes of information at large scale while still providing applications with a consistent and on-time data delivery.

In this work, a set of synthetic workloads from YCSB (Yahoo! Cloud Service Benchmark) was configured to simulate random reads/writes and measure their impact into the overall energy consumption of a well-known distributed data store, HBase. The cluster is comprised of 40 servers and the results have been confirmed with several configurations and runs on the Grid5000 experimental platform. The results indicate that certain write-intensive workloads can be a bottleneck in terms of throughput, further deepening the problem of having an energy-efficient consistency management. Regarding read-intensive workloads, we observe similar patterns but with a very different impact on their energy footprint. We plan to further investigate how to leverage energy-aware mechanisms that overcome the energy-consistency trade-off, while taking into account the selected configuration.

6.5. Scalable I/O and Visualization for Exascale Systems

6.5.1. CALCioM: mitigating cross-application I/O interference

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

As larger supercomputers are used by an increasing number of applications in a concurrent manner, the interference produced by multiple applications accessing a shared parallel file system in contention becomes a major problem. Interference often breaks single-application I/O optimizations (such as access patterns preliminarily optimized to improve data locality on disks), thereby dramatically degrading application I/O performance, increasing run-time variability and, as a result, lowering machine-wide efficiency. We addressed this challenge by proposing CALCioM [15], a framework that aims to mitigate I/O interference through the dynamic selection of appropriate scheduling policies. CALCioM allows several applications running on a supercomputer to communicate and coordinate their I/O strategy in order to avoid interfering with one another. We examined four I/O strategies that can be accommodated in this framework: serializing, interrupting, interfering and coordinating. Experiments on Argonne's BG/P Surveyor machine and on several clusters of Grid'5000 showed that CALCioM can be used to improve the scheduling strategy efficiently and transparently between several otherwise interfering applications, given specified metrics of machine-wide efficiency. This work led to a publication at the IPDPS 2014 conference.

6.5.2. Omnisc'IO: Predicting the I/O patterns of HPC applications

Participants: Matthieu Dorier, Shadi Ibrahim, Gabriel Antoniu.

Many I/O optimizations including prefetching, caching, and scheduling, have been proposed to improve the performance of the I/O stack. In order to optimize these techniques, modeling and predicting spatial and temporal I/O patterns of HPC applications as they run, have become crucial. In this direction we introduced Omnisc'IO [16], an original approach that aims to make a step forward toward an intelligent I/O management of HPC applications in next-generation, post-Petascale supercomputers. It builds a grammar-based model of the I/O behavior of any HPC application, and uses this model to predict when future I/O operations will occur, as well as where and how much data will be accessed. Omnisc'IO is transparently integrated into the POSIX and MPI-I/O stacks and does not require any modification to application sources or to high-level I/O libraries. It works without prior knowledge of the application, and converges to accurate predictions within a couple of iterations only. Its implementation is efficient both in computation time and in memory footprint. Omnisc'IO was evaluated with four real HPC applications — CM1, Nek5000, GTC, and LAMMPS — using a variety of I/O backends ranging from simple POSIX to Parallel HDF5 on top of MPI-I/O. Our experiments showed that Omnisc'IO achieves from 79 % to 100 % accuracy in spatial prediction and an average precision of temporal predictions ranging from 0.2 seconds to less than a millisecond. This work was published at the SC14 conference and initiated the development of the Omnisc'IO software.

6.5.3. Smart In-Situ Visualization

Participants: Lokman Rahmani, Matthieu Dorier, Gabriel Antoniu.

The increasing gap between computational power and I/O performance in new supercomputers has started to drive a shift from an offline approach to data analysis to an inline approach, termed *in-situ visualization* (ISV). While most visualization software now provides ISV, they typically visualize large dumps of unstructured data, by rendering everything at the highest possible resolution. This often negatively impacts the performance of simulations that support ISV, in particular when ISV is performed interactively, as in-situ visualization requires synchronization with the simulation. In this ongoing work, we investigate a smarter method of performing ISV. Our approach consists in adapting the resolution of regions of the visualization area based on how much their data are *relevant* with regards to the physical phenomena being simulated. In this direction, we first provide a generic definition of relevant data subsets based on *data variability*. Following this definition, we investigate various filtering algorithms to detect relevant data subsets automatically. The proposed filtering algorithms are derived from information theory, statistics and image processing. Our work is validated in the context of climate simulation, where we show an up to 40% improvement of time-to-solution without any significant loss regarding the quality of visualization (QoV). QoV loss is *quantified* using the structural similarity index metric (SSIM) that takes in consideration human visual system to compute visual errors.

6.6. Data Streaming and Small Data

6.6.1. JetStream: enabling high-performance event streaming across cloud data-centers

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

The easily-accessible computation power offered by cloud infrastructures coupled with the revolution of Big Data are expanding the scale and speed at which data analysis is performed. In their quest for extracting value out of the 3 Vs of Big Data, applications process larger data sets, within and across clouds. Enabling fast data transfers across geographically distributed sites becomes particularly important for applications which manage continuous streams of events in real time. Scientific applications (e.g. the Ocean Observatory Initiative or the ATLAS experiment) as well as commercial ones (e.g. Microsoft's Bing and Office 365 large-scale services) operate on tens of data-centers around the globe and follow similar patterns: they aggregate monitoring data, assess the QoS or run global data mining queries based on inter-site event stream processing.

In [22] we propose a set of strategies for efficient transfers of events between cloud data-centers and we introduce JetStream: a prototype implementing these strategies as a high-performance, batch-based streaming middleware. 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. The prototype was validated on tens of nodes from US and Europe data-centers of the Windows

Azure cloud using synthetic benchmarks and with application code in the context of the Alice experiment at CERN. The results show an increase in transfer rate of 250 times over individual event streaming. Besides, introducing an adaptive transfer strategy brings an additional 25 % gain. Finally, the transfer rate can further be tripled thanks to the use of multi-route streaming.

6.6.2. Efficient management of many small data objects

Participants: Pierre Matri, Alexandru Costan, Gabriel Antoniu.

Large-scale intensive applications must often manage millions or even billions of small objects. Twitter, for example, has to record on average 5700 new tweets every second. Each of these objects are typically smaller than a kilobyte, and as a result, the database has to store billions of these objects. The sheer amount of objects and the small data sizes can also be found in many other applications, like sensor networks, or graph processing. Another important aspect are the access patterns of these applications where reads dominate over writes, which means the storage system has to be heavily optimized towards read performance.

To address these challenges, we are designing a novel storage system offering fast data access with minimal overhead. Learning from BlobSeer [33], we introduce a more efficient way to manage metadata. To this end, we propose to remove the centralised version manager and to distribute versions across the whole cluster using a distributed hash table. This greatly reduces the response times by allowing single-hop reads for most usage patterns. Additionally, this approach distributes the load over the whole cluster, thus providing a better horizontal scalability and fault tolerance.

MESCAL Project-Team

6. New Results

6.1. Simulation of distributed architectures

- Simgrid is a toolkit providing core functionalities for the simulation of distributed applications in heterogeneous distributed environments. It models fine-grain detail of the studied platform. In [6], we present quantitative results that show that SimGrid compares favorably to state-of-the-art domain-specific simulators in terms of scalability, accuracy, or the trade-off between the two. In [37], [29], we develop an hybrid approach of simulation and emulation of applications that use starPU. By using this approach, Simgrid calibrates the time to run specific subtasks at runtime and simulates all system calls of the application. This approach allows us to obtain performance results that are within one percent of measured results.
- In [33], [18], we study the problem of sampling the stationary distribution of a random walker in $\{0 \cdots N\}^d$ using simulation. This algorithm combines the rejection method and coupling from the past of a set of trajectories of the Markov chain that generalizes the classical sandwich approach. We also provide a complexity analysis of this approach in several cases showing a coupling time in $O(N^2 d \log d)$ when no arc is forbidden and an experimental study of its performance.

6.2. Interactive Analysis and Visualization of Large Distributed Systems

- In [13], we review the methodology that we use to visualize information for large-scale data-set. Our approach uses tools from information theory to define a trade-off between the loss of information and the compactness of the representation. This methodology is applied to spatio-temporal representation of traces of execution in [30], [16], [17], [32]. In these papers, we show how to build a concise overview of the trace behavior as the result of a spatio-temporal data aggregation process. The experimental results show that this approach can help the quick and accurate detection of anomalies in traces containing up to two hundred million events.
- Trace analysis graphical user environments have to provide different views on trace data, to really help provide insights on the traced application behavior. In [22], [35], we propose an open and modular software architecture, the FrameSoC workbench, that defines clear principles for view engineering and for view consistency management. The FrameSoC workbench has been successfully applied in real trace analysis use-cases. This work has also been tested on real scenario coming from a collaboration with ST Microelectronic [25].
- In [7], we design a novel prediction method with Bayes model to predict a load fluctuation pattern over a long-term interval, in the context of Google data centers. All of the prediction methods are evaluated using Google trace with 10,000+ heterogeneous hosts. Experiments show that our Bayes method improves the long-term load prediction accuracy by up 5 to 50%, compared to other state-of-the-art methods.

6.3. Management of Parallel Architectures

- In [12], we present a topology-aware load balancing algorithm for parallel multi-core machines and its proof of asymptotic convergence to an optimal solution. The algorithm, named HwTopoLB, takes into account the properties of current parallel systems composed of multi-core compute nodes, namely their network interconnection, and their complex and hierarchical core topology. We have implemented HwTopoLB using the Charm++ Parallel Runtime System and evaluated its performance with two different benchmarks and one application. Our experimental results confirms that HwTopoLB outperform existing load balancing strategies on different multi-core systems.

- Large scale distributed systems typically comprise hundreds to millions of entities that have only a partial view of resources. How to fairly and efficiently share such resources between entities in a distributed way has thus become a critical question. In [31], we develop a possible answer based on Lagrangian optimization and distributed gradient descent. Under certain conditions, the resource sharing problem can be formulated as a global optimization problem, which can be solved by a distributed self-stabilizing demand and response algorithm.
- The management of resources on testbeds, including their description, reservation and verification, is a challenging issue, especially on of large scale testbeds such as those used for research on High Performance Computing or Clouds. In [23], we present the solution designed for the Grid'5000 testbed in order to: (1) provide users with an in-depth and machine-parsable description of the testbed's resources; (2) enable multi-criteria selection and reservation of resources using a HPC resource manager; (3) ensure that the description of the resources remains accurate. In [24], we present Kascade, a solution for the broadcast of data to a large set of compute nodes. We evaluate Kascade using a set of large scale experiments in a variety of experimental settings, and show that Kascade: (1) achieves very high scalability by organizing nodes in a pipeline; (2) can almost saturate a 1 Gbit/s network, even at large scale; (3) handles failures of nodes during the transfer seamlessly because of its fault-tolerant design.

6.4. Reproducible experiments and papers

- In the field of large-scale distributed systems, experimentation is particularly difficult. The studied systems are complex, often nondeterministic and unreliable, software is plagued with bugs, whereas the experiment workflows are unclear and hard to reproduce. In [5], 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.
- Experiment reproducibility is a milestone of the scientific method. Reproducibility of experiments in computer science would bring several advantages such as code re-usability and technology transfer. The reproducibility problem in computer science has been solved partially, addressing particular class of applications or single machine setups. In [26], we present our approach oriented to setup complex environments for experimentation, environments that require a lot of configuration and the installation of several software packages. The main objective of our approach is to enable the exact and independent reconstruction of a given software environment and the reuse of code. We present a simple and small software appliance generator that helps an experimenter to construct a specific software stack that can be deployed on different available testbeds. [14],
- In [28], [45], 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.

6.5. Game Theory and Distributed Optimization

- In wireless networks, channel conditions of and user quality of service (QoS) requirements vary, often quite arbitrarily, with time (e.g. due to user mobility, fading, etc.) In this dynamic setting, static solution concepts (such as Nash equilibrium) are no longer relevant. 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 [21], we examine the performance of the seminal Foschini–Miljanic (FM) power control scheme in a random environment. 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. We introduce

an adjusted version of the FM algorithm which retains the convergence and no-regret properties of the original algorithm in this constrained setting. In [27], we examine the problem of cost / energy-efficient power allocation in uplink multi-carrier orthogonal frequency-division multiple access wireless networks. We use tools from stochastic convex programming to develop a learning scheme that retains its convergence properties irrespective of the magnitude of the observational errors. In [11], we consider a cognitive radio network where wireless users with multiple antennas communicate over several non-interfering frequency bands. We draw on the method of matrix exponential learning and online mirror descent techniques to derive a no-regret policy that relies only on local channel state information.

- In game theory, the best-response strategy of a player is a strategy that maximizes the selfish payoff of this player. A natural and popular question is, when players update their strategy over time, do they converge to a Nash equilibrium. In [15], we characterize the revision sets in different variants of the best response algorithm that guarantee convergence to pure Nash Equilibria in potential games. We prove that if the revision protocol is separable, then the greedy version as well as smoothed versions of the algorithm converge to pure Nash equilibria. If the revision protocol is not separable, then convergence to Nash Equilibria may fail in both cases. In [43], we investigate a class of reinforcement learning dynamics in which each player plays a "regularized best response" to a score vector consisting of his actions' cumulative payoffs. 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.

6.6. Agent-based modeling and applications to Smart Energy and Transportation Systems

- Renewable energy sources, such as wind, are characterized by non-dispatchability, high volatility, and non-perfect forecasts. Energy storage or electric loads that have a flexible consumption are viewed as a way to mitigate these effects. In [9], [19], we study centralized and distributed algorithms for solving this problem. We provide theoretical bounds on the trade-off between energy loss and the use of reserves. We develop a centralized algorithm that attains this bound in [9]. In [19], we study a distributed optimization problem by modeling a two-stage electricity market. We show that the market is efficient: the players' selfish responses to prices coincide with a socially optimal policy. We develop a distributed solution technique based on the Alternating Direction Method of Multipliers (ADMM) and trajectorial forecasts to compute the Nash-equilibrium.
- Bike-sharing systems are becoming important for urban transportation. In these systems, users arrive at a station, pick up a bike, use it for a while, and then return it to another station of their choice. In [8], we propose a stochastic model of an homogeneous bike-sharing system and study the effect of the randomness of user choices on the number of problematic stations. Even in a homogeneous city, the system exhibits a poor performance: the minimal proportion of problematic stations is of the order of the inverse of the capacity. We show that simple incentives, such as suggesting users to return to the least loaded station among two stations, improve the situation by an exponential factor.
- In [10], we discuss the validation of an agent-based model of emergent city systems with heterogeneous agents. We transform our model into an analytically tractable discrete Markov model, and we examine the city size distribution. We show that the Markov chains lead to a power-law distribution when the ranges of migration options are randomly distributed across the agent population. We also identify sufficient conditions under which the Markov chains produce the Zipf's Law, which has never been done within a discrete framework. The conditions under which our simplified model yields the Zipf's Law are in agreement with, and thus validate, the configurations of the original heterogeneous agent-based model.

MOAIS Project-Team

5. New Results

5.1. Scheduling semi-malleable jobs to minimize mean flow time

This paper [9] deals with the problem of scheduling n_A malleable and n_B non-malleable jobs to be executed together on two parallel identical machines to minimize mean flow time. We propose a set of dominant schedules for this problem, and a dynamic programming algorithm that finds an optimal schedule in this dominant set in time $O(n_A^2 n_B)$.

5.2. Elements of Design for Containers and Solutions in the LinBox Library

We describe in this paper [12] new design techniques used in the exact linear algebra library LinBox, intended to make the library safer and easier to use, while keeping it generic and efficient. First, we review the new simplified structure for containers, based on our *founding scope allocation* model. We explain design choices and their impact on coding: unification of our matrix classes, clearer model for matrices and submatrices,... Then we present a variation of the *strategy* design pattern that is comprised of a controller-plugin system: the controller (solution) chooses among plug-ins (algorithms) that always call back the controllers for subtasks. We give examples using the solution `mul`. Finally we present a benchmark architecture that serves two purposes: Providing the user with easier ways to produce graphs; Creating a framework for automatically tuning the library and supporting regression testing.

5.3. Scheduling Data Flow Program in XKaapi: A New Affinity Based Algorithm for Heterogeneous Architectures

Efficient implementations of parallel applications on heterogeneous hybrid architectures require a careful balance between computations and communications with accelerator devices. Even if most of the communication time can be overlapped by computations, it is essential to reduce the total volume of communicated data. The literature therefore abounds with ad hoc methods to reach that balance, but these are architecture and application dependent. We propose [12] here a generic mechanism to automatically optimize the scheduling between CPUs and GPUs, and compare two strategies within this mechanism: the classical Heterogeneous Earliest Finish Time (HEFT) algorithm and our new, parametrized, Distributed Affinity Dual Approximation algorithm (DADA), which consists in grouping the tasks by affinity before running a fast dual approximation. We ran experiments on a heterogeneous parallel machine with twelve CPU cores and eight NVIDIA Fermi GPUs. Three standard dense linear algebra kernels from the PLASMA library have been ported on top of the XKaapi runtime system. We report their performances. It results that HEFT and DADA perform well for various experimental conditions, but that DADA performs better for larger systems and number of GPUs, and, in most cases, generates much lower data transfers than HEFT to achieve the same performance.

5.4. Evaluation of OpenMP Dependent Tasks with the KASTORS Benchmark Suite

The recent introduction of task dependencies in the OpenMP specification provides new ways of synchronizing tasks. Application programmers can now describe the data a task will read as input and write as output, letting the runtime system resolve fine-grain dependencies between tasks to decide which task should execute next. Such an approach should scale better than the excessive global synchronization found in most OpenMP 3.0 applications. As promising as it looks however, any new feature needs proper evaluation to encourage application programmers to embrace it. This paper [26] introduces the KASTORS benchmark suite designed to evaluate OpenMP tasks dependencies. We modified state-of-the-art OpenMP 3.0 benchmarks and data-flow parallel linear algebra kernels to make use of tasks dependencies. Learning from this experience, we propose extensions to the current OpenMP specification to improve the expressiveness of dependencies. We eventually evaluate both the GCC/libGOMP and the CLANG/libIOMP implementations of OpenMP 4.0 on our KASTORS suite, demonstrating the interest of task dependencies compared to taskwait-based approaches.

5.5. Sparse Polynomial Interpolation Codes and their decoding beyond half the minimal distance

We present [21] algorithms performing sparse univariate polynomial interpolation with errors in the evaluations of the polynomial. Based on the initial work by Comer, Kaltofen and Pernet [Proc. ISSAC 2012], we define the sparse polynomial interpolation codes and state that their minimal distance is precisely the length divided by twice the sparsity. At ISSAC 2012, we have given a decoding algorithm for as much as half the minimal distance and a list decoding algorithm up to the minimal distance. Our new polynomial-time list decoding algorithm uses sub-sequences of the received evaluations indexed by a linear progression, allowing the decoding for a larger radius, that is, more errors in the evaluations while returning a list of candidate sparse polynomials. We quantify this improvement for all typically small values of number of terms and number of errors, and provide a worst case asymptotic analysis of this improvement. For instance, for sparsity T = 5 with up to 10 errors we can list decode in polynomial-time from 74 values of the polynomial with unknown terms, whereas our earlier algorithm required 2T (E + 1) = 110 evaluations. We then propose two variations of these codes in characteristic zero, where appropriate choices of values for the variable yield a much larger minimal distance: the length minus twice the sparsity.

5.6. A Spatiotemporal Data Aggregation Technique for Performance Analysis of Large-scale Execution Traces

Analysts commonly use execution traces collected at runtime to understand the behavior of an application running on distributed and parallel systems. These traces are inspected post mortem using various visualization techniques that, however, do not scale properly for a large number of events. This issue, mainly due to human perception limitations, is also the result of bounded screen resolutions preventing the proper drawing of many graphical objects. This paper [21] proposes a new visualization technique overcoming such limitations by providing a concise overview of the trace behavior as the result of a spatiotemporal data aggregation process. The experimental results show that this approach can help the quick and accurate detection of anomalies in traces containing up to two hundred million events.

5.7. Scheduling independent tasks on multi-cores with GPU accelerators

More and more computers use hybrid architectures combining multi-core processors and hardware accelerators like GPUs (Graphics Processing Units). We present in this paper [3] a new method for scheduling efficiently parallel applications with m CPUs and k GPUs, where each task of the application can be processed either on a core (CPU) or on a GPU. The objective is to minimize the maximum completion time (makespan). The corresponding scheduling problem is NP-hard, we propose an efficient approximation algorithm which achieves an approximation ratio of $4/3 + 1/3k$. We first detail and analyze the method, based on a dual approximation scheme, that uses dynamic programming to balance evenly the load between the heterogeneous resources. Then, we present a faster approximation algorithm for a special case of the previous problem, where all the tasks are accelerated when affected to GPU, with a performance guarantee of $3/2$ for any number of GPUs. We run some simulations based on realistic benchmarks and compare the solutions obtained by a relaxed version of the generic method to the one provided by a classical scheduling algorithm (HEFT). Finally, we present an implementation of the $4/3$-approximation and its relaxed version on a classical linear algebra kernel into the scheduler of the XKaapi runtime system.

5.8. A Flexible Framework for Asynchronous In Situ and In Transit Analytics for Scientific Simulations

High performance computing systems are today composed of tens of thousands of processors and deep memory hierarchies. The next generation of machines will further increase the unbalance between I/O capabilities and processing power. To reduce the pressure on I/Os, the in situ analytics paradigm proposes to process the data as closely as possible to where and when the data are produced. Processing can be embedded

in the simulation code, executed asynchronously on helper cores on the same nodes, or performed in transit on staging nodes dedicated to analytics. Today, software environnements as well as usage scenarios still need to be investigated before in situ analytics become a standard practice. In this paper [3] we introduce a framework for designing, deploying and executing in situ scenarios. Based on a component model, the scientist designs analytics workflows by first developing processing components that are next assembled in a dataflow graph through a Python script. At runtime the graph is instantiated according to the execution context, the framework taking care of deploying the application on the target architecture and coordinating the analytics workflows with the simulation execution. Component coordination, zero-copy intra-node communications or inter-nodes data transfers rely on per-node distributed daemons. We evaluate various scenarios performing in situ and in transit analytics on large molecular dynamics systems simulated with Gromacs using up to 1664 cores. We show in particular that analytics processing can be performed on the fraction of resources the simulation does not use well, resulting in a limited impact on the simulation performance (less than 6%). Our more advanced scenario combines in situ and in transit processing to compute a molecular surface based on the Quicksurf algorithm.

5.9. Generic Deterministic Random Number Generation in Dynamic-Multithreaded Platforms

On dynamic multithreaded platforms with on-line scheduling such as work-stealing, randomized computations raise the issue of reproducibility. Compliant with de facto standard sequential Deterministic Random Number Generators (DRNGs) noted R, we propose [23] a parallel DRNG implementation for finite computations that provides deterministic parallel execution. It uses the stateless sub-stream approach, enabling the use of efficient DRNG such as Mersenne Twister or Linear Congruential. We demonstrate that if R provides fast jump ahead in the random sequence, the re-seeding overhead is small, polylog in expectation, independently from the parallel computation's depth. Experiments benchmark the performance of randomized algorithms employing our solution against the stateful DRNG DotMix, tailored to the Cilk Plus dynamic multithreading runtime. The overhead of our implementation ParDRNG compares favorably to the linear overhead of DotMix re-seedings.

ROMA Team

6. New Results

6.1. Highlights of the Year

Yves Robert was awarded the 2014 IEEE Technical Committee on Scalable Computing (TCSC) Award for Excellence.

In October 2014, CERFACS, ENS Lyon, INPT, Inria and University of Bordeaux launched a consortium around the software package MUMPS (see http://mumps-consortium.org).

6.2. Cost-Optimal Execution of Boolean DNF Trees with Shared Streams

Several applications process queries expressed as trees of Boolean operators applied to predicates on sensor data streams, e.g., mobile apps and automotive apps. Sensor data must be retrieved from the sensors, which incurs a cost, e.g., an energy expense that depletes the battery of a mobile device, a bandwidth usage. The objective is to determine the order in which predicates should be evaluated so as to shortcut part of the query evaluation and minimize the expected cost. This problem has been studied assuming that each data stream occurs at a single predicate. In this work [17], [27] we study the case in which a data stream occurs in multiple predicates, either when each predicate references a single stream or when a predicate can reference multiple streams. In the single-stream case we give an optimal algorithm for a single-level tree and show that the problem is NP-complete for DNF trees. For DNF trees we show that there exists an optimal predicate evaluation order that is depth-first, which provides a basis for designing a range of heuristics. In the multi-stream case we show that the problem is NP-complete even for single-level trees. As in the single stream case, for DNF trees we show that there exists a depth-first leaf evaluation order that is optimal and we design efficient heuristics.

6.3. Efficient checkpoint/verification patterns for silent error detection

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 (RR UT-EECS-14-729), 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%.

6.4. Assessing general-purpose algorithms to cope with fail-stop and silent errors

In this work (RR-Inria-8599), we combine the traditional checkpointing and rollback recovery strategies with verification mechanisms to address both fail-stop and silent errors. The objective is to minimize either makespan or energy consumption. While DVFS is a popular approach for reducing the energy consumption, using lower speeds/voltages can increase the number of errors, thereby complicating the problem. We consider

an application workflow whose dependence graph is a chain of tasks, and we study three execution scenarios: (i) a single speed is used during the whole execution; (ii) a second, possibly higher speed is used for any potential re-execution; (iii) different pairs of speeds can be used throughout the execution. For each scenario, we determine the optimal checkpointing and verification locations (and the optimal speeds for the third scenario) to minimize either objective. The different execution scenarios are then assessed and compared through an extensive set of experiments.

6.5. Scheduling the I/O of HPC applications under congestion

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 (RR-Inria-8519), we analyze the effects of interference on application I/O bandwidth, and propose several scheduling techniques to mitigate congestion. 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.

6.6. Power-aware replica placement in tree networks with multiple servers per client

In this work (RR-Inria-8474), we revisit the well-studied problem of replica placement in tree networks. Rather than minimizing the number of servers needed to serve all client requests, we aim at minimizing the total power consumed by these servers. In addition, we use the most general (and powerful) server assignment policy, where the requests of a client can be served by multiple servers located in the (unique) path from this client to the root of the tree. We consider multi-modal servers that can operate at a set of discrete speeds, using the dynamic voltage and frequency scaling (DVFS) technique. The optimization problem is to determine an optimal location of the servers in the tree, as well as the speed at which each server is operated. A major result is the NP-completeness of this problem, to be contrasted with the minimization of the number of servers, which has polynomial complexity. Another important contribution is the formulation of a Mixed Integer Linear Program (MILP) for the problem, together with the design of several polynomial-time heuristics. We assess the efficiency of these heuristics by simulation. For mid-size instances (up to 30 nodes in the tree), we evaluate their absolute performance by comparison with the optimal solution (obtained via the MILP). The most efficient heuristics provide satisfactory results, within 20% of the optimal solution.

6.7. Parallel scheduling of task trees with limited memory

This work [28] investigates the execution of tree-shaped task graphs using multiple processors. Each edge of such a tree represents some large data. A task can only be executed if all input and output data fit into memory, and a data can only be removed from memory after the completion of the task that uses it as an input data. Such trees arise, for instance, in the multifrontal method of sparse matrix factorization. The peak memory needed for the processing of the entire tree depends on the execution order of the tasks. With one processor the objective of the tree traversal is to minimize the required memory. This problem was well studied and optimal polynomial algorithms were proposed.

Here, we extend the problem by considering multiple processors, which is of obvious interest in the application area of matrix factorization. With multiple processors comes the additional objective to minimize the time needed to traverse the tree, i.e., to minimize the makespan. Not surprisingly, this problem proves to be much harder than the sequential one. We study the computational complexity of this problem and provide inapproximability results even for unit weight trees. We design a series of practical heuristics achieving different trade-offs between the minimization of peak memory usage and makespan. Some of these heuristics are able to process a tree while keeping the memory usage under a given memory limit. The different heuristics are evaluated in an extensive experimental evaluation using realistic trees.

6.8. Scheduling Trees of Malleable Tasks for Sparse Linear Algebra

Scientific workloads are often described as directed acyclic task graphs. In this work [30], 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 [62], [63] for matrix computations, 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 parameter which does not depend on the task. We first motivate the relevance of this model for our application with actual experiments on multicore platforms. Then, we study the optimal allocation proposed by Prasanna and Musicus for makespan minimization using optimal control theory. We largely 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 multicore platforms. There, a task cannot be distributed among several distributed nodes. In such a distributed setting (homogeneous or heterogeneous), we prove the NP-completeness of the corresponding scheduling problem, and propose some approximation algorithms. We finally assess the relevance of our approach by simulations on realistic trees. We show 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).

6.9. Non-clairvoyant reduction algorithms for heterogeneous platforms

In this work [6], we have revisited the classical problem of the reduction collective operation in a heterogeneous environment. We have discussed and evaluated four algorithms that are non-clairvoyant, i.e., they do not know in advance the computation and communication costs. On the one hand, Binomial-stat and Fibonacci-stat are static algorithms that decide in advance which operations will be reduced, without adapting to the environment; they were originally defined for homogeneous settings. On the other hand, Tree-dyn and Non-Commut-Tree-dyn are fully dynamic algorithms, for commutative or non-commutative reductions. We have shown that these algorithms are approximation algorithms with constant or asymptotic ratios. We assessed the relative performance of all four non-clairvoyant algorithms with heterogeneous costs through a set of simulations. Our conclusions hold for a variety of distributions.

6.10. Memory-aware tree traversals with pre-assigned tasks

We have studied the complexity of traversing tree-shaped workflows whose tasks require large I/O files. We target a heterogeneous architecture with two resource types, each with a different memory, such as a multicore node equipped with a dedicated accelerator (FPGA or GPU). The tasks in the workflow are colored according to their type and can be processed if all there input and output files can be stored in the corresponding memory. The amount of used memory of each type at a given execution step strongly depends upon the ordering in which the tasks are executed, and upon when communications between both memories are scheduled. The objective is to determine an efficient traversal that minimizes the maximum amount of memory of each type needed to traverse the whole tree. In this study [11], we establish the complexity of this two-memory scheduling problem, and provide inapproximability results. In addition, we design several heuristics, based on both post-order and general traversals, and we evaluate them on a comprehensive set of tree graphs, including random trees as well as assembly trees arising in the context of sparse matrix factorizations.

6.11. Analysis of Dynamic Scheduling Strategies for Matrix Multiplication on Heterogeneous Platforms

The tremendous increase in the size and heterogeneity of supercomputers makes it very difficult to predict the performance of a scheduling algorithm. Therefore, dynamic solutions, where scheduling decisions are made at runtime have overpassed static allocation strategies. The simplicity and efficiency of dynamic schedulers such as Hadoop are a key of the success of the MapReduce framework. Dynamic schedulers such as StarPU, PaRSEC or StarSs are also developed for more constrained computations, e.g. task graphs coming from linear

algebra. To make their decisions, these runtime systems make use of some static information, such as the distance of tasks to the critical path or the affinity between tasks and computing resources (CPU, GPU,. . .) and of dynamic information, such as where input data are actually located. In this study [16], we concentrate on two elementary linear algebra kernels, namely the outer product and the matrix multiplication. For each problem, we propose several dynamic strategies that can be used at runtime and we provide an analytic study of their theoretical performance. We prove that the theoretical analysis provides very good estimate of the amount of communications induced by a dynamic strategy and can be used in order to efficiently determine thresholds used in dynamic scheduler, thus enabling to choose among them for a given problem and architecture.

6.12. Determining the optimal redistribution

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 optimise 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.

6.13. On the hierarchically structured bin packing problem

We study the hierarchically structured bin packing problem [14]. 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 investigate an existing algorithm and make a correction to the analysis of its approximation ratio. Further results regarding the structure of an optimal solution and a strengthened inapproximability result are given.

6.14. Heuristics for the bipartite matching problem

We propose two heuristics for the bipartite matching problem that are amenable to shared-memory parallelization [18]. The first heuristic is very intriguing from parallelization perspective. It has no significant algorithmic synchronization overhead and no conflict resolution is needed across threads. We show that this heuristic has an approximation ratio of around 0.632. The second heuristic is designed to obtain a larger matching by employing the well-known Karp-Sipser heuristic on a judiciously chosen subgraph of the original graph. We show 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 exploit the structure of the selected subgraphs to propose a specialized implementation which demonstrates a very good scalability. Based on our experiments and theoretical evidence, we conjecture that this second heuristic obtains matchings with cardinality of at least 0.866 of the maximum cardinality. We discuss parallel implementations of the proposed heuristics on shared memory systems. Experimental results, for demonstrating speed-ups and verifying the theoretical results in practice, are provided.

6.15. Fill-in reduction in sparse matrix factorizations using hypergraphs

We discuss the use of hypergraph partitioning based methods in fill-reducing orderings of sparse matrices for Cholesky, LU and QR factorizations [33]. For the Cholesky factorization, we investigate a recent result on pattern-wise decomposition of sparse matrices, generalize the result, and develop algorithmic tools to obtain more effective ordering methods. The generalized results help us formulate the fill-reducing ordering

problem for LU factorization as we do for the Cholesky case, without ever symmetrizing the given matrix A as $|A| + |A^T|$ or $|A^T||A|$. For the QR factorization, we adopt a recently proposed technique to use hypergraph models in a fairly standard manner. The method again does not form the possibly much denser matrix $|A^T||A|$. We also discuss alternatives for LU and QR factorization cases where the symmetrized matrix can be used. We provide comparisons with the most common alternatives in all three cases.

6.16. On partitioning two dimensional finite difference meshes for distributed memory parallel computers

We investigate the problem of partitioning finite difference meshes in two dimensions among the processors of a parallel computer [20]. The objective is to achieve a perfect load balance while minimizing the communication cost. There are well-known graph, hypergraph, and geometry-based partitioning algorithms for this problem. The known geometric algorithms have linear running time and obtain the best results for very special mesh sizes and processor numbers. We propose another geometric algorithm. The proposed algorithm is linear; is applicable to much more cases than some well-known alternatives; obtains better results than the graph partitioning algorithms; obtains better results than the hypergraph partitioning algorithms almost always. Our algorithm also obtains better results than a known asymptotically-optimal algorithm for some small number of processors. We also catalog related theoretical results.

6.17. A symmetry preserving algorithm for matrix scaling

We present an iterative algorithm which asymptotically scales the ∞-norm of each row and each column of a matrix to one [12]. This scaling algorithm preserves symmetry of the original matrix and shows fast linear convergence with an asymptotic rate of 1/2. We discuss extensions of the algorithm to the one-norm, and by inference to other norms. For the 1-norm case, we show again that convergence is linear, with the rate dependent on the spectrum of the scaled matrix. We demonstrate experimentally that the scaling algorithm improves the conditioning of the matrix and that it helps direct solvers by reducing the need for pivoting. In particular, for symmetric matrices the theoretical and experimental results highlight the potential of the proposed algorithm over existing alternatives.

6.18. Direct solvers for sparse linear systems

In the context of the MUMPS sparse direct solver (see Section 5.1), we worked in 2014 on: block-low-rank solvers and shared-memory parallelism [4], [13], hybrid (shared-distributed) parallelism and efficient collective communications in asynchronous environments [2], and scheduling strategies to decrease the memory-usage of multifrontal solvers. Quite significant performance gains have been obtained on up to 2000 cores of a Bullx DLC system (CALMIP mesocentre), some of the corresponding developments will be made available in the next release of our solver. We also worked on setting up a consortium of industrial users to fund engineers working on MUMPS (see Section 7.1). These activities were done in collaboration with INP Toulouse and with CERFACS, CNRS, ENS Lyon, Univ. Bordeaux, EDF, LSTC (Livermore, California) and EMGS (Norway).

RUNTIME Team

6. New Results

6.1. Highlights of the Year

- This year we started very large collaborations with the BULL/Atos company. WE started one European project, one PIA french project and one PhD thesis. The amount of Person Year funded with this project exceed 10. The research we will do with Bull covers resource management, process placement, platform modeling, application modeling, affinity abstraction.
- The StarPU software is used by CEA for automatically distributing linear algebra on their cluster of 144 hybrid nodes.

6.2. Task scheduling over heterogeneous architectures

We continued our work on extending STARPU to master exploitation of Heterogeneous Platforms through dynamic task scheduling, with a now-imminent release of StarPU 1.2.

We have improved the simulation support with SIMGRID, to augment the accuracy of the simulated execution according to the hardware capabilities [30].

We have collaborated with various research projects to leverage the potential of STARPU. We have improved the support for the PASTIX and QR-MUMPS sparse matrix solvers, thus obtaining competitive performance on CPUs and on CPUs+GPUs [25]. We have improved the MPI communication engine of STARPU to get better performance with the EADS hmatrix solver.

We have obtained very good performance and scalability with a Cholesky factorization distributed over a cluster of 144 heterogeneous nodes hosted at CEA.

We have studied the theoretical performance bound that can be achieved for the Cholesky factorization, reproduced the performance of a theorically optimal scheduled, shown that the classical HEFT heuristic is far from it, that more application-specific heuristics allow to get performance closer to the peak, and that the peak is not reachable with simple heuristics, because it requires non-trivial task order inversions.

In relationship with the ADT K'Star effort of building the KLANG-OMP OpenMP compiler and putting together the KASTORS benchmark suite, StarPU has been extended to provide an OpenMP-enabled runtime support for KLANG-OMP. In particular, the StarPU OpenMP Runtime Support implements *preemptible* tasks required for OpenMP, using the concept of continuations, while maintaining interoperability with StarPU regular, non-blocking tasks, and while preserving the heterogeneous, performance model-based scheduling capabilities of StarPU.

The KLANG-OMP C/C++ OpenMP compiler co-developed with Inria Team MOAIS enables plain OpenMP applications to run un-modified on top of the StarPU runtime system, thus significantly increasing the performance portability potential of StarPU.

6.3. Modeling hierarchical platform memory performance with microbenchmarks

Bertrand PUTIGNY developed a new memory performance model based on micro-benchmarks during his PhD. He transforms parallel codes such as OpenMP into memory access skeleton before predicting memory buffer states in caches and using benchmarks outputs to predict the runtime. This model successfully predict the performance behavior of several memory-bound kernels [26].

We also used this model to study the impact of memory caches on the performance on intra-node MPI communication [27].

6.4. Static modeling of clusters of multicore and heterogeneous nodes

We improved the hwloc software to better manage clusters of nodes. This first includes the management of HPC node I/O devices by providing easy ways to retrieve the locality of GPUs and network interfaces. A scalable global view of clusters can be built by factorizing the common topology information that is usually shared by many similar nodes [20]. Finally the topology of the network assembling all these nodes can be exposed in a generic technology-independent manner using the new netloc tool [21] that is now part of hwloc.

6.5. Multithreaded communications

We have proposed a full rewrite of the PIOMAN software, to make it rely on system threads rather than on the now obsolete MARCEL thread scheduler. It makes it more portable, composable with any runtime system used for multithreading, and more scalable. We have shown [19][18] that it features good properties with regard to asynchronous communication progression and multithreaded communications in applications.

6.6. Toplogy-aware load balancing in Charm++

Charm++ implements a fine-grained paradigm based on migratable computing objects. This programming model is designed to run large-scale experiments and provide a dynamic load balancing system to optimize it. Our previous Charm++ load balancer designed for communication-bound applications was improved to scale on large platforms. More precisely, we worked on the network awareness of this algorithm by using LibTopoMap. Our topology-aware load balancing algorithm was also restructured to be parallel and distributed. These enhancements were validated on the Blue Waters supercomputer at Urbana-Champaign, IL. Finally, We have begun to carry out experiments on real application modeling seismic wave propagation.

6.7. Topology-aware ressource allocation

On the one hand SLURM already provides topology aware placement techniques to promote the choice of group of nodes that are placed on the same network level, connected under the same network switch or even placed close to each other so as to avoid long distance communications. On the other hand users can map tasks in a parallel application to the physical processors on the chosen nodes, based on the communication topology.

Our goal is to take in account, in SLURM, placement process, hardware topology, and application communication pattern too. We have implemented a new selection option for the cons_res plugin in SLURM 2.6.5. In this case the usually best fit algorithm used to choose nodes is replaced by Treematch, an algorithm to find the best placement among the free nodes list in light of a given application communication matrix. Tests and evaluation of this feature are in progress.

6.8. Scheduling of dynamic streaming applications on hybrid embedded MPSoCs

The work on the dataflow scheduler has continued so as to improve it: it is now simper and more efficient. Moreover, an H.264 video decoder implementation from STMicroelectronics has been ported onto the developed execution model to conduct more significant experiments. This application exhibits a higher level of complexity and variability, which is the reason why it is well suited for assessing the scheduler's reactivity. Furthermore, an important groundwork has been carried out to enable software support for parts of the application, which enlarges considerably the design space and allow to benefit from better flexibility. In parallel, some earlier work on list scheduling under memory constraints has been extended and published in an international journal [11].

6.9. Performance model for multithreaded applications on multi-core processors

Concerning data locality, researches have shown a tradeof in groupement strategy for process mapping. We have to deal with balanced improvement of several aspects such as threads synchronizations or resource exploitation. Weighting those criterias can only be achieved according to a certain knowledge of both the application and the machine.

Thus, we are working on modeling threads affinity and weights on machines topology to improve a placement method based on the TreeMatch algorithm using new metrics. Several experiences have lead us to the conclusion that it is very hard to identify the key hints and to understand application needs.

Consequently, we are developping a visual tool which displays hardware counters aggregated and mapped on the system topology to identify dynamically those hardware narrows during execution, and understand processes placement effects on them. We hope to achieve a better comprehension of process placement consequences on resources usage by applications.

TYREX Project-Team

6. New Results

6.1. Automated Refactoring for Size Reduction of CSS Style Sheets

Cascading Style Sheets (CSS) is a standard language for stylizing and formatting web documents [17]. Its role in web user experience becomes increasingly important. However, CSS files tend to be designed from a result-driven point of view, without much attention devoted to the CSS file structure as long as it produces the desired results. Furthermore, the rendering intended in the browser is often checked and debugged with a document instance. Style sheets normally apply to a set of documents, therefore modifications added while focusing on a particular instance might affect other documents of the set.

We present a first prototype and a new CSS semantic analyzer and optimizer that is capable of automatically detecting and removing redundant property declarations and rules. We build on earlier work on tree logics to locate redundancies due to the semantics of selectors and properties. Existing purely syntactic CSS optimizers can be used in conjunction with our tool, for performing complementary (and orthogonal) size reduction, toward the common goal of providing smaller and cleaner CSS files. We have been able to detect large numbers of unnecessary property declarations in complex web pages; and we have also found mistakes in the style sheets of some of the most popular web sites. The number of safe modifications can easily grow as more components of CSS are supported and more features are implemented, such as property inheritance, translation of pseudo-classes into query languages, analysis of media queries, merging of equivalent selectors or containment involving grouped selectors.

6.2. Equipping IDEs with XML-Path Reasoning Capabilities

One of the challenges in Web development is to achieve a good level of quality in terms of code size and runtime performance for popular domain-specific languages such as XQuery, XSLT, and XML Schema. We developed an IDE augmented with static detection of inconsistent XPath expressions that assists the programmer with simplifying development and debugging of any application involving XPath expressions [12]. The tool is based on newly developed formal verification techniques based on expressive modal logics, which are now efficient enough to be used in the process of software development. We applied this to a full XQuery compiler for which we introduced an analysis for identifying and eliminating dead code automatically.

6.3. XQuery and Static Typing: Tackling the Problem of Backward Axes

XQuery is a functional language dedicated to XML data querying and manipulation. As opposed to other W3C-standardized languages for XML (e.g. XSLT), it has been intended to feature strong static typing. Currently, however, some expressions of the language cannot be statically typed with any precision. This is due to a discrepancy between the semantics of the language and its type algebra: namely, the values of the language are (possibly inner) tree nodes, which may have siblings and ancestors in the data. The types on the other hand are regular tree types, as usual in the XML world: they describe sets of trees. The type associated to a node then corresponds to the subtree whose root is that node and contains no information about the rest of the data. This makes navigation expressions using “backward axes,” which return e.g. the siblings of a node, impossible to type.

We show how to handle this discrepancy by improving the type system. We describe a logic-based language of extended types able to represent inner tree nodes and show how it can dramatically increase the precision of typing for navigation expressions. We describe how inclusion between these extended types and the classical regular tree types can be decided, allowing a hybrid system combining both type languages. The result is a net increase in precision of typing [20].

6.4. A Core Calculus for XQuery 3.0: Combining Navigational and Pattern Matching Approaches

XML processing languages can be classified according to whether they extract XML data by paths or patterns. The strengths of one category correspond to the weaknesses of the other. In this work, we propose to bridge the gap between these two classes by considering two languages, one in each class: XQuery (for path-based extraction) and CDuce (for pattern-based extraction). To this end, we extend CDuce so as it can be seen as a succinct core λ-calculus that captures XQuery 3.0. The extensions we consider essentially allow CDuce to implement XPath-like navigational expressions by pattern matching and precisely type them. The elaboration of XQuery 3.0 into the extended CDuce provides a formal semantics and a sound static type system for XQuery 3.0 programs [18].

6.5. Session Types as Generic Process Types

Behavioural type systems ensure more than the usual safety guarantees of static analysis [15]. They are based on the idea of "types-as-processes", providing dedicated type algebras for particular properties, ranging from protocol compatibility to race-freedom, lock-freedom, or even responsiveness. Two successful, although rather different, approaches, are session types and process types. The former allows to specify and verify (distributed) communication protocols using specific type (proof) systems; the latter allows to infer from a system specification a process abstraction on which it is simpler to verify properties, using a generic type (proof) system. What is the relationship between these approaches? Can the generic one subsume the specific one? At what price? And can the former be used as a compiler for the latter? This work is a step towards answers to such questions. Concretely, we have defined a stepwise encoding of a pi-calculus with sessions and session types (the system of Gay and Hole) into a pi-calculus with process types (the Generic Type System of Igarashi and Kobayashi). We encode session type environments, polarities (which distinguish session channels end-points), and labelled sums. We show forward and reverse operational correspondences for the encodings, as well as typing correspondences. To faithfully encode session subtyping in process types subtyping, one needs to add to the target language record constructors and new subtyping rules. This work shows how the programming convenience of session types as protocol abstractions can be combined with the simplicity and power of the pi-calculus, taking advantage in particular of the framework provided by the Generic Type System.

6.6. Personal Shopping and Navigator System for Visually Impaired People

We have developed a a personal assistant and navigator system for visually impaired people [14]. This system has been built using a set of domain specific languages based on XML such as OpenStreetMap extended for Augmented Reality. It demonstrate how partially sighted people could be aided by the technology in performing an ordinary activity, like going to a mall and moving inside it to find a specific product. We propose an Android application that integrates Pedestrian Dead Reckoning and Computer Vision algorithms, using an off-the-shelf Smartphone connected to a Smart-watch. The detection, recognition and pose estimation of specific objects or features in the scene derive an estimate of user location with sub-meter accuracy when combined with a hardware-sensor pedometer. The proposed prototype interfaces with a user by means of Augmented Reality, exploring a variety of sensorial modalities other than just visual overlay, namely audio and haptic modalities, to create a seamless immersive user experience. The interface and interaction of the preliminary platform have been studied through specific evaluation methods. The feedback gathered will be taken into consideration to further improve the proposed system.

ASCOLA Project-Team

6. New Results

6.1. Highlights of the Year

Nicolas Tabareau was 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."

Jonathan Pastor has won the joint 1st prize at the Grid5000 Scale challenge, an international challenge for large-scale experiments on geographically-distributed cluster environments. Jonathan has shown with a colleague how to deploy and manage thousands of VMs in such an environment using his approach to fully distributed virtual machine management.

This year we have provided major research results in two domains. First, we have developed several new approaches for the formal reasoning over software in the domains of theorem proving [31], as well as reasoning over distributed interaction protocols [32] and software compositions [24]. Second, we have developed new methods supporting dynamic computations over the cloud, both by means of more elastic cloud applications [27] and better locality management for the dynamic placement of virtual machines in Cloud infrastructures [29].

6.2. Programming Languages

Participants: Ronan-Alexandre Cherrueau, Rémi Douence, Hervé Grall, Thomas Ledoux, 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 published new results extending previous type theories: we have introduced a notion of universe polymorphism for the theorem prover Coq and new type-based mechanisms for the definition and analysis of program equivalences. We have also shown how to harness capabilities, well-known in the security domain, in the context of the functional programming language Haskell. These results are detailed in the current section.

Furthermore, we have applied formal methods and typing in the context of aspect oriented programming ([12], [16], [24]) and in the context of distributed programming (aspectual session types [32]). We have also developed a framework for the formal definition and analysis of accountability properties based on temporal logics. These different results are detailed in Sec. 6.3 for details.

6.2.1.1. Universe Polymorphism in Coq

Universes are used in type theory to ensure consistency by checking that definitions are well-stratified according to a certain hierarchy. In the case of the Coq proof assistant, based on the predicative Calculus of Inductive Constructions (pCIC), this hierarchy is built from an impredicative sort Prop and an infinite number of predicative Type universes. A cumulativity relation represents the inclusion order of universes in the core theory. Originally, universes were thought to be floating levels, and definitions to implicitly constrain these levels in a consistent manner. This works well for most theories, however the globality of levels and constraints precludes generic constructions on universes that could work at different levels. We have introduced universe polymorphism [31] that extends this setup by adding local bindings of universes and constraints, supporting generic definitions over universes, reusable at different levels. This provides the same kind of code reuse facilities as ML-style parametric polymorphism. However, the structure and hierarchy of universes is more complex than bare polymorphic type variables.

6.2.1.2. A Logical Study of Program Equivalence

Proving program equivalence for a functional language with references is a notoriously difficult problem. The goal of the thesis of Guilhem Jaber on "A Logical Study of Program Equivalence" [G. Jaber, Mines Nantes, July 14] was to propose a logical system in which such proofs can be formalized, and in some cases inferred automatically. In the first part, a generic extension method of dependent type theory has been proposed, based on a forcing interpretation seen as a presheaf translation of type theory. This extension equips type theory with guarded recursive constructions, which are subsequently used to reason on higher-order references. In the second part, he has defined a nominal game semantics for a language with higher-order references. It marries the categorical structure of game semantics with a trace representation of denotations of programs, which can be computed operationally and thus have good modularity properties. Using this semantics, he has proven completeness of Kripke logical relations defined in a direct way, using guarded recursive types, without using biorthogonality. The problem of contextual equivalence is then reduced to the satisfiability of an automatically generated formula defined in this logic, that is, to the existence of a world validating this formula. Under some conditions, this satisfiability can be decided using a SMT solver.

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 [25] 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.

6.2.2. Language Mechanisms

In 2014, we have proposed new general language-based mechanisms for concurrent event-based systems and sequential programming languages. Moreover, we have investigated domain-specific languages that support aspect-oriented programming and provide control over propagation strategies in constraint solvers. These results are detailed in the remainder of this section.

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. Concurrent Event-Based Programming

Advanced concurrency abstractions overcome the drawbacks of low-level techniques such as locks and monitors, freeing programmers that implement concurrent applications from the burden of concentrating on low-level details. However, with current approaches the coordination logic involved in complex coordination schemas is fragmented into several pieces including join patterns, data emissions triggered in different places of the application, and the application logic that implicitly creates dependencies among communication channels, hence indirectly among join patterns. In [33], we have presented 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. We have validated the approach with case studies and provided a first performance assessment.

6.2.2.2. Lazy imperative programming

Laziness is a powerful concept in functional programming that permits the reuse of general functions in a specific context, while keeping performance close to the efficiency of dedicated definitions. Lazy evaluation can be used in imperative programming too. Twenty years ago, John Launchbury was already advocating for lazy imperative programming, but the level of laziness of his framework remained limited. Twenty years after, the picture has not changed.

We have proposed an Haskell framework to specify computational effects of imperative programs as well as their dependencies [23]. We have presented a semantics of a call-by-need lambda-calculus extended with imperative strict and lazy features and proved the correctness of our approach. While originally motivated by a less rigid use of foreign functions, we have shown that our approach is fruitful for a simple scenario based on sorted mutable arrays. Furthermore, we can take advantage of equations between algebraic operations to dynamically optimize compositions of imperative computations.

6.2.2.3. 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 [37]. 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.

6.2.2.4. Controlling constraint propagation

Constraint propagation is at the heart of constraint solvers. Two main trends co-exist for its implementation: variable-oriented propagation engines and constraint-oriented propagation engines. These two approaches ensure the same level of local consistency but their efficiency (computation time) can be quite different depending on the problem instances to be solved. However, it is usually accepted that there is no best approach in general, and modern constraint solvers implement only one of them.

In the context of Charles Prud'homme's PhD Thesis [15], we have gone a step further providing a solver independent language at the modeling stage to enable the design of propagation engines. We have validated our proposal with a reference implementation based on the Choco solver and the MiniZinc constraint modeling language.

6.3. Software Composition

Participants: Diana Allam, Walid Benghabrit, Ronan-Alexandre Cherrueau, Rémi Douence, Hervé Grall, Thomas Ledoux, Jean-Claude Royer, Mohamed Sellami, 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 2014, we have developed two general approaches for the definition and enforcement of accountability properties.

6.3.1.1. Logic-based accountability properties

We have proposed a framework for the representation of cloud accountability policies [19]. Such policies offer end-users a clear view of the privacy and accountability obligations asserted by the entities they interact with, as well as means to represent their preferences. This framework comes with two novel accountability policy languages; an abstract one, which is devoted for the representation of preferences/obligations in an human

readable fashion, a concrete one for the mapping to concrete enforceable policies. We motivate our solution with concrete use case scenarios. [30] discusses issues related to data privacy and big data technologies and advocate the use of the framework to support accountability.

We have provided an abstract language for the representation of accountability obligations [20]. We define its semantics using first-order temporal logic and a specific modality for accountability is introduced. We analyze a healthcare use case to illustrate the efficiency of our approach in representing accountability obligations in realistic situations. The use of such services-based applications usually implies the flow of personal data on-line between several parties. In [21], we consider this issue at the design-time of the software and we propose some foundations for an accountable software design. Accountability for a software is a property describing, among other aspects, its liability to end-users for the usage of the data it has been entrusted. We propose to enrich software's component design by accountability clauses using an abstract accountability language (introduced in [20]). We also define conditions for the well-formedness of an accountable component design and show how they can be checked using the μ-CRL model-checker.

6.3.1.2. Defining and enforcing multi-level accountability properties

Many accountability policies require access to all levels of the software stack of service-based applications. Furthermore, they should include explicit means for the definition of cross-domain policies and provide constructive means for the implementation of a wide variety of of accountability properties. These features, in particular, multi-level support, are missing in existing approaches.

We have provided an approach that addresses these objectives explicitly through a language for the definition of expressive regular policies over accountability predicates applicable at all levels of the service stack [22]. Furthermore, we have presented hierarchies of constructive schemes for the implementation of policies for transparency and remediation properties that are implemented in terms of our accountability policy language. Finally, we have shown how to harness the accountability schemes to tackle real-world violations of accountability properties arising from security vulnerabilities of OAuth-based authorization and authentication protocols.

6.3.2. Aspect-Oriented Programming

We have produced in 2014 a range of results enabling reasoning over aspect languages and investigated the use of execution levels. These results are presented in the remainder of this section.

We have also applied ideas from aspect oriented programming in the context of distributed programming (aspectual session types [32]), see Sec. 6.4 .

6.3.2.1. Reasoning about aspect interference using effective aspects

Aspect-oriented programming (AOP) aims at enhancing modularity and reusability in software systems by offering an abstraction mechanism to deal with crosscutting concerns. But, in most general-purpose aspect languages aspects have almost unrestricted power, eventually conflicting with these goals. To tame aspects, we have proposed Effective Aspects: a novel approach to embed the pointcut/advice model of AOP in a statically-typed functional programming language like Haskell; along two main contributions. First, we have defined a monadic embedding of the full pointcut/advice model of AOP [16].

Type soundness is guaranteed by exploiting the underlying type system, in particular phantom types and a new anti-unification type class. In this model aspects are first-class, can be deployed dynamically, and the pointcut language is extensible, therefore combining the flexibility of dynamically-typed aspect languages with the guarantees of a static type system. Monads (which allow the definition of sequences of computations in functional programs) enable us to directly reason about computational effects both in aspects and base programs using traditional monadic techniques. Using this we extend the notion of Open Modules with effects, and also with protected pointcut interfaces to external advising. These restrictions are enforced statically using the type system. Also, we adapt the techniques of EffectiveAdvice to reason about and enforce control flow properties as well as to control effect interference. We show that the parametricity-based approach to effect interference falls short in the presence of multiple aspects and propose a different approach using monad views, a novel technique for handling the monad stack, developed by Schrijvers and Oliveira. Then, we

exploit the properties of our model to enable the modular construction of new semantics for aspect scoping and weaving. Our second contribution [24] builds upon a powerful model to reason about mixin-based composition of effectful components and their interference, based on equational reasoning, parametricity, and algebraic laws about monadic effects. Our contribution is to show how to reason about interference in the presence of unrestricted quantification through pointcuts. We show that global reasoning can be compositional, which is key for the scalability of the approach in the face of large and evolving systems. A comprehensive version of those two works appears in Ismael Figueroa PhD thesis [12].

6.3.2.2. Execution Levels for AOP: from program design to applications

In AOP languages, advice evaluation is usually considered as part of the base program evaluation. This is also the case for certain pointcuts, such as if pointcuts in AspectJ, or simply all pointcuts in higher-order aspect languages like AspectScheme. While viewing aspects as part of base level computation clearly distinguishes AOP from reflection, it also comes at a price: because aspects observe base level computation, evaluating pointcuts and advice at the base level can trigger infinite regression. To avoid these pitfalls, aspect languages propose ad hoc mechanisms, which increase the complexity for programmers while being insufficient in many cases. We have proposed to clarify the situation by introducing levels of execution in the programming language [18], thereby allowing aspects to observe and run at specific, possibly different, levels. We have adopted a defensive default that avoids infinite regression, and gives advanced programmers the means to override this default using level-shifting operators.

6.3.3. Service provisioning

This year, we have provided results on two fundamental problems of service-oriented architectures: service interoperability and service mediation.

6.3.3.1. Service interoperability

Web service support a document-oriented style for clients to interact with a server and promote an environment for systems that is loosely coupled and interoperable. Two models exist for implementing Web services: A process-oriented Web services model, SOAP, and a resource-oriented Web services model, RESTful. Service components are mainly based on description interfaces. These interfaces are often known as structural standardized interfaces like WSDL for SOAP and WADL for RESTful. The implementation of Web services is increasingly based on object-oriented (OO) frameworks, at the client and the server sides. Using these frameworks, developers can transform an object code into a Web service, or access a remote Web service, at the touch of a button. In this context, two levels are present: an object level built over a service level.

Diana Allam's PhD thesis [11] has focused on two properties of these frameworks:

- The loose coupling between the two levels, which allows the complex technical details of the service level to be hidden at the object level and the service level to be evolved with a minimal impact on the object level.
- The interoperability induced by the substitution principle associated to subtyping in the object level, which allows to freely convert a value of a subtype into a value of a supertype.

The thesis provides three contributions in this context. We propose a unified formal model for web services based on message passing and enabling first class channels. It is equipped with a powerful type-checking allowing union, intersection and negation operations as well as subtyping. The type checking algorithm relies on the semantic approach defined by G. Castagna. This type system is also protected against attackers. The second contribution is a concrete refinement of the model into RESTful and SOAP frameworks as well as a unified API for service discovery. To define such an API, we have first shown how the details of the standard interfaces (WSDL and WADL) could be simplified and abstracted and then we rely on subtyping in the discovery mechanism. Finally, to solve some of the interoperability issues between the OO level and the service level a formalization of the binding using categorical concepts (commutative diagrams) is proposed. Based on this an analysis of the mismatch problems has been done and a new specification of the data binding has been formalized. The document then discusses some variations in the implementation of the data binding solution and a prototype for the Apache CXF framework.

Mayleen Lacouture's PhD thesis "A Chemical Programming Language for Orchestrating Services - Application to Interoperability Problems" [M. Lacouture, MN/U. Nantes, Oct. 14] proposes a framework easing interoperability in the form of an architecture that integrates different orchestration languages with heterogeneous service providers around a pivot language. The pivot language is implemented as a new orchestration language based on the chemical programming paradigm. Concretely, the dissertation presents a language called Criojo that implements and extends the Heta-calculus, an original calculus associated to a chemical abstract machine dedicated to service-oriented computing. The consequence of adopting this approach would be an improvement in the interoperability of services and orchestration languages, thus easing the development of composite services. The high level of abstraction of Criojo could allow developers to write very concise orchestrations since message exchanges are represented in a natural and intuitive way.

6.3.3.2. Service mediation

Service composition is a major advance service-oriented computing brings to enable the development of distributed applications. However, the distributed nature of services hampers their composition with data heterogeneity problems. We address these problems with a decentralized Mediation-as-a-Service architecture that solves data inconsistencies occurring during the composition of business services [17]. As an extension to our previous work that focused on data interpretation problems, we present in this paper a solution to solve data inconsistencies at the syntactic, structural and semantic levels. We show how syntactic, structural and semantic mediation techniques can be combined, and how semantic mediation provides useful information that helps structural and syntactic mediation. We demonstrate how our architecture enables decentralized publication and discovery of mediation services. We motivate our work with a concrete scenario and validate our proposal with experiments.

6.3.4. Software product line architectures

Software product lines were designed from the product line tested out by H. Ford at the beginning of the 20TH century, which led to the success of his automotive production. For 15 years, these methods have been visible in several software application fields: telephony at Nokia, televisions at Philips, print software at HP and flight applications at Boeing, among others. The concept of architecture is crucial for classic software applications, and this concept is even more important at the level of domain engineering in product lines. In a product line, the so-called reference architecture generically describes the architectures of all the products in the family. The chapter [34] describes the technical means and methods for defining a reference architecture for a software product line. It also presents the methods for operating this architecture through, for example, techniques emerging from model and software component engineering, or aspect-oriented programming. These concepts and techniques are illustrated using a case study.

6.4. Cloud applications and infrastructures

Participants: Adrien Lebre, Thomas Ledoux, Yousri Kouki, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Flavien Quesnel, Mario Südholt.

In 2014, we have provided solutions for Cloud-based and distributed programming, virtual environments and data centers, in particular concerning energy-optimal Cloud applications.

6.4.1. Cloud and distributed programming

This year we have published results on a broker that provides better guarantees on service-level agreements in the Cloud. Furthermore, we have extended a class of formally-defined protocols, session types.

6.4.1.1. Service-level agreement for the Cloud

Elasticity is the intrinsic element that differentiates Cloud Computing from traditional computing paradigms, since it allows service providers to rapidly adjust their needs for resources to absorb the demand and hence guarantee a minimum level of Quality of Service (QoS) that respects the Service Level Agreements (SLAs) previously defined with their clients. However, due to non-negligible resource initiation time, network fluctuations or unpredictable workload, it becomes hard to guarantee QoS levels and SLA violations may occur.

We propose a language support for Cloud elasticity management that relies on CSLA (Cloud Service Level Agreement) [27]. CSLA offers new features such as QoS/functionality degradation and an advanced penalty model that allow providers to finely express contracts so that services self-adaptation capabilities are improved and SLA violations minimized. The approach was evaluated with a real infrastructure and application testbed. Experimental results show that the use of CSLA makes Cloud services capable of absorbing more peaks and oscillations by trading-off the QoS levels and costs due to penalties.

6.4.1.2. AO session types for distributed protocols

Multiparty session types allow the definition of distributed processes with strong communication safety properties. A global type is a choreographic specification of the interactions between peers, which is then projected locally in each peer. Well-typed processes behave accordingly to the global protocol specification. Multiparty session types are however monolithic entities that are not amenable to modular extensions. Also, session types impose conservative requirements to prevent any race condition, which prohibit the uniform application of extensions at different points in a protocol. We have proposed a means to support modular extensions with aspectual session types [32], a static pointcut/advice mechanism at the session type level. To support the modular definition of crosscutting concerns, we have augmented the expressivity of session types to allow harmless race conditions. As a result, aspectual session types make multiparty session types more flexible, modular, and extensible.

6.4.2. Virtualization and data centers

In 2014, we have produced a variety of results on a new model for utility computing that addresses fundamental shortcomings of today's Cloud computing model. Furthermore, we have provided more powerful techniques for the virtualization of computations and the management of cluster-based environments, such as data centers.

6.4.2.1. Next generation utility computing

To accommodate the ever-increasing demand for Utility Computing (UC) resources while taking into account both energy and economical issues, the current trend consists in building larger and larger data centers in a few strategic locations. Although such an approach enables to cope with the actual demand while continuing to operate UC resources through centralized software system, it is far from delivering sustainable and efficient UC infrastructures. Throughout the Discovery initiative [0], we investigate how UC resources can be managed differently, considering locality as a primary concern. Concretely, we study how it can be possible to leverage any facilities available through the Internet in order to deliver widely distributed UC platforms that can better match the geographical dispersal of users as well as the unending resource demand. Critical to the emergence of such locality-based UC (LUC) platforms is the availability of appropriate operating mechanisms. We presented a prospective vision of a unified system driving the use of resources at an unprecedented scale by turning a complex and diverse infra structure into a collection of abstracted computing facilities that is both easy to operate and reliable [35]. By deploying and using such a LUC Operating System on backbones, our ultimate vision is to make possible to host/operate a large part of the Internet by its internal structure itself: A scalable and nearly infinite set of resources delivered by any computing facilities forming the Internet, starting from the larger hubs operated by ISPs, governments and academic institutions to any idle resources that may be provided by end-users. We highlight that this work is conducted through a collaboration between the ASAP, ASCOLA, AVALON and MYRIADS Inria Project-teams.

6.4.2.2. Adding locality capabilities to virtual machine schedulers

Through the DVMS proposal, we showed in 2013 the benefit of leveraging peer-to-peer algorithms to design and implement virtual machines (VMs) scheduling algorithms. Although P2P based proposals considerably improve the scalability, leading to the management of hundreds of thousands of VMs over thousands of physical machines (PMs), they do not consider the network overhead introduced by multi-site infrastructures. This over- head can have a dramatic impact on the performance if there is no mechanism favoring intra-site v.s. inter-site manipulations. This year, we extended our DVMS mechanism with a new building block designed on top of the Vivaldi coordinates mechanism. We showed its benefits by discussing several experiments performed

[0] http://beyondtheclouds.github.io

on four distinct sites of the Grid'5000 testbed. With our proposal and without changing the scheduling decision algorithm, the number of inter-site operations has been reduced by 72% [29]. This result provides a glimpse of the promising future of using locality properties to improve the performance of massive distributed Cloud platforms. We highlight that this work has been performed in collaboration with the ASAP, ASCOLA, AVALON and MYRIADS Inria Project-teams.

6.4.2.3. WAN-wide elasticity capabilities for distributed file systems

Applications dealing with huge amounts of data suffer significant performance impacts when they are deployed on top of an hybrid platform (i.e the extension of a local infrastructure with external cloud resources). More precisely, through a set of preliminary experiments we shew that mechanisms which enable on demand extensions of current Distributed File Systems (DFSes) are required. These mechanisms should be able to leverage external storage resources while taking into account the performance constraints imposed by the physical network topology used to interconnect the different sites. To address such a challenge we presented the premises of the Group Based File System, a glue providing the elasticity capability for storage resources by federating on demand any POSIX file systems [28].

6.4.3. Energy optimization

Demand for Green services is increasing considerably as people are getting more environmental conscious to build a sustainable society. Therefore, enterprise and clients want to shift their workloads towards green Cloud environment offered by the Infrastructure-as-a-Service (IaaS) provider. The main challenge for an IaaS provider is to determine the best trade-off between its profit while using renewable energy and customers satisfaction. In order to address this issue, we propose a *Cloud energy broker* [26], which can adjust the availability and price combination to buy Green energy dynamically from the market to make datacenter green. Our energy broker tries to maximize of using renewable energy under strict budget constraint whereas it also tries to minimize the use of brown energy by capping the limit of overall energy consumption of datacenter. The energy broker was evaluated with a real workload traced by PlanetLab. Experimental results show that our energy broker successfully enables meeting the best trade-off.

DIVERSE Project-Team

6. New Results

6.1. Highlights of the Year

"Globalizing Modeling Languages" appears in IEEE Computer Magazine. This paper synthesizes our vision of how domain-specific languages form the foundations of global software development. Its appearance in a highly visible venue is major milestone for the dissemination and impact of our work about the diversity of languages.

DiverSE extremely present at the SPLC conference. SPLC is the main international conference for software product line engineering. In 2014, the DiverSE team had a very strong presence at this conference, presenting novel scientific contributions, results of industrial collaborations, and demonstrations of latest software tools.

6.2. Results on Software Language Engineering

The engineering of systems involves many different stakeholders, each with their own domain of expertise. Hence more and more organizations are adopting Domain Specific Languages (DSLs) to allow domain experts to express solutions directly in terms of relevant domain concepts. This new trend raises new challenges about designing DSLs, evolving a set of DSLs and coordinating the use of multiple DSLs for both DSL designers and DSL users. In [56] we present the overall vision that we develop in the DiverSE team about Software Language Engineering. The main results on this topic are presented below.

6.2.1. Globalization of Domain Specific Languages

In the software and systems modeling community, research on domain-specific modeling languages (DSMLs) focuses on technologies for developing languages and tools to increase the effectiveness of domain experts. Yet, there is a lack of support to explicitly relate concepts expressed in different DSMLs, which prevents software and system engineers to reason about information spread across models describing different system aspects. Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages [20]. In such a context, we develop a research initiative that broadens the DSML research focus beyond the development of independent DSMLs to one that supports globalized DSMLs, that is, DSMLs that facilitate coordination of work across different domains of expertise. We also provid a formal framework to prove the correctness of model driven engineering composition operators [57].

6.2.2. Meta-Language for the Concurrency Concern in DSLs

Concurrency is of primary interest in the development of complex software-intensive systems, as well as the deployment on modern platforms. However, reifying the definition of the DSL concurrency remains a challenge. This hinders: a) the development of a complete understanding of the DSL semantics; b) the effectiveness of concurrency-aware analysis techniques; c) the analysis of the deployment on parallel architectures. In this context, we present MoCCML, a dedicated meta-language for formally specifying the concurrency concern within the definition of a DSL [44]. 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. Mo CCML is illustrated on the definition of an lightweight extension of SDF (SynchronousData Flow).

6.2.3. Automating Variability Model Inference for Component-Based Language Implementations

Componentized language frameworks, coupled with variability modeling, have the potential to bring language development to the masses, by simplifying the configuration of a new language from an existing set of reusable components. However, designing variability models for this purpose requires not only a good understanding of these frameworks and the way components interact, but also an adequate familiarity with the problem domain. In [68] we propose an approach to automatically infer a relevant variability model from a collection of already implemented language components, given a structured, but general representation of the domain. We describe techniques to assist users in achieving a better understanding of the relationships between language components, and find out which languages can be derived from them with respect to the given domain.

6.2.4. Metamorphic Domain-Specific Languages

External or internal domain-specific languages (DSLs) or (fluent) APIs? Whoever you are – a developer or a user of a DSL – you usually have to choose side; you should not! What about metamorphic DSLs that change their shape according to your needs? Our 4-years journey of providing the "right" support (in the domain of feature modeling), led us to develop an external DSL, different shapes of an internal API, and maintain all these languages. A key insight is that there is no one-size-fits-all solution or no clear superiority of a solution compared to another. On the contrary, we found that it does make sense to continue the maintenance of an external and internal DSL. Based on our experience and on an analysis of the DSL engineering field, the vision that we foresee for the future of software languages is their ability to be self-adaptable to the most appropriate shape (including the corresponding integrated development environment) according to a particular usage or task. We call metamorphic DSL such a language, able to change from one shape to another shape [27].

6.2.5. Adapting mutation testing for model transformations

Due to the specificities of models and transformations, classical software testing techniques have to be adapted. Among these techniques, mutation analysis has been ported and a set of mutation operators has been defined. However, mutation analysis currently requires a considerable manual work and is hampered by the test data set improvement activity. This activity is seen by testers as a difficult and time-consuming job, and reduces the benefits of the mutation analysis.

We provide a model transformation traceability mechanism, in conjunction with a model of mutation operators and a dedicated algorithm, to automatically or semi-automatically produce test models that detect new faults [18].

6.2.6. Efficient model cloning for analysis

We propose an original approach that exploits the fact that operations rarely modify a whole model. Given a set of immutable properties, our cloning approach determines the objects and fields that can be shared between the runtime representations of a model and its clones. Our generic cloning algorithm is parameterized with three strategies that establish a trade-off between memory savings and the ease of clone manipulation. We evaluated memory footprints and computation overheads with 100 randomly generated metamodels and models [40]. We have also drawn the research roadmap to exploit these efficient clone operations to analyze multidimensional execution traces [41].

6.3. Results on Variability Modeling and Engineering

6.3.1. Engineering Interactive Systems

In agreement with our permanent effort to validate the techniques we propose on real use cases in various domains, we applied seminal MDE to interactive systems engineering. This led to two collaborations. The first one has been conducted with 3D Collaborative Virtual Environments (3D CVE) researchers. Despite the increasing use of 3D CVE, their development is still a cumbersome task. The various concerns to consider

(distributed system, 3D graphics, *etc.*) complexify their development as well as their evolution. We propose to leverage MDE for developing 3D CVEs [45]. We have shown how a 3D CVE framework benefits from a DSL we built using state-of-the-art MDE technologies. The benefits are multiple: 3D CVEs designers can focus on the behavior of their virtual objects without bothering with distributed and graphics features; configuring the content of 3D CVEs and their deployment on various software and hardware platforms can be automated through code generation.

The second collaboration is international and has been conducted with software visualization researchers. Current metamodel editing tools are based on standard visualization and navigation features, such as physical zooms. However, as soon as metamodels become larger, navigating through large metamodels becomes a tedious task that hinders their understanding. In this work, we promote the use of model slicing techniques [102] to build visualization techniques dedicated to metamodels [37]. This approach is implemented in a metamodel visualizer, called `Explain`.

6.3.2. Variability management in regulatory requirements and system engineering

Nuclear power plants are some of the most sophisticated and complex energy systems ever designed. These systems perform safety critical functions and must conform to national safety institutions and international regulations. In many cases, regulatory documents provide very high level and ambiguous requirements that leave a wide margin for interpretation. As the French nuclear industry is now seeking to spread its activities outside France, it is but necessary to master the ins and the outs of the variability between countries safety culture and regulations. This sets both an industrial and a scientific challenge to introduce and propose a product line engineering approach to an unaware industry whose safety culture is made of interpretations, specificities, and exceptions. We have developed two contributions within the French R&D project CONNEXION, while introducing variability modeling to the French nuclear industry [66], [34].

As part of the VaryMDE project (a bilateral collaboration between Thales and Inria) we have developed techniques to generate counter-examples (also called anti-patterns) of model-based product lines [22]. The goal is to infer (1) guidelines or domain-specific rules to avoid earlier the specification of incorrect mappings (2) testing oracles for increasing the robustness of derivation engines given a modeling language. We have applied the approach in the context of a real industrial scenario with Thales involving a large-scale metamodel.

6.3.3. Handling testing challenges in product line engineering

Testing techniques in industry are not yet adapted for product line engineering (PLE).

We have developed original contributions to adapt model-based testing for PLE [65], [63], [13]. We equip usage models, a widely used formalism in MBT, with variability capabilities. Formal correspondences are established between a variability model, a set of functional requirements, and a usage model. An algorithm then exploits the traceability links to automatically derive a usage model variant from a desired set of selected features. The approach is integrated into the MBT tool MaTeLo and is currently used in industry.

We have also developed a variability-based testing approach to derive video sequence variants. The ideas of our VANE approach are i) to encode in a variability model what can vary within a video sequence; ii) to exploit the variability model to generate testable configurations; iii) to synthesize variants of video sequences corresponding to configurations. VANE computes T-wise covering sets while optimizing a function over attributes [50], [25].

6.3.4. Reverse engineering variability models

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 [42]. We have performed an empirical evaluation on hundreds of feature models, coming from the SPLOT repository and Wikipedia [108]. 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 [19]). Beyond the reverse engineering of variability, our work has numerous practical applications (e.g., merging multiple product lines, slicing a configuration process).

6.3.5. 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. Despite their apparent simplicity, PCMs contain heterogeneous, ambiguous, uncontrolled and partial information that hinders their efficient exploitations. We have first elaborated our vision and identify research challenges for an exploitation of PCMs when engineering comparators, configurators, or other services [67].

We have formalized PCMs through model-based automated techniques and developed additional tooling to support the edition and re-engineering of PCMs [43]. 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 http://www.opencompare.org

6.4. Results on Heterogeneous and dynamic software architectures

This year, we focused on the challenges that use *models@runtime* for resource-constrained and resource-aware systems. Our main results are in the following four subdomains:

- We designed an adaptive monitoring framework for component-based systems in which we highlight the benefits of using models@runtime for adaptive monitoring.
- We improved models@runtime technologies for resource-constrained devices.
- We designed efficient reasoning techniques for dynamic software architecture, focusing in particular on resource consumption optimization challenges.
- We performed several experiments on the Internet of Things application domain.

The next section details our experiments.

6.4.1. Resource-aware dynamic architecture

Modern component frameworks support continuous deployment and simultaneous execution of multiple software components on top of the same virtual machine. However, isolation between the various components is limited. A faulty version of any one of the software components can compromise the whole system by consuming all available resources. We propose a solution to efficiently identify faulty software components running simultaneously in a single virtual machine. It is based on an optimistic adaptive monitoring system to identify the faulty component. Suspected components are instrumented to obtain fined grain data for deeper analysis by the monitoring system, but only when required. Unsuspected components are left untouched and execute normally. Thus, we perform localized, just-in-time monitoring that decreases the accumulated overhead of the monitoring system. We evaluated our approach against a state-of-the-art monitoring system and we have shown that our technique correctly detects faulty components, while reducing overhead by an average of 80% [52]. Based on this work, we have presented two tutorials at the CBSE/QoSA conference [49] and at the Middleware conference [51].

6.4.2. Technology enablers for resource-aware dynamic software architecture

Models@runtime provides semantically rich reflection layers enabling intelligent systems to reason about themselves and their surrounding context. Most reasoning processes require not only to explore the current state, but also the past history to take sustainable decisions e.g. to avoid oscillating between states. Models@runtime and model-driven engineering in general lack native mechanisms to efficiently support the notion of history, and current approaches usually generate redundant data when versioning models, which reasoners need to navigate. Because of this limitation, models fail in providing suitable and sustainable abstractions to

deal with domains relying on history-aware reasoning. This work tackles this challenge by considering history as a native concept for modeling foundations. Integrated in conjunction with lazy load/storage techniques into the Kevoree Modeling Framework, we demonstrated onto a energy-aware smart grid case study that this mechanisms enable a sustainable reasoning about massive historized models [53].

In this field we also created a specific extension to the `docker.io` open-source project to support a dynamic resource reservation of running containers [9]

6.4.3. *Efficient reasoning techniques for dynamic software architecture*

Providing software with the capacity of adapting itself according to its environment requires effective techniques to reason and decide on what adaptation to undertake over the running system. To decide on a system adaptation, we have to characterize the value of the system in its corresponding execution environment. A system cannot be characterized by a single dimension, but only using several dimensions such as performance, energy consumption, security and so on. In this context, we have proposed various techniques to leverage multi-objective evolutionary algorithms both at deployment time [46], [21] and at runtime [47] to enable system optimization using multidimensional optimization. We have also proposed a technique to adapt a system proactively based on predictions in order to prevent failures [60]

6.4.4. *The Internet of Things application domain*

We apply our techniques for heterogeneous and dynamic software architecture more specifically to the Internet of Things application domain. We have two main contributions: (1) an application of the models@runtime concepts on embedded nodes with very limited resources for memory, CPU and battery [30], and (2) a study on the problem of renewable energy production and consumption at home [39]. Domestic microgeneration is the onsite generation of low and zero-carbon heat and electricity by private households to meet their own needs. In this paper we explore how an everyday household routine (doing laundry) can be augmented by digital technologies to help households with photovoltaic solar energy generation to make better use of self-generated energy. We present an 8 month in the field study that involved 18 UK households in longitudinal energy data collection, prototype deployment and participatory data analysis [38]. Through a series of technology interventions mixing energy feedback, proactive suggestions and direct control, the study uncovered opportunities, potential rewards and barriers for families to shift energy consuming household activities. The study highlights how digital technology can act as a mediator between household laundry routines and energy demand-shifting behaviors. Finally, the study provides insights into how a "smart" energy-aware washing machine shapes organization of domestic life and how people "communicate" with their washing machine.

6.5. 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 new approaches to increase diversity in software systems.

6.5.1. *Automatic synthesis of computationally diverse program variants*

The predictability of program execution provides attackers with a rich source of knowledge that they can exploit to spy or remotely control the program. Moving target defense addresses this issue by constantly switching between many diverse variants of a program, thus reducing the certainty that an attacker can have about the program execution. The effectiveness of this approach relies on the availability of a large number of software variants that exhibit different executions. However, current approaches rely on the natural diversity provided by off-the-shelf components, which is very limited. We have explored the automatic synthesis of large sets of program variants, called *sosies* [32]. Sosies provide the same expected functionality as the original program, while exhibiting different executions. They are said to be computationally diverse.

6.5.2. Software Evolution for Diversity Emergence

We aim at favoring spontaneous diversification in software systems, to increase their adaptive capacities. This objective is founded on three observations: (1) software has to constantly evolve to face unpredictable changes in its requirements, execution environment or to respond to failure (bugs, attacks, etc.); (2) the emergence and maintenance of high levels of diversity are essential to provide adaptive capacities to many forms of complex systems, ranging from ecological and biological systems to social and economical systems; (3) diversity levels tend to be very low in software systems. In this work [33], we consider evolution as a driver for diversity as a means to increase resilience in software systems. In particular, we are inspired by bipartite ecological relationships to investigate the automatic diversification of the server side of a client-server architecture.

6.5.3. Analyzing the diversity of development practices in open source projects

Decentralized version control systems allow a rich structure of commit histories, which presents features that are typical of complex graph models. We bring some evidences of how the very structure of these commit histories carries relevant information about the distributed development process. By means of a novel data structure that we formally define, we analyze the topological characteristics of commit graphs of a sample of git projects. Our findings point out the existence of common recurrent structural patterns that identically occur in different projects and can be considered building blocks of distributed collaborative development [36], [35].

FOCUS Project-Team

6. New Results

6.1. Highlights of the Year

Valeria Vignudelli has received the AILA (Associazione Italiana di Logica e sue Applicazioni) award for her 2014 master thesis.

6.2. Service-oriented computing

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

6.2.1. 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. Coloured Petri nets (CPN) offer a formal yet easy tool for modelling interactions in SOAs, however mapping abstract SOAs into executable ones requires a non-trivial and time-costly analysis. In [34], we propose a methodology that maps CPN-modelled SOAs into Jolie SOAs (our target language), exploiting a collection of recurring control-flow patterns, called Workflow Patterns, as composable blocks of the translation. We validate our approach with a realistic use case. In addition, we pragmatically assess the expressiveness of Jolie with respect to the considered WPs.

6.2.2. 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 [18] we surveyed the work on choreographies and behavioural contracts in multiparty interactions. In [28] we explored the notion of deadlock freedom (the system never gets stuck), and the related notions of lock freedom (each action is eventually executed under a fair scheduling) and progress (each session never gets stuck). Previous work studied how to define progress in an open setting by introducing the notion of catalysers, execution contexts generated from the type of a process. We refined the notion of catalysers leading to a novel characterization of progress in terms of the standard notion of lock-freedom. We applied our results both to binary session types and in an untyped session-based setting. We combined our results with existing techniques for lock-freedom, obtaining a new methodology for proving progress. Our methodology captures new processes w.r.t. previous progress analysis based on session types. The two following works consider the extension of choreographies, which traditionally have a static structure, to deal with adaptation, i.e., dynamic changes of the structure of choreographies. A preliminary analysis of adaptable choreographies at the level of types is presented in [27]. This work considers both updates from inside the system (self-adaptation), and external updates. Adaptable choreographies as a programming language are considered in [33], where we presented AIOCJ, a framework for programming distributed adaptive applications. AIOCJ allows the programmer to specify which parts of the application can be adapted. Adaptation takes place at run-time by means of rules, which can change during the execution to tackle possibly unforeseen adaptation needs. AIOCJ relies on a solid theory that ensures applications to be deadlock free by construction also after adaptation.

6.3. Models for reliability

Participants: Mario Bravetti, Elena Giachino, Ivan Lanese, Michael Lienhardt, Gianluigi Zavattaro.

6.3.1. Reversibility

We have continued the study of causal-consistent reversibility started in the past years. In [17] we presented an overview of causal-consistent reversibility, summarizing the main approaches in the literature, and the related results and applications. An interesting application is debugging. Reversible debugging provides developers with a way to execute their applications both forward and backward, seeking the cause of a misbehaviour. In a concurrent setting, reversing actions in the exact reverse order they have been executed may lead to undo many actions that were not related to the bug under analysis. On the other hand, undoing actions in some order that violates causal dependencies may lead to states that could not be reached in a forward execution. In [36] we proposed a new approach, where each action can be reversed if all its consequences have already been reversed. The main feature of the approach is that it allows the programmer to easily individuate and undo exactly the actions that caused a given misbehaviour till the corresponding bug is reached. We discussed the appropriate primitives for causal-consistent reversible debugging and presented their prototype implementation in the CaReDeb tool.

6.3.2. Fault models

We have continued the study of primitives for fault handling in a concurrent setting. In [19] we critically discussed the different choices that have to be made when defining a fault model for a concurrent object-oriented programming language. We consider in particular the ABS language, and analyse the interplay between the fault model and the main features of ABS, namely the cooperative concurrency model, based on asynchronous method invocations whose return results via futures, and its emphasis on static analysis based on invariants.

6.4. Cloud Computing

Participants: Roberto Amadini, Maurizio Gabbrielli, Elena Giachino, Saverio Giallorenzo, Claudio Guidi, Cosimo Laneve, Michael Lienhardt, Tudor Alexandru Lascu, Jacopo Mauro, Gianluigi Zavattaro.

6.4.1. Cloud application deployment

Configuration and management of applications in the cloud is a complex task that requires novel methodologies and tools. In [16] we have performed a foundational study of the complexity boundaries for the automatic deployment problem, showing that in the general case this problem is undecidable, it is decidable but non-primitive recursive if capacity constraints are not taken into account, while it turns out to be polynomial time if also conflicts between software components are not considered. Starting from these foundational observations, we have investigated the exploitability in this specific context of state-of-the-art constraint optimization techniques, a well established approach for the modeling and solution of complex optimization problems. In particular, in [23], [24] we have studied how the "portfolio technique" approach can be applied to optimization problems, combining and exploiting the performances of existing solvers to get a global, more robust and fast solver. Encouraged by these results, we have developed SUNNY-CP [13], [22]: a porftolio constraint solver for constraint satisfaction and optimization problems. SUNNY-CP has proven to have remarkable performances, ranking 4th in the annual MiniZinc challenge (i.e., the international competition to evaluate the performances of constraint solvers) and receiving a ‘honorable’ mention by the challenge organizers.

6.4.2. Cloud resource management

The management of cloud resources from client programs requires the definition of Application Programming Interfaces (APIs) that expose specific functionalities to external invokers. Programs can be built that compose existing APIs in order to obtain new functionalities. However API composition easily becomes a frustrating and time-costly task that hinders API reuse. The issue derives from technology-dependent features of API composition such as the need of extensive documentation, protocol integration, security issues, etc.. In [39] we introduce the perspective of the API-as-a-Service (APIaaS) layer as tool to ease the development and deployment of applications based on API composition, abstracting communication protocols and message formats. We elicit the desirable features of such a layer and provide a proof-of-concept prototype implemented using a service-oriented language.

Another critical aspect in this context deals with the problem of dynamic reallocation of resources. In [38] we study a type-based technique for modeling and analysis of systems in which concurrent object-oriented programs dynamically create and move resources. The type of a program is behavioural, namely it expresses the resource deployments over periods of (logical) time. Our technique admits the inference of types and may underlie the optimisation of the costs and consumption of resources.

6.5. Resource Control

Participants: Michele Alberti, Alberto Cappai, Ugo Dal Lago, Simone Martini, Giulio Pellitta, Davide Sangiorgi, Marco Solieri, Valeria Vignudelli.

6.5.1. Probabilistic higher-order calculi

The first results of our efforts on probabilistic higher-order systems and languages have started to appear in 2014. In particular, we have focused our attention on the impact of probability to the classical notion of context equivalence for the lambda-calculus, showing that applicative bisimilarity continues to be a congruence [31], and that it even coincides with context equivalence when evaluation is done in the call-by-value order [29]. The expressive power of higher-order concurrent contexts has been compared to the expressive power of lambda-calculi contexts and put in relation with other equivalences when the observed process is either an ordinary Labelled Transition Systems (LTS) or a reactive probabilistic transition system [25]. The obtained spectrum of equivalences for reactive probabilistic processes has been shown to be finer than the one for classic LTSs. We have also analysed the expressive power of different first-order testing equivalences (with nondeterministic tests, probabilistic tests, and both nondeterministic and probabilistic tests) in the spectrum for reactive probabilistic processes [26].

6.5.2. Resource consumption

The main result about resource consumption has been about an open problem on the λ-calculus: we proved that the number of leftmost-outermost steps to normal form is indeed an invariant cost model in the sense of Slot and van Emde Boas' weak invariance thesis [21]. We also introduced a new recursion theoretic framework for probabilistic computation in which one is able to capture probabilistic polynomial time through Leivant's Tiering [32].

6.5.3. Geometry of interaction

Novel results have been obtained for Geometry of Interaction (GoI), itself a semantics framework for linear logic introduced by Jean-Yves Girard thirty years ago. In particular, we have shown how the most concrete presentations of GoI, namely so-called token machines, can go *parallel*, thus exploiting the potential parallelism in functional programs (through the Curry-Howard Correspondence). This has been made concrete by studying extensions of multiplicative linear logic in which synchronization becomes an operator where tokens can indeed synchronize [30]. This has been later shown to be necessary to model quantum computation [44]. A simple, minimalistic GoI model of the resource λ-calculus has also been introduced [43].

6.6. Verification techniques for extensional properties

Participants: Daniel Hirschkoff, Elena Giachino, Michael Lienhardt, Cosimo Laneve, Jean-Marie Madiot, Davide Sangiorgi.

Extensional refers to properties that have to do with behavioural descriptions of a system (i.e., how a system looks like from the outside). Examples of such properties include classical functional correctness and deadlock freedom. Related to techniques for extensional properties are the issues of decidability (the problem of establishing whether certain properties are computationally feasable).

6.6.1. 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. We have transported this theory onto languages whose bisimilarity and LTS go beyond those of first-order models [40]. The approach consists in exhibiting fully abstract translations of the more sophisticated LTSs and bisimilarities onto the first-order ones. This allows us to reuse directly the large corpus of up-to techniques that are available on first-order LTSs. We have investigated the method on the π-calculus, the Higher-Order π-calculus, and a (call-by-value) λ-calculus with references.

In [20], mostly a tutorial paper, a few forms of bisimulation and of coinductive techniques that have been proposed for higher-order languages are discussed, beginning with the pure lambda-calculus and then moving to extensions of it, notably those with non-determinism and probabilities.

6.6.2. Deadlock detection

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 [37], [35], in a basic model featuring recursion and fresh name generation, called Lam. We then have designed a type system that associates Lams to processes. As a byproduct of these two techniques, we have an algorithm that is more powerful than previous ones and that can be easily integrated into the current release of TyPiCal, a type-based analyser for π-calculus.

6.6.3. Expressiveness and decidability in actor-like systems

Refining work in previous years, we have studied [15] the expressive power of an actor-like language, featuring concurrent objects and asynchronous message-passing. We have identified the presence/absence of fields as a crucial feature: the dynamic creation of names in combination with fields gives rise to Turing completeness. On the other hand, restricting to stateless actors gives rise to systems for which properties such as termination are decidable. This decidability result still holds for actors with states when the number of actors is bounded and the state is read-only.

INDES Project-Team

6. New Results

6.1. Web programming

Participants: Yoann Couillec, Vincent Prunet, Tamara Rezk, 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 [17].

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 nodejs, 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.

A prototype version of Hop.js is currently used by several academic and SME R&D teams to jointly develop an assistive robotic platform and a set of distributed applications.

We plan to release the first public Hop.js version by the end of the first semester of 2015, as we plan to start describing in forthcoming papers.

6.1.2. Multitier Debugging

Debugging Web applications is difficult because of their distributed nature and because the server-side and the client-side of the application are generally treated separately. The multitier approach, which reunifies the two ends of the application inside a unique execution environment, helps the debugging process because it lets the debugger access more runtime informations.

Based on our previous work on the Hop multitier debugger [17], we have built a multitier debugger for Hop.js, our multitier extension of JavaScript. Its advantage over most debuggers for the Web is that it reports the full stack trace containing all the server-side and client-side frames that have conducted to an error. Errors are reported on their actual position on the source code, wherever they occur on the server or on the client. This paper presents this debugger and sketches its implementation. This work is described in a yet unpublished paper, which will appear in 2015.

6.1.3. Datasource

We extended the HOP.JS language with an embedded language, inspired by PLINQ and ORC, called DATASOURCE. It allows programmers request multiple data sources with queries written in a unique language. We used a plinq-like language to express queries and an orc-like language to orchestrate them. Our query language and the orchastration languages can be used simultaneously or separately. We implemented bindings between DATASOURCE and some representative types of data sets such as SPARQL endpoints, relationnal databases, WEB services, and WEB pages. We are extending HOP.JS by supporting EcmaScript 6 array comprehensions in order to write a unique query over multiple data sources in a unified formalism. The query is then compiled into database specific queries. We linked all the bindings made for HOP with HOP.JS. We implemented another binding for a document oriented data base, MONGODB.

6.2. Distributed programming

Participant: Bernard Serpette [correspondant].

6.2.1. Logical behavioural semantics of Esterel

We have formalised, with the Coq system, the logical behavioural semantics of Esterel as described in Gérard Berry's book. In order to define the properties of reactivity and determinism, we have defined a new semantics using contexts with a proven correspondence between the two semantics.

The specification and the proofs of the correspondence take 3500 lines of Coq.

6.2.2. Abstract distributed machine

We have experimented an abstract machine composed of distributed nodes. Each node has exactly two named links to other nodes and an instruction able to modify one link of a reachable node. This instruction is executed when a token is received, once the instruction is achieved the token is transmitted to another reachable node.

This abstract machine is turing complete. The λ-calculus and the π-calculus can be compiled to the instruction set of this machine.

The execution of one individual node may involve paths of arbitrary length, for example, when compiling the λ-calculus or the π-calculus, the path length for accessing a variable is proportional to it's de Bruijn index and therefore is not bounded. Given a machine with instructions of unbounded paths, we can build an equivalent machine where all the paths are bounded by two: a node is only able to access it's own links and the links of its neighbour. Moreover, this transformation uses only 6 different instructions.

6.3. Security and Privacy

Participants: Ilaria Castellani, José Fragoso Santos, Nataliia Bielova, Tamara Rezk [correspondant].

6.3.1. Security of Dynamically Evolving Systems of Communicating Processes

We have started to address security issues in the context of dynamically evolving systems of communicating processes, which are able to adapt themselves in reaction to particular events (for instance, security attacks or changes in security policies). We present initial results on a simple model of processes communicating via structured interactions (sessions), in which self-adaptation and security concerns are jointly addressed. In this model, security violations occur when processes attempt to read or write messages of inappropriate security level within a structured interaction. Such violations trigger adaptation mechanisms that prevent the violations to occur and/or to propagate their effect in the choreography. Our model is equipped with local and global mechanisms for reacting to security violations; type soundness results ensure that the global protocols are still correctly executed while the system adapts itself to preserve its security.

6.3.2. Browser Randomisation against Fingerprinting: a Quantitative Information Flow Approach

Web tracking companies use device fingerprinting to distinguish the users of the websites by checking the numerous properties of their machines and web browsers. One way to protect the users' privacy is to make them switch between different machine and browser configurations. We propose a formalisation of this privacy enforcement mechanism.

We use information-theoretic channels to model the knowledge of the tracker and the fingerprinting program, and show how to synthesise a randomisation mechanism that defines the distribution of configurations for each user. This mechanism provides a strong guarantee of *privacy* (the probability of identifying the user is bounded by a given threshold) while maximising *usability* (the user switches to other configurations rarely). To find an optimal solution, we express the enforcement problem of randomisation by a linear program. We investigate and compare several approaches to randomisation and find that more efficient privacy enforcement would often provide lower usability. Finally, we relax the requirement of knowing the fingerprinting program in advance, by proposing a randomisation mechanism that guarantees privacy for an arbitrary program.

This work has been published and presented at the Nordic Conference on Secure IT Systems (NordSec 2014) [12]. The extended version of the paper has been published as a technical report [20].

6.3.3. Crying Wolf? On the Price Discrimination of Online Airline Tickets

Price discrimination refers to the practice of dynamically varying the prices of goods based on a customer's purchasing power and willingness to pay. Motivated by several anecdotal accounts, we report on a three week experiment, conducted in search of price discrimination in airline tickets. Despite presenting the companies with multiple opportunities for discriminating us, and contrary to our expectations, we did not find any evidence for systematic price discrimination. At the same time, we witnessed the highly volatile prices of certain airlines which make it hard to establish cause and effect. Finally, we provided alternative explanations for the observed price differences.

This work has been published and presented at the Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2014) [19].

6.3.4. Stateful Declassification Policies for Event-Driven Programs

We propose a novel mechanism for enforcing information flow policies with support for declassification on event-driven programs. Declassification policies consist of two functions. First, a projection function specifies for each confidential event what information in the event can be declassified directly. This generalizes the traditional security labelling of inputs. Second, a stateful release function specifies the aggregate information about all confidential events seen so far that can be declassified. We provide evidence that such declassification policies are useful in the context of JavaScript web applications. An enforcement mechanism for our policies is presented and its soundness and precision is proven. Finally, we give evidence of practicality by implementing and evaluating the mechanism in a browser. This work has been published at Computer Security Foundations (CSF'14) [18].

6.3.5. An Information Flow Monitor for a Core of DOM

We propose and prove sound a novel, purely dynamic, flow sensitive monitor for securing information flow in an imperative language extended with DOM-like tree operations, that we call Core DOM. In Core DOM, as in the DOM API, tree nodes are treated as first-class values. We take advantage of this feature in order to implement an information flow control mechanism that is finer-grained than previous approaches in the literature. Furthermore, we extend Core DOM with additional constructs to model the behavior of live collections in the DOM Core Level 1 API. We show that this kind of construct effectively augments the observational power of an attacker and we modify the proposed monitor so as to tackle newly introduced forms of information leaks. This work has been published at the 9th International Symposium on Trustworthy Global Computing (TGC) [11].

6.3.6. An Information Flow Monitor-Inlining Compiler for Securing a Core of JavaScript

Web application designers and users alike are interested in isolation properties for trusted JavaScript code in order to prevent confidential resources from being leaked to untrusted parties. Noninterference provides the mathematical foundation for reasoning precisely about the information flows that take place during the execution of a program. Due to the dynamicity of the language, research on mechanisms for enforcing noninterference in JavaScript has mostly focused on dynamic approaches. We present the first information flow monitor inlining compiler for a realistic core of JavaScript. We prove that the proposed compiler enforces termination-insensitive noninterference and we provide an implementation that illustrates its applicability.

This work has been published at the 29th IFIP International Information Security and Privacy Conference (IFIP SEC) [14].

6.3.7. From Static to Hybrid Typing Secure Information Flow in a Core of JavaScript

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.

PHOENIX Project-Team

6. New Results

6.1. Highlights of the Year

- A best paper award was obtained at ASSETS 2014 (The 16th International ACM SIGACCESS Conference on Computers and Accessibility), by the 5 authors of the paper "Tablet-Based Activity Schedule for Children with Autism in Mainstream Environment" .

BEST PAPERS AWARDS :
[26] **ASSETS 2014 - The 16th International ACM SIGACCESS Conference on Computers and Accessibility**. C. FAGE, L. POMMEREAU, C. CONSEL, E. BALLAND, H. SAUZÉON.

6.2. Technological Support for Self-Regulation of Children with Autism

Children with Autism Spectrum Disorders (ASD) have difficulties to self-regulate emotions, impeding their inclusion in a range of mainstreamed environments. Self-regulating emotions has been shown to require recognizing emotions and invoking specific coping strategies.

In the context of the School+ research project, we have developed an application dedicated to self-regulating emotions in children with ASD. Ten children with ASD have experimentally tested this tablet-based application over a period of three months in a mainstreamed school. A collaborative learning approach, involving parents, teachers and a school aid, was used 1) to train students to operate the tablet and our application autonomously, and 2) to facilitate the adoption of our intervention tool.

This study shows that our application was successful in enabling students with ASD to self-regulate their emotions in a school environment. Our application helped children with autism to recognize and name their emotions, and to regulate them using idiosyncratic, parent-child, coping strategies, supported by multimedia contents.

This work is in the context of the School+ national research project funded by the French Ministry of National Education.

A best paper award was obtained at ASSETS 2014 (The 16th International ACM SIGACCESS Conference on Computers and Accessibility) for this work in October 2014, by the 5 authors of the paper "Tablet-Based Activity Schedule for Children with Autism in Mainstream Environment" [26]: Charles Fage, Léonard Pommereau, Charles Consel, Emilie Balland, and Hélène Sauzéon.

6.3. A Low-Cost approach to the Verification of Daily Activities of Elders

Activities of Daily Living (ADL) are abilities defining the functional status of an individual. Verifying what ADLs are performed by an elder is a decisive factor to determine what kinds and what levels of assistance are needed for an individual and whether aging in place is desirable. The importance of this issue has led a number of researchers to develop a range of Ubicomp approaches that can monitor activities.

In this study, we take these prior results one step further and apply them to the needs of caregiver professionals to monitor elders at their home. Specifically, our approach relies on the following key observation: as people age their daily activities are increasingly organized according to a routine to optimize their daily functioning. As a result, their activities do not need to be recognized but should rather be verified. Deviations are a warning sign of degradation.

We have developed an approach to activity verification. This approach relies on a technological infrastructure that is simple, low-cost and non-intrusive. This infrastructure was deployed in four homes of elders of 83 years of age on average. The same set of sensors was used in the four homes and was placed at strategic locations with respect to their routines to verify the target activities. The analysis of the data collected during five weekdays show that they follow very strict routines that can easily be associated with their main activities.

This work is in the context of the DomAssist project, funded by the following partners: UDCCAS, CG33, CRA, CNSA, Chambre des métiers. A report of the work has been published at the 16th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2014) [25].

6.4. Using virtual reality for studying everyday-like memory and its cognitive correlates

This work consisted in a pilot-study with a comparison approach between aging and traumatic brain injury (TBI) to investigate everyday object memory patterns using a virtual HOMES test.

- Methods: Sixteen young controls, 15 older adults and 15 TBI patients underwent the HOMES test and traditional tests.
- Results: Older adults and TBI patients exhibited similar HOMES performances: poor recall, a greater recognition benefit, high false recognitions, but intact clustering and proactive interference effects. The age-related differences for HOMES measures were mainly mediated by executive functioning, while the HOMES performances in the TBI group were correlated with memory measures.
- Conclusion: The differential cognitive mediating effects for a similar everyday-like memory pattern have been discussed by highlighting the need for more cautious interpretations of cognitive mechanisms behind similar behavioral patterns in different populations especially in clinical and rehabilitation settings.
- Implication for Rehabilitation:
  - Virtual reality might provide ecological scenarios to assess the multiple processes of everyday memory in elderly people as well as in TBI patients.
  - A similar pattern of Everyday-like memory failures might result from different cognitive origins among different neuropsychological patients.
  - The assessment of specific cognitive origins of Everyday-like memory impairments deserves consideration for drawing up relevant rehabilitative programs that match the specific cognitive needs of patients for performing everyday memory tasks.

This work has been published in the journal "Disability and Rehabilitation: Assistive Technology" in November 2014 [14].

RMOD Project-Team

6. New Results

6.1. Highlights of the Year

- Pharo 3.0 has been released in April 2014.
- Moose 5.0 has been released in December 2014.
- The book Deep into Pharo has been released publicly http://www.deepintopharo.com.
- RMOD entered in a sponsoring agreement with LAM Research, Inc.

6.2. Tools for understanding applications

Remodularization Analysis Using Semantic Clustering. We report an experience on using and adapting Semantic Clustering to evaluate software remodularizations. Semantic Clustering is an approach that relies on information retrieval and clustering techniques to extract sets of similar classes in a system, according to their vocabularies. We adapted Semantic Clustering to support remodularization analysis. We evaluate our adaptation using six real-world remodularizations of four software systems. We report that Semantic Clustering and conceptual metrics can be used to express and explain the intention of the architects when performing common modularization operators, such as module decomposition. [37]

Towards a new package dependency model. Smalltalk originally did not have a package manager. Each Smalltalk implementation defined its own with more or less functionalities. Since 2010, Monticello/Metacello[Hen09] one package manager is available for open-source Smalltalks. It allows one to load source code packages with their dependencies. This package manager does not have all features we can find in well-known package managers like those used for the Linux operating system. We identify the missing features and propose a solution to reach a full-featured package manager. A part of this solution is to represent packages and dependencies as first-class objects, leading to the definition of a new dependency model. [32]

A Domain Specific Aspect Language for IDE Events. Integrated development environments (IDEs) have become the primary way to develop software. Besides just using the built-in features, it becomes more and more important to be able to extend the IDE with new features and extensions. Plugin architectures exist, but they show weaknesses related to unanticipated extensions and event handling. We argue that a more general solution for extending IDEs is needed. We present and discuss a solution, motivated by a set of concrete examples: a domain specific aspect language for IDE events. In it, join points are events of interest that may trigger the advice in which the behavior of the IDE extension is called. We show how this allows for the development of IDE plugins and demonstrate the advantages over traditional publish/subscribe systems. [21]

AspectMaps: Extending Moose to visualize AOP software. When using aspect-oriented programming the application implicitly invokes the functionality contained in the aspects. Consequently program comprehension of such a software is more intricate. To alleviate this difficulty we developed the AspectMaps visualization and tool. AspectMaps extends the Moose program comprehension and reverse engineering platform with support for aspects, and is implemented using facilities provided by Moose. We present the AspectMaps tool, and show how it can be used by performing an exploration of a fairly large aspect-oriented application. We then show how we extended the FAMIX meta-model family that underpins Moose to also provide support for aspects. This extension is called ASPIX, and thanks to this enhancement Moose can now also treat aspect-oriented software. Finally, we report on our experiences using some of the tools in Moose; Mondrian to implement the visualization, and Glamour to build the user interface. We discuss how we were able to implement a sizable visualization tool using them and how we were able to deal with some of their limitations. [20]

6.3. Software Quality: Taming Software Evolution

APIEvolutionMiner: Keeping API Evolution under Control. During software evolution, source code is constantly refactored. In real-world migrations, many methods in the newer version are not present in the old version (e.g., 60% of the methods in Eclipse 2.0 were not in version 1.0). This requires changes to be consistently applied to reflect the new API and avoid further maintenance problems. We propose a tool to extract rules by monitoring API changes applied in source code during system evolution. In this process, changes are mined at revision level in code history. Our tool focuses on mining invocation changes to keep track of how they are evolving. We also provide three case studies in order to evaluate the tool. [34]

Towards an Automation of the Mutation Analysis Dedicated to Model Transformation. A major benefit of Model Driven Engineering (MDE) relies on the automatic generation of artefacts from high-level models through intermediary levels using model transformations. In such a process, the input must be well-designed and the model transformations should be trustworthy. Due to the specificities of models and transformations, classical software test techniques have to be adapted. Among these techniques, mutation analysis has been ported and a set of mutation operators has been defined. However, mutation analysis currently requires a considerable manual work and suffers from the test data set improvement activity. This activity is seen by testers as a difficult and time-consuming job, and reduces the benefits of the mutation analysis. This work addresses the test data set improvement activity. Model transformation traceability in conjunction with a model of mutation operators, and a dedicated algorithm allow to automatically or semi-automatically produce test models that detect new faults. The proposed approach is validated and illustrated in a case study written in Kermeta. [17]

Predicting software defects with causality tests. We propose a defect prediction approach centered on more robust evidences towards causality between source code metrics (as predictors) and the occurrence of defects. More specifically, we rely on the Granger causality test to evaluate whether past variations in source code metrics values can be used to forecast changes in time series of defects. Our approach triggers alarms when changes made to the source code of a target system have a high chance of producing defects. We evaluated our approach in several life stages of four Java-based systems. We reached an average precision greater than 50% in three out of the four systems we evaluated. Moreover, by comparing our approach with baselines that are not based on causality tests, it achieved a better precision. [19]

6.4. Software Quality: History and Changes

Tracking dependencies between code changes: An incremental approach. Merging a change often leads to the question of knowing what are the dependencies to other changes that should be merged too to obtain a working system. This question also arises with code history trackers – Code history trackers are tools that react to what the developer do by creating first-class objects that represent the change made to the system. We evaluate the capacity of different code history trackers to represent, also as first-class objects, the dependencies between those changes. We also present a representation for dependencies that works with the event model of Epicea, a fine-grained and incremental code history tracker. [32]

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 aheuristic-based approach for architecture conformance can be used to rapidly raise architectural warnings, without deeply involving experts in the process. [22]

6.5. Reconciling Dynamic Languages and Isolation

Delegation Proxies: The Power of Propagation. Scoping behavioral variations to dynamic extents is useful to support non-functional requirements that otherwise result in cross-cutting code. Unfortunately, such variations are difficult to achieve with traditional reflection or aspects. We show that with a modification of dynamic proxies, called delegation proxies, it becomes possible to reflectively implement variations that propagate to all objects accessed in the dynamic extent of a message send. We demonstrate our approach with examples of variations scoped to dynamic extents that help simplify code related to safety, reliability, and monitoring. [38]

Reifying the Reflectogram. Reflective facilities in OO languages are used both for implementing language extensions (such as AOP frameworks) and for supporting new programming tools and methodologies (such as object-centric debugging and message-based profiling). Yet controlling the run-time behavior of these reflective facilities introduces several challenges, such as computational overhead, the possibility of meta-recursion and an unclean separation of concerns between base and meta-level. We present five dimensions of meta-level control from related literature that try to remedy these problems. These dimensions are namely: temporal and spatial control, placement control, level control and identity control. We argue that the reification of the descriptive notion of the reflectogram, can unify the control of meta-level execution in all these five dimensions. We present a model for the reification of the reflectogram and validate our approach through a prototype implementation in the Pharo programming environment. Finally we detail a case study on run-time tracing illustrating our approach. [35]

Bootstrapping Reflective Systems: The Case of Pharo. Bootstrapping is a technique commonly known by its usage in language definition by the introduction of a compiler written in the same language it compiles. This process is important to understand and modify the definition of a given language using the same language, taking benefit of the abstractions and expression power it provides. A bootstrap, then, supports the evolution of a language. However, the infrastructure of reflective systems like Smalltalk includes, in addition to a compiler, an environment with several self-references. A reflective system bootstrap should consider all its infrastructural components. We propose a definition of bootstrap for object-oriented reflective systems, we describe the architecture and components it should contain and we analyze the challenges it has to overcome. Finally, we present a reference bootstrap process for a reflective system and Hazelnut, its implementation for bootstrapping the Pharo Smalltalk-inspired system. [26]

6.6. Dynamic Languages: Virtual Machines

Benzo: Reflective Glue for Low-level Programming. The goal of high-level low-level programming is to bring the abstraction capabilities of high-level languages to the system programming domain, such as virtual machines (VMs) and language runtimes. However, existing solutions are bound to compilation time and expose limited possibilities to be changed at runtime and from language-side. They do not fit well with fully reflective languages and environments. We propose Benzo1, a lightweight framework for high-level low-level programming that allows developers to generate and execute at runtime low-level code. It promotes the implementation, and dynamic modification, of system components with high-level language tools outperforming existing dynamic solutions. Since Benzo is a general framework we choose three applications that cover an important range of the spectrum of system programming for validating the infrastructure: a Foreign Function Interface (FFI), primitives instrumentation and a just-in-time bytecode compiler (JIT). With Benzo we show that these typical VM-level components are feasible as reflective language-side implementations. Due to its unique combination of high-level reflection and low-level programming, Benzo shows better performance for these three applications than the comparable high-level implementations. [30]

A bytecode set for adaptive optimizations. The Cog virtual machine features a bytecode interpreter and a baseline Just-in-time compiler. To reach the performance level of industrial quality virtual machines such as Java HotSpot, it needs to employ an adaptive inlining com-piler, a tool that on the fly aggressively optimizes frequently executed portions of code. We decided to implement such a tool as a bytecode to bytecode optimizer, implemented above the virtual machine, where it can be written and developed in Smalltalk. The optimizer we plan needs to extend the operations encoded in the bytecode set and its quality heavily depends on the bytecode

set quality. The current bytecode set understood by the virtual machine is old and lacks any room to add new operations. We decided to implement a new bytecode set, which includes additional bytecodes that allow the Just-in-time compiler to generate less generic, and hence simpler and faster code sequences for frequently executed primitives. The new bytecode set includes traps for validating speculative inlining de-cisions and is extensible without compromising optimization opportunities. In addition, we took advantage of this work to solve limitations of the current bytecode set such as the maximum number of instance variable per class, or number of literals per method. We plan to have it in production in the Cog virtual machine and its Pharo, Squeak and Newspeak clients in the coming year. [43]

6.7. Traits

Trait-oriented Programming in Java 8 Java 8 was released recently. Along with lambda expressions, a new language construct is introduced: default methods in interfaces. The intent of this feature is to allow interfaces to be extended over time preserving backward compatibility. We show a possible, different use of interfaces with default methods: we introduce a trait-oriented programming style based on an interface-as-trait idea, with the aim of improving code modularity. Starting from the most common operators on traits, we introduce some programming patterns mimicking such operators and discuss this approach. [29]

6.8. Tailoring Applications

In the context of the PhD of G. Polito, we developed Tornado, a way to generate specialized and minimal runtime. Using a run-fail-grow approach, which tries to execute an expression in an empty world, and on failure copies the missing program elements from a mother environment to the currently empty world, we could grew 11k full reflective application adding two numbers or 18k for the 100 factorial expression. We also usd this approach to generate specialized webserver in around 500kb. These results show that we can generate hyperspecialized kernels.

TACOMA Team

6. New Results

6.1. Self-describing objects and tangible data structures

Participants: Nebil Ben Mabrouk, Paul Couderc [contact], Arnab Sinha.

A development in the line of the coupled objects principles 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, coupled 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 analyzing 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 "ÃÂÃÂself-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. On this topic, a generic graph structure applicable to RFID systems for supporting self-describing objects is proposed in Arnab Sinha's thesis document [1], and was published in [4].

6.2. Pervasive support for RFIDs

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

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 mis-read (by using data redundancy) and to help the reading protocol by integrating hints about the tag set collection being read.

Our objective here is to study 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 can enable faster are more efficient inventory protocols.

An important milestone was completed in 2014, with the implementation of an experiment test bed in order to support the experiment campaign. This task involved a significant development and engineering effort. This testbed is currently deployed at the IETR (http://www.ietr.fr) building, and features a multi-axis mobile RFID antenna system driven by a software platform.

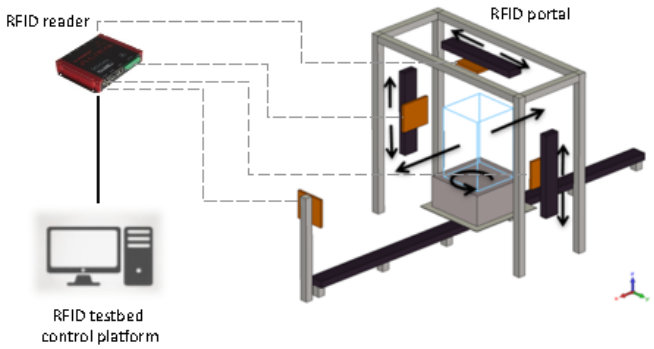

Figure 1. RFID testbed

This system allows both interactive testing as well as long running experiments of RFID reading protocols. The software platform was designed to allow fine control over all dynamic aspects influencing RFID readings: movements for target and antenna, RFID reader configuration, and smart antenna configuration (diversity and power control). Given this flexibility, this platform should be able to reproduce most of the situations found in real applications. In particular, it can be used to design custom reading set up optimized for various RFID portal applications [3].

6.3. Context-aware dynamic Smart Home Platform

Participants: Andrey Boytsov, Aurélien Richez, Yoann Maurel, Frédéric Weis [contact].

Tacoma group is focussed on the conception and implementation of innovative services for the Smart Home. The range of considered services is broad : from "optimizing the energy consumption" to "helping users to find their way in a building". To provide such services, 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.

Building efficient and accurate context awareness was and is still a great challenge but 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 decided to use the Belief Function Theory (BFT) [8] [9] as it allows to express uncertainty

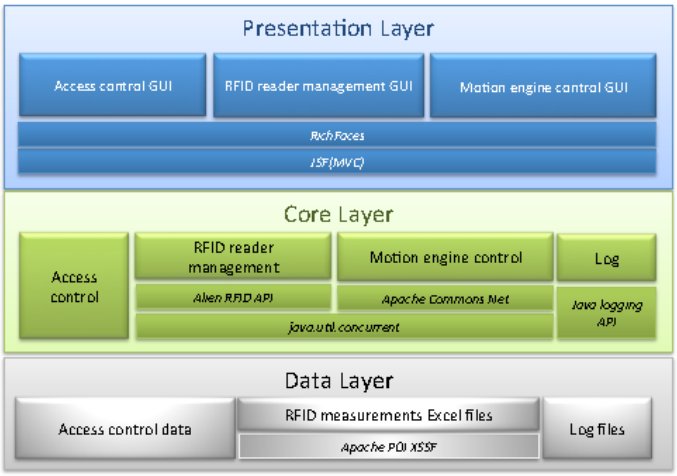

Figure 2. Software architecture of the RFID testbed

and imprecision. Although these results are very promising, great challenges still lied in (i) the support of the dynamic reconfiguration to face evolving hardware or software conditions and (ii) the deployment and the configuration of the layered architecture and sensors to allow the use of our approach in unknown environments.

One of our goals is to build a pervasive platform with constrained performance and cost [7]. The cost is particularly critical for sensors and actuators: we choose to limit our scope to inexpensive and non-invasive sensors *i.e.* no video camera. This past months, Tacoma has been working on the conception and implementation of a Smart Home Platform based on earlier prototypes inherited from ACES team. The prototypes were implemented as an hard-to-maintain monolithic code. The code also suffered from a lot of redundancy. More importantly the platform hardly supported dynamism and provided no support for reconfiguration and adaptation at runtime. With this in mind, during the re-writing of the platform the emphasis has been placed on the following aspects:

- supporting the dynamic discovery of heterogeneous sensors;
- enabling the dynamic deployment of applications at runtime ;
- enabling context-awareness by providing contextual information to these applications;
- enhancing the separation of concerns and code-reuse.

Our goal is to design and build a platform that is:

- **evolutive**: the Home environment is ever-changing and thus it is important to allow users to add new sensors or new services dynamically at runtime. It is also mandatory to recalibrate the sensors to face the change in the Home. This is mainly why we based our platform on OSGi;
- **maintainable and administrable**: we raised the maintainability by using a modular approach using C-modules or iPOJO components; the platform is itself modular to achieve a good separation of concerns (e.g., communication, module loading, discovery). We also built in-production monitoring interfaces that provides information on the belief functions that are used, the fusion process and the sensors values;
- **easy to configure**: 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 mass functions, required by the theory, that have to be provided by each sensor.

6.3.1. Towards dynamism using OSGI

The development of our initial platform in C proved to be costly and hard to maintain. The dynamism is hard to achieved with a low-level language and requires an heavy development process. This led the team to investigate the use of OSGi as a based for our execution platform. OSGi is the specification of an execution framework developed on top of Java. It relies on the Java's dynamic features (dynamic and on demand class loading through class loaders) to provide a coarse-grained level of modularity. This choice was reinforced by our collaboration with the Adele team (LIG Laboratory in Grenoble). This team is using OSGi as a core for building Smart Home applications. Using OSGi would ease collaboration and code sharing.

One main concerns regarding the use of Java was the limited performances of the targeted hardware (raspberry pi). The Belief Function Theory (BFT) requires heavy computations and the embedded CPU could have been the bottleneck. Moreover, the JVM supported by the raspberry pi is limited compared to standard JVM. As a preliminary study, we choose to implement the core of the BFT library in Java and to compare the performances with the C implementation. Unexpectedly the Java implementation performed better than the C implementation in most of the case. This can be explained by three factors. First, the BFT theory is tedious to implement in low-level language. The C-implementation could probably be optimized but this will lower the readability of the source code and impact the maintainability. Second and conversely, using Java raised the code readability and allowed us to performed some optimization. Third, the JIT (Just In Time) compiler provided by the VM have been improved these past years and the optimization performed by the VM are sufficient to bring on par performances with the C implementation. As, the performances of the C platform were largely sufficient, this preliminary phase validated our decision to switch to OSGi.

6.3.2. Automated configuration of sensors

A previous defended in the group in december 2013 has shown promising results applying the BFT theory to the Smart Home Domain. It is currently possible to collect sensor values and extract belief functions from them. The platform can then extract a context from the belief functions and offer services to the user depending on what is happening. For instance, the user may be notified of an open window when he leaves the house.

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. 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 model 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.

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. The group is currently studying the possible use of clustering and classifications algorithm in order to ease the calibration of sensors. Yoann Maurel and Frédéric Weis supervised a project with a group of ENS student on this subject. The goal is to generate our belief model from a training set of sensor data. We mainly focus on two algorithms: k-nearest neighbors (KNN) and overlapping k-mean (OKM). A first experimentation with KNN and motion sensors showed that this algorithm is promising. We used a training data set to compute the presence belief model. We acquired a first 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.

6.4. 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. During the first year, we focused on identifying the needs of the industrial partners and public authorities that fund the chair.

This activity is centered on the concept of metamorphic housing (see section 4.2). During this year, we introduced a solution of metamorphic housing addressing the goals of saving space and energy in an apartment building, while preserving residents' comfort: the on-demand room. 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, as illustrated in Figure 3 . 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.

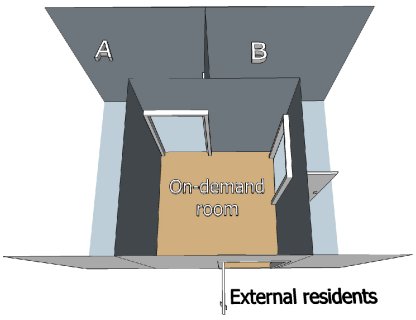

Figure 3. Floor plan for a metamorphic house

The underlying research problems are numerous. Dynamically "plugging" the room into a different apartment requires replacing the owner of the room's equipment, including appliances, heating, ventilation and air conditioning systems (HVAC), sensors, etc. The rights to control them and receive information from them must be dynamically reallocated. This must be done in a transparent fashion, so that off-the-shelf devices and appliances can be used.

In some cases, devices require dynamic reprogramming, like HVAC systems, because they must adapt to occupants' preferences and settings (e.g., ambient temperature set point).

Another research problem is the automatic learning of a schedule for the on-demand room. Regularities in users' requests for the room, duration of their occupation and privacy level can be discovered and learned. In this way, users do not have to manually book the room and usage conflicts can be prevented. We started investigating these research problems with an interdisciplinary approach and in collaboration with companies and public authorities [6]. We also started working on a prototype of the on-demand room solution, which will be presented as an immersive interactive virtual-reality application, leveraging the Immersia platform http://www.irisa.fr/immersia/.

COATI Project-Team

6. New Results

6.1. Network Design and Management

Participants: Jean-Claude Bermond, David Coudert, Frédéric Giroire, Frédéric Havet, Alvinice Kodjo, Aurélien Lancin, Bi Li, Fatima Zahra Moataz, Christelle Molle-Caillouet, Joanna Moulierac, Nicolas Nisse, Stéphane Pérennes, Truong Khoa Phan.

More information on several results presented in this section may be found in the PhD thesis of A. Kodjo [13], B. Li [15] and T. K. Phan [18].

6.1.1. Optimization in backbone networks

6.1.1.1. 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 [52], we extend these notions to characterize new cases in which these optimization problems can be solved in polynomial time or are fixed parameter tractable. We also investigate on 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.

6.1.1.2. Dynamic Routing and Spectrum Assignment in Optical Networks

Elastic Optical Networks (EONs) promises a better utilization of the spectrum in optical networks. In fact, as the optical transmission spectrum is carved into fixed-length bands in the traditional WDM networks, small bit rates are over-provisioned and very high bit rates do not fit. EONs are moving away from this fixed-grid and allow the spectrum to be divided flexibly: each request is allocated exactly the resources it needs. In [34], we present two exact algorithms to route and allocate spectrum to a new request in an EON using only Non-Disruptive Defragmentation (Push-Pull). In the first algorithm, we find the shortest routing path for the new request (i.e., the shortest path from source to destination where contiguous spectrum to satisfy the request can be freed) and then find the position that gives the overall minimum delay on that path. In the second algorithm, we find at the same time a routing path and a position in the spectrum, that minimize the delay of insertion (over all other paths and positions). Both algorithms are polynomial in the size of the network, its bandwidth and the number of provisioned requests.

6.1.2. Microwave Backhaul networks

6.1.2.1. Chance-Constrained Optimization of Reliable Backhaul networks

In [25], we extend our former investigation on conceiving reliable fixed point-to-point wireless networks under outage probability constraints. We consider the problem of determining the minimum cost bandwidth assignment of a network, while guaranteeing a reliability level of the solution. If the optimal bandwidth assignment and routing of traffic demands are accomplished, the reliability criterion requires that network flows remain feasible with high probability, regarding that the performance of microwave links is prone to

variations due to external factors, e.g., weather. We introduce a *chance-constrained programming* approach to tackle this problem and we present reformulations to standard integer linear programming models, including a budget-constrained formulation. To improve the solving performance, we propose new valid inequalities and a primal heuristic. Computational results present a performance analysis of the valid inequalities and the heuristic. Further, the outperformance of the novel model compared to more traditional approaches is documented.

6.1.2.2. Robust optimization in multi-operators microwave backhaul networks

In [41], we consider the problem of sharing the infrastructure of a backhaul network for routing. We investigate on the revenue maximization problem for the physical network operator (PNO) when subject to stochastic traffic requirements of multiple virtual network operators (VNO) and prescribed service level agreements (SLA). We use robust optimization to study the tradeoff between revenue maximization and the allowed level of uncertainty in the traffic demands. This mixed integer linear programming model takes into account end-to-end traffic delays as example of quality-of-service requirement in a SLA. To show the effectiveness of our model, we present a study on the price of robustness, i.e. the additional price to pay in order to obtain a feasible solution for the robust scheme, on realistic scenarios.

6.1.3. Energy efficiency

6.1.3.1. Robust Optimization for Energy-aware Routing with Redundancy Elimination

Many studies in literature have shown that energy-aware routing (EAR) can significantly reduce energy consumption for backbone networks. Also, as an arising concern in networking research area, the protocol-independent traffic redundancy elimination (RE) technique helps to reduce (a.k.a compress) traffic load on backbone network. In [35], [50], we present an extended model of the classical multi-commodity flow problem with compressible flows. Our model is robust with fluctuation of traffic demand and compression rate. In details, we allow any set of a predefined size of traffic flows to deviate simultaneously from their nominal volumes or compression rates. As an applicable example, we use this model to combine redundancy elimination and energy-aware routing to increase energy efficiency for a backbone network. Using this extra knowledge on the dynamics of the traffic pattern, we are able to significantly increase energy efficiency for the network. We formally define the problem and model it as a Mixed Integer Linear Program (MILP). We then propose an efficient heuristic algorithm that is suitable for large networks. Simulation results with real traffic traces on Abilene, Geant and Germany50 networks show that our approach allows for 16-28% extra energy savings with respect to the classical EAR model.

6.1.3.2. Optimizing IGP Link Weights for Energy-efficiency in a Changing World

Recently, due to the increasing power consumption and worldwide gas emissions in ICT (Information and Communication Technology), energy efficient ways to design and operate backbone networks are becoming a new concern for network operators. Since these networks are usually overprovisioned and since traffic load has a small influence on power consumption of network equipments, the most common approach to save energy is to put unused line cards that drive links between neighbouring routers into sleep mode. To guarantee QoS, 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. We propose in [62] 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. We formally define the problem and model it as Mixed Integer Linear Program (MILP). We then propose an efficient heuristic algorithm that is suitable for large networks. Simulation results with real traffic traces on three different networks show that our approach achieves high energy saving while keeping the networks in stable state (less changes in network configuration).

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

A routing R of a connected graph G is a collection that contains simple paths connecting every ordered pair of vertices in G. The edge-forwarding index with respect to R (or simply the forwarding index with respect to R) $\pi(G, R)$ of G is the maximum number of paths in R passing through any edge of G. The forwarding index $\pi(G)$ of G is the minimum $\pi(G, R)$ over all routings R's of G. This parameter has been studied for different graph classes. Motivated by energy efficiency, we look in [57], for different numbers of edges, at the best spanning graphs of a square grid, namely those with a low forwarding index.

6.1.4. Software-Defined Networks

6.1.4.1. Rule Placement in Software-Defined Networks for Energy-aware Routing

Software-defined Networks (SDN), in particular OpenFlow, is a new networking paradigm enabling innovation through network programmability. Over past few years, many applications have been built using SDN such as server load balancing, virtual-machine migration, traffic engineering and access control. We focus on using SDN for energy-aware routing (EAR). SDN can collect traffic matrix and then computes routing solutions satisfying QoS while being minimal in energy consumption (with minimal number of active links). However, prior works on EAR have assumed that the table of OpenFlow switch can hold an infinite number of rules. In practice, this assumption does not hold since the flow table is implemented with Ternary Content Addressable Memory (TCAM) which is expensive and power-hungry. In [39], [56], we propose an optimization method to minimize energy consumption for a backbone network while respecting capacity constraints on links and rule space constraints on routers. In details, we present an exact formulation using Integer Linear Program (ILP) and introduce efficient greedy heuristic algorithm. Based on simulations, we show that using this smart rule space allocation, it is possible to save almost as much power consumption as the classical EAR approach.

6.1.4.2. Compressing Two-dimensional Routing Tables with Order

A communication in a network is a pair of nodes (s, t). The node s is called the source source and t the destination. A communication set is a set of distinct communications, i.e. two communications might have the same source or the same destination, but they cannot have both same source and same destination. A routing of a communication (s, t) is a path in the network from s to t. A routing of a communication set is a union of routings of its communications. At each node, there is a set X of communications whose routing path goes through this node. The node needs to be able to find for each communication (s, t) in X, the port that the routing path of (s, t) uses to leave it. An easy way of doing it is to store the list of all triples (s, t, k), where $(s, t) \in X$ and k is the port used by the (s, t)-path to leave the node. Such triples are called communication triples. However, such a list might be very large. Motivated by routing in telecommunication network using Software Defined Network Technologies, we consider in [55] the problem of compacting this list using aggregation rules. Indeed, SDN routers use specific memory which is expensive and of small capacity. Hence, in addition, we can use some additional triples, called *-triples. As an example, a t-destination triple $(*, t, p)$, means that every communication with destination t leaves on port p. We carry out in this work a study of the problem complexity, providing results of NP-completeness, of Fixed-Parameter Tractability and approximation algorithms.

6.1.5. Data gathering in radio networks

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 we consider a binary interference model where two calls interfere in a step, if the sender of one call is at distance at most d_I from the receiver of the other call. 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 [23], 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 $d_I \in \{0, 1, 2\}$. In particular, we present 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. 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.

6.2. Graph Algorithms

Participants: Julio Araújo, Jean-Claude Bermond, David Coudert, Guillaume Ducoffe, Frédéric Giroire, Aurélien Lancin, Bi Li, Fatima Zahra Moataz, Christelle Molle-Caillouet, 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. More information on several results presented in this section may be found in PhD thesis of B. Li [15] and A. Lancin [14], and in the Habilitation thesis of N. Nisse [17].

6.2.1. Complexity and Computation of Graph Parameters

We use graph theory to model various network problems. In general 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.

6.2.1.1. Hyperbolicity

The Gromov hyperbolicity is an important parameter for analyzing complex networks since it expresses how the metric structure of a network looks like a tree. In other words, it provides bounds on the stretch resulting from the embedding of a network topology into a weighted tree. It is therefore used to provide bounds on the expected stretch of greedy-routing algorithms in Internet-like graphs. However, the best known algorithm for computing this parameter has time complexity in $O(n^{3.69})$, which is prohibitive for large-scale graphs.

In [47], we investigate some relations between the hyperbolicity of a graph and the hyperbolicity of its *atoms*, that are the subgraphs resulting from the decomposition of the graph according its clique minimal separators. More precisely, we prove that the maximum hyperbolicity taken over all the atoms is at least the hyperbolicity of *Gminus one*. We also give an algorithm to slightly modify the atoms, which is at no extra cost than computing the atoms themselves, and so that the maximum hyperbolicity taken over all the resulting graphs is *exactly* the hyperbolicity of G. An experimental evaluation of our methodology is provided for large collaboration networks. Finally, we deduce from our theoretical results the first *linear-time* algorithm to compute the hyperbolicity of an outerplanar graph.

The shortest-path metric d of a connected graph G is 1/2-hyperbolic if, and only if, it satisfies $\mathrm{d}(u, v) + \mathrm{d}(x, y) \leq \max\{\mathrm{d}(u, x) + \mathrm{d}(v, y), \mathrm{d}(u, y) + \mathrm{d}(v, x)\} + 1$, for every 4-tuple u, x, v, y of G. We show in [26], [48] that the problem of deciding whether an unweighted graph is 1/2-hyperbolic is subcubic equivalent to the problem of determining whether there is a chordless cycle of length 4 in a graph. An improved algorithm is also given for both problems, taking advantage of fast rectangular matrix multiplication. In the worst case it runs in $O(n^{3.26})$-time.

6.2.1.2. Branch and Bound Algorithm for computing Pathwidth

It is well known that many NP-hard problems are tractable in the class of bounded pathwidth graphs. In particular, path-decompositions of graphs are an important ingredient of dynamic programming algorithms for solving such problems. Therefore, computing the pathwidth and associated path-decomposition of graphs has both a theoretical and practical interest. In [36], [51], we design a Branch and Bound algorithm that computes the exact pathwidth of graphs and a corresponding path-decomposition. Our main contribution consists of

several non-trivial techniques to reduce the size of the input graph (pre-processing) and to cut the exploration space during the search phase of the algorithm. We evaluate experimentally our algorithm by comparing it to existing algorithms of the literature. It appears from the simulations that our algorithm offers a significative gain with respect to previous work. In particular, it is able to compute the exact pathwidth of any graph with less than 60 nodes in a reasonable running-time (10 min.). Moreover, our algorithm also achieves good performance when used as a heuristic (i.e., when returning best result found within bounded time-limit). Our algorithm is not restricted to undirected graphs since it actually computes the vertex-separation of digraphs (which coincides with the pathwidth in case of undirected graphs).

6.2.1.3. To satisfy impatient Web surfers is hard

Prefetching is a basic mechanism for faster data access and efficient computing. An important issue in prefetching is the tradeoff between the amount of network's resources wasted by the prefetching and the gain of time. For instance, in the Web, browsers may download documents in advance while a Web surfer is surfing. Since the Web surfer follows the hyperlinks in an unpredictable way, the choice of the Web pages to be prefetched must be computed online. The question is then to determine the minimum amount of resources used by prefetching that ensures that all documents accessed by the Web surfer have previously been loaded in the cache. In [28], we model this problem as a two-player game similar to Cops and Robber Games in graphs. Let $k \geq 1$ be any integer. The first player, a fugitive, starts on a marked vertex of a (di)graph G. The second player, an observer, marks at most k vertices, then the fugitive moves along one edge/arc of G to a new vertex, then the observer marks at most k vertices, etc. The fugitive wins if it enters an unmarked vertex, and the observer wins otherwise. The surveillance number of a (di)graph is the minimum k such that the observer marking at most k vertices at each step can win against any strategy of the fugitive. We also consider the connected variant of this game, i.e., when a vertex can be marked only if it is adjacent to an already marked vertex. We study the computational complexity of the game. All our results hold for both variants, connected or unrestricted. We show that deciding whether the surveillance number of a chordal graph is at most 2 is NP-hard. We also prove that deciding if the surveillance number of a DAG is at most 4 is PSPACE-complete. Moreover, we show that the problem of computing the surveillance number is NP-hard in split graphs. On the other hand, we provide polynomial-time algorithms computing surveillance numbers of trees and interval graphs. Moreover, in the case of trees, we establish a combinatorial characterization of the surveillance number.

6.2.2. Tree-decompositions

6.2.2.1. Minimum Size Tree-Decompositions

Tree-Decompositions are the corner-stone 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 based on tree-decompositions is dominated by the size (number of bags) of the tree-decompositions. It is then interesting to try to minimize the size of the tree-decompositions. In [42], [60], we consider the problem of computing a tree-decomposition of a graph with width at most k and minimum size. More precisely, we focus on the following problem: given a fixed $k \geq 1$, what is the complexity of computing a tree-decomposition of width at most k with minimum size in the class of graphs with treewidth at most k? We prove that the problem is NP-complete in planar graphs for any fixed $k \geq 4$ and polynomial for $k \leq 2$. We also show that for $k = 3$ the problem can be solved in polynomial time in the class of trees and 2-connected outerplanar graphs.

6.2.2.2. Exclusive Graph Searching vs. Pathwidth

In Graph Searching, a team of searchers aims at capturing an invisible fugitive moving arbitrarily fast in a graph. Equivalently, the searchers try to clear a contaminated network. The problem is to compute the minimum number of searchers required to accomplish this task. Several variants of Graph Searching have been studied mainly because of their close relationship with the pathwidth of a graph. Blin et al. defined the Exclusive Graph Searching where searchers cannot "jump" and no node can be occupied by more than one searcher. In [61], we study the complexity of this new variant. We show that the problem is NP-hard in

planar graphs with maximum degree 3 and it can be solved in linear time in the class of cographs. We also show that monotone Exclusive Graph Searching is NP-complete in split graphs where Pathwidth is known to be solvable in polynomial time. Moreover, we prove that monotone Exclusive Graph Searching is in P in a subclass of star-like graphs where Pathwidth is known to be NP-hard. Hence, the computational complexities of monotone Exclusive Graph Searching and Pathwidth cannot be compared. This is the first variant of Graph Searching for which such a difference is proved.

6.2.2.3. Diameter of Minimal Separators in Graphs

In [49], 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 graphH* 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.

6.2.3. Distributed computing with mobile agents

6.2.3.1. Stigmergy of Anonymous Agents in Discrete Environments

Communication by stigmergy consists, for agents/robots devoid of other dedicated communication devices, in exchanging information by observing each other's movements, similar to how honeybees use a dance to inform each other on the location of food sources. Stigmergy, while a popular technique in soft computing (e.g., swarm intelligence and swarm robotics), has received little attention from a computational viewpoint, with only one study proposing a method in a continuous environment. An important question is whether there are limits intrinsic to the environment on the feasibility of stigmergy. While it is not the case in a continuous environment, we show that the answer is quite different when the environment is discrete. In [53], [37], we consider stigmergy in graphs and identifies classes of graphs in which robots can communicate by stigmergy. We provide two algorithms with different tradeoffs. One algorithm achieves faster stigmergy when the density of robots is low enough to let robots move independently. This algorithm works when the graph contains some particular pairwise-disjoint subgraphs. The second algorithm, while slower solves the problem under an extremely high density of robots assuming that the graph admits some large cycle. Both algorithms are described in a general way, for any graph that admits the desired properties and with identified nodes. We show how the latter assumption can be removed in more specific topologies. Indeed, we consider stigmergy in the grid which offers additional orientation information not available in a general graphs, allowing us to relax some of the assumptions. Given an $N \times M$ anonymous grid, we show that the first algorithm requires $O(\mathcal{M})$ steps to achieve communication by stigmergy, where $\mathcal{M}$ is the maximum length of a communication message, but it works only if the number of robots is less than $\lfloor \frac{N \cdot M}{9} \rfloor$. The second algorithm, which requires $O(k^2)$ steps, where k is the number of robots, on the other hand, works for up to $N \cdot M - 5$ robots. In both cases, we consider very weak assumptions on the robots capabilities: i.e., we assume that the robots are anonymous, asynchronous, uniform, and execute deterministic algorithms.

6.2.3.2. Gathering and Exclusive Searching on Rings under Minimal Assumptions

Consider a set of mobile robots with minimal capabilities placed over distinct nodes of a discrete anonymous ring. Asynchronously, each robot takes a snapshot of the ring, determining which nodes are either occupied by robots or empty. Based on the observed configuration, it decides whether to move to one of its adjacent nodes or not. In the first case, it performs the computed move, eventually. The computation also depends on the required task. In [38], we solve both the well-known Gathering and Exclusive Searching tasks. In the former problem, all robots must simultaneously occupy the same node, eventually. In the latter problem, the

aim is to clear all edges of the graph. An edge is cleared if it is traversed by a robot or if both its endpoints are occupied. We consider the exclusive searching where it must be ensured that two robots never occupy the same node. Moreover, since the robots are oblivious, the clearing is perpetual, i.e., the ring is cleared infinitely often. In the literature, most contributions are restricted to a subset of initial configurations. Here, we design two different algorithms and provide a characterization of the initial configurations that permit the resolution of the problems under minimal assumptions.

6.2.4. Enhancing the Web's Transparency

Today's Web services – such as Google, Amazon, and Facebook – leverage user data for varied purposes, including personalizing recommendations, targeting advertisements, and adjusting prices. At present, users have little insight into how their data is being used. Hence, they cannot make informed choices about the services they choose.

To increase transparency, we developed *XRay* [40], the first fine-grained, robust, and scalable personal data tracking system for the Web. XRay predicts which data in an arbitrary Web account (such as emails, searches, or viewed products) is being used to target which outputs (such as ads, recommended products, or prices). XRay's core functions are service agnostic and easy to instantiate for new services, and they can track data within and across services. To make predictions independent of the audited service, XRay relies on the following insight: by comparing outputs from different accounts with similar, but not identical, subsets of data, one can pinpoint targeting through correlation. We show both theoretically, and through experiments on Gmail, Amazon, and YouTube, that XRay achieves high precision and recall by correlating data from a surprisingly small number of extra accounts.

6.2.5. Algorithm design in biology

In COATI, we have recently started a collaboration with EPI ABS (Algorithms Biology Structure) from Sophia Antipolis on minimal connectivity complexes in mass spectrometry based macro-molecular complex reconstruction [63]. This problem turns out to be a minimum color covering problem (minimum number of colors to cover colored edges with connectivity constraints on the subgraphs induced by the colors) of the edges of a graph, and is surprizingly similar to a capacity maximization problem in a multi-interfaces radio network we were studying.

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 of finding the most plausible contacts between these subunits, and minimum connectivity inference (MCI) is the variant consisting of 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 [43] two novel algorithms, ALGO-MILP-W and ALGO-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 the yeast exosome, for which both a high resolution crystal structure and a large set of oligomers is known, show that our algorithms predict contacts with high specificity and sensitivity, yielding a very significant improvement over previous work. The software accompanying this paper is made available, and should prove of ubiquitous interest whenever connectivity inference from oligomers is faced.

6.3. Structural Graph Theory

Participants: Jean-Claude Bermond, Frédéric Havet, Nicolas Nisse, Ana Karolinna Maia de Oliveira, Stéphane Pérennes.

More information on several results presented in this section may be found in PhD thesis of A. K. Maia de Oliveira [16], and in the Habilitation thesis of N. Nisse [17].

6.3.1. Graph colouring and applications

Graph colouring is a central problem in graph theory and it has a huge number of applications in various scientific domains (telecommunications, scheduling, bio-informatics, ...). We mainly study graph colouring problems that model ressource allocation problems.

6.3.1.1. Backbone colouring

A well-known channel assignment problem is the following: we are given a graph G, whose vertices correspond to transmitters, together with an edge-weighting w. The weight of an edge corresponds to the minimum separation between the channels on its endvertices to avoid interferences. (If there is no edge, no separation is required, the transmitters do not interfere.) We need to assign positive integers (corresponding to channels) to the vertices so that for every edge e the channels assigned to its endvertices differ by at least $w(e)$. The goal is to minimize the largest integer used, which corresponds to minimizing the *span* of the used bandwidth. We studied a particular, yet quite general, case, called *backbone colouring*, in which there are only two levels of interference. So we are given a graph G and a subgraph H, called *the backone*. Two adjacent vertices in H must get integers at least q apart, while adjacent vertices in G must get integers at distance at least 1. The minimum span in this case is called the q-backbone chromatic number and is denoted $BBC_q(G, H)$. In [30] and [45], we focus on the case when G is planar and H is a forest. In [30], we give a series of NP-hardness results as well as upper bounds for $BBC_q(G, H)$, depending on the type of the forest (matching, galaxy, spanning tree). We also discuss a circular version of the problem. In [45], we give some upper bounds when G is planar and has no cycles of length 4 an 5, and G is a tree, and we relate those results to the celebrated Steinberg's Conjecture stating that every planar graphs with no cycles of length 4 or 5 is 3-colourable.

In [29], we consider the list version of this problem (in which each vertex is given a particular list of admissible colours), with particular focus on colours in $\mathbb{Z}_p$ – this problem is closely related to the problem of circular choosability. We first prove that the list circular q-backbone chromatic number of a graph is bounded by a function of the list chromatic number. We then consider the more general problem in which each edge is assigned an individual distance between its endpoints, and provide bounds using the Combinatorial Nullstellensatz. Through this result and through structural approaches, we achieve good bounds when both the graph and the backbone belong to restricted families of graphs.

6.3.1.2. On-line colouring graphs with few P_4s

Various on-line colouring procedures are used. The most widespread ones is the greedy one, which results in a greedy colouring. Given a graph $G = (V; E)$, a *greedy colouring* of G is a proper colouring such that, for each two colours $i < j$, every vertex of $V(G)$ coloured j has a neighbour with colour i. A second optimization procedure consists from time to time to consider the present colouring and to free some colour when possible: if each vertex of a colour class has another colour that is not used by its neighbours, we can recolour each vertex in the calls by another colour. This procedure results in a b-colouring of the graph. A *b-colouring* of a graph G is a proper colouring such that every colour class contains a vertex which is adjacent to at least one vertex in every other colour class. One of the performance measure of such graph is the maximum numbers of colours they could possibly use. The greatest k such that G has a greedy colouring with k colours is the *Grundy number* of G. The greatest integer k for which there exists a b-colouring of G with k colours is its *b-chromatic number*. Determining the Grundy number and the b-chromatic number of a graph are NP-hard problems in general. For a fixed q, the $(q; q - 4)$-graphs are the graphs for which no set of at most q vertices induces more than $q - 4$ distinct induced P_4s paths of order 4). In [24], we obtain polynomial-time algorithms to determine the Grundy number and the b-chromatic number of $(q; q - 4)$-graphs, for a fixed q. They generalize previous results obtained for cographs and P_4-sparse graphs, classes strictly contained in the $(q; q - 4)$-graphs.

6.3.1.3. Weighted colouring

We also studied weighted colouring which models various problems of shared resources allocation. Given a vertex-weighted graph G and a (proper) r-colouring $c = \{C_1, ..., C_r\}$ of G , the weight of a colour class C_i is the maximum weight of a vertex coloured i and the weight of c is the sum of the weights of its colour classes. The objective of the Weighted Colouring Problem is, given a vertex-weighted graph G , to determine the

minimum weight of a proper colouring of G, that is, its *weighted chromatic number*. In [21], [33], we prove that the Weighted Coloring Problem admits a version of the Hajós' Theorem and so we show a necessary and sufficient condition for the weighted chromatic number of a vertex-weighted graph G to be at least k , for any positive real k. The Weighted Colouring Problem problem remains NP-complete in some particular graph classes as bipartite graphs. In their seminal paper, Guan and Zhu asked whether the weighted chromatic number of bounded tree-width graphs (partial k-trees) can be computed in polynomial-time. Surprisingly, the time-complexity of computing this parameter in trees is still open. We show in [21] that, assuming the Exponential Time Hypothesis (3-SAT cannot be solved in sub-exponential time), the best algorithm to compute the weighted chromatic number of n-node trees has time-complexity $n^{\Theta(\log n)}$. Our result mainly relies on proving that, when computing an optimal proper weighted colouring of a graph G, it is hard to combine colourings of its connected components, even when G is a forest.

6.3.1.4. Inducing proper colourings

Frequently, the proper colouring of the graph must be induced by some other parameters that a vertex can compute locally, for example on looking on the labels assigned to its incident edges or to their orientations.

For a connected graph G of order $|V(G)| \geq 3$ and a k-labelling $c : E(G) \to \{1, 2, ..., k\}$ of the edges of G, the *code* of a vertex v of G is the ordered k-tuple $(\ell_1, \ell_2, ..., \ell_k)$, where ℓ_i is the number of edges incident with v that are labelled i. The k-labelling c is *detectable* if every two adjacent vertices of G have distinct codes. The minimum positive integer k for which G has a detectable k-labelling is the *detection number*$det(G)$ of G. In [31], we show that it is NP-complete to decide if the detection number of a cubic graph is 2. We also show that the detection number of every bipartite graph of minimum degree at least 3 is at most 2. Finally, we give some sufficient condition for a cubic graph to have detection number 3.

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 $po(G)$, is the minimum of the maximum indegree over all its proper orientations. In [32], [44], we prove that $po(G) \leq \left(\Delta(G) + \sqrt{\Delta(G)}\right)/2 + 1$ if G is a bipartite graph, and $po(G) \leq 4$ if G is a tree. It is well-known that $po(G) \leq \Delta(G)$, for every graph G. However, we prove that deciding whether $po(G) \leq \Delta(G) - 1$ 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 $po(G) \leq 2$, for planar *subcubic* graphs G. Moreover, we prove that it is NP-complete to decide whether $po(G) \leq 3$, for planar bipartite graphs G with maximum degree 5.

6.3.2. Directed graphs

Graph theory can be roughly partitioned into two branches: the areas of undirected graphs and directed graphs (digraphs). Even though both areas have numerous important applications, for various reasons, undirected graphs have been studied much more extensively than directed graphs. One of the reasons is that many problems for digraphs are much more difficult than their analogues for undirected graphs.

6.3.2.1. Finding a subdivision of a digraph

One of the cornerstones of modern (undirected) graph theory is minor theory of Robertson and Seymour. Unfortunately, we cannot expect an equivalent for directed graphs. Minor theory implies in particular that, for any fixed F, detecting a subdivision of F in an input graph G can be performed in polynomial time by the Robertson and Seymour linkage algorithm. In contrast, the analogous subdivision problem for digraph can be either polynomial-time solvable or NP-complete, depending on the fixed digraph F. In [16], a number of examples of polynomial instances, several NP-completeness proofs as well as a number of conjectures and open problems are given. In addition, it is conjectured that, for every integer k greater than 1, the directed cycles of length at least k have the Erdős-Pósa Property : for every n, there exists an integer t_n such that for every digraph D, either D contains n disjoint directed cycles of length at least k, or there is a set T of t_n vertices that meets every directed cycle of length at least k. This generalizes a celebrated result of Reed, Robertson, Seymour and Thomas which is the case $k = 2$ of this conjecture. We prove the conjecture

for $k = 3$. We also show that the directed k-Linkage problem is polynomial-time solvable for digraphs with circumference at most 2. From these two results, we deduce that if F is the disjoint union of directed cycles of length at most 3, then one can decide in polynomial time if a digraph contains a subdivision of F.

6.3.2.2. The complexity of finding arc-disjoint branching flows

The concept of arc-disjoint flows in networks is a very general framework within which many well-known and important problems can be formulated. In particular, the existence of arc-disjoint branching flows, that is, flows which send one unit of flow from a given source s to all other vertices, generalizes the concept of arc-disjoint out-branchings (spanning out-trees) in a digraph. A pair of out-branchings $B^+_{s,1}$, $B^+_{s,2}$ from a root s in a digraph $D = (V, A)$ on n vertices corresponds to arc-disjoint branching flows x_1,x_2 (the arcs carrying flow in x_i are those used in $B^+_{s,i}$, $i = 1, 2$) in the network that we obtain from D by giving all arcs capacity $n-1$. It is then a natural question to ask how much we can lower the capacities on the arcs and still have, say, two arc-disjoint branching flows from the given root s. In [46], we prove that for every fixed integer ≥ 2 it is

- an NP-complete problem to decide whether a network $\mathcal{N} = (V, A, u)$ where $u_{ij} = k$ for every arc ij has two arc-disjoint branching flows rooted at s.
- a polynomial problem to decide whether a network $\mathcal{N} = (V, A, u)$ on n vertices and $u_{ij} = n - k$ for every arc ij has two arc-disjoint branching flows rooted at s.

The algorithm for the later result generalizes the polynomial algorithm, due to Lovász, for deciding whether a given input digraph has two arc-disjoint out-branchings rooted at a given vertex. Finally we prove that under the so-called Exponential Time Hypothesis (ETH), for every $\epsilon > 0$ and for every $k(n)$ with $(\log(n))^{1+\epsilon} \leq k(n) \leq \frac{n}{2}$ (and for every large i we have $k(n) = i$ for some n) there is no polynomial algorithm for deciding whether a given digraph contains two arc-disjoint branching flows from the same root so that no arc carries flow larger than $n - k(n)$.

6.3.2.3. Splitting a tournament into two subtournaments with given minimum outdegree

A (k_1, k_2)-*outdegree-splitting* of a digraph D is a partition (V_1, V_2) of its vertex set such that $D[V_1]$ and $D[V_2]$ have minimum outdegree at least k_1 and k_2, respectively. In [58], we show that there exists a minimum function f_T such that every tournament of minimum outdegree at least $f_T(k_1, k_2)$ has a (k_1, k_2)-outdegree-splitting, and $f_T(k_1, k_2) \leq k_1^2/2 + 3k_1/2 + k_2 + 1$. We also show a polynomial-time algorithm that finds a (k_1, k_2)-outdegree-splitting of a tournament if one exists, and returns 'no' otherwise. We give better bound on f_T and faster algorithms when $k_1 = 1$.

6.3.2.4. Eulerian and Hamiltonian dicycles in directed hypergraphs

In [19], we generalize the concepts of Eulerian and Hamiltonian digraphs to directed hypergraphs. A *dihypergraphH* is a pair $(\mathcal{V}(H), \mathcal{E}(H))$, where $\mathcal{V}(H)$ is a non-empty set of elements, called *vertices*, and $\mathcal{E}(H)$ is a collection of ordered pairs of subsets of $\mathcal{V}(H)$, called *hyperarcs*. It is Eulerian (resp. Hamiltonian) if there is a dicycle containing each hyperarc (resp. each vertex) exactly once. We first present some properties of Eulerian and Hamiltonian dihypergraphs. For example, we show that deciding whether a dihypergraph is Eulerian is an NP-complete problem. We also study when iterated line dihypergraphs are Eulerian and Hamiltonian. Finally, we study when the generalized de Bruijn dihypergraphs are Eulerian and Hamiltonian. In particular, we determine when they contain a complete Berge dicycle, i.e. an Eulerian and Hamiltonian dicycle.

DANTE Team

6. New Results

6.1. Highlights of the Year

6.1.1. The Internet of Things: A new equipments of excellence

Inaugurated last autumn, the very large scale IoT-LAB platform (https://www.iot-lab.info) is strengthening the capabilities of the FIT equipment of excellence dedicated to the Internet of Things. Offering a unique wide-ranging collection of equipment, these laboratories are available to both researchers and commercial companies alike.

IoT-LAB is a large-scale experimental platform for communicating objects and networks of sensors. It enables the rapid deployment of experiments and the collection of large amounts of data. It includes over 2700 sensor nodes, distributed over six sites in France, offering a wide range of different processor architectures and radio components. IoT-LAB is available for use on line. It is already used by over 300 users in forty countries, including around ten commercial companies. As of the end of October 2014, some 10 000 experiments had already been carried out.

6.1.2. Graph-based signal processing

Our first results towards the definition of a digital framework for signal processing on graphs constitutes an important outcome of DANTE's activity in 2014. Our participation to this emerging discipline was marked with several scientific recognitions: publication in the main DSP conference [14], involvement in the first ANR project focusing on this theme and retained for funding (2015-2019), we are in charge of the organisation of a Special Session dedicated to "Methodologies for signal processing on graphs" at Eusipco conference (2015).

6.1.3. Complex contagion process

Diffusion of innovation can be interpreted as a social spreading phenomena governed by the impact of media and social interactions. Although these mechanisms have been identified by quantitative theories, their role and relative importance are not entirely understood, since empirical verification has so far been hindered by the lack of appropriate data. Here we analyse a dataset recording the spreading dynamics of the world's largest Voice over Internet Protocol service to empirically support the assumptions behind models of social contagion. We show that the rate of spontaneous service adoption is constant, the probability of adoption via social influence is linearly proportional to the fraction of adopting neighbors, and the rate of service termination is time-invariant and independent of the behavior of peers. By implementing the detected diffusion mechanisms into a dynamical agent-based model, we are able to emulate the adoption dynamics of the service in several countries worldwide. This approach enables us to make medium-term predictions of service adoption and disclose dependencies between the dynamics of innovation spreading and the socioeconomic development of a country. This work was recently published in the Journal of the Royal Society Interface.

6.2. Diffusion and dynamic of complex networks

Participants: Márton Karsai [correspondant], Éric Fleury, Christophe Crespelle.

Time varying networks and the weakness of strong ties We analyse a mobile call dataset and find a simple statistical law that characterize the temporal evolution of users' egocentric networks. We encode this observation in a reinforcement process defining a time-varying network model that exhibits the emergence of strong and weak ties. We study the effect of time-varying and heterogeneous interactions on the classic rumor spreading model in both synthetic, and real-world networks. We observe that strong ties severely inhibit information diffusion by confining the spreading process among agents with recurrent communication patterns. This provides the counterintuitive evidence that strong ties may have a negative role in the spreading of information across networks.

Complex contagion process in spreading of online innovation [8]. Here we analyse a dataset recording the spreading dynamics of the world's largest Voice over Internet Protocol service to empirically support the assumptions behind models of social contagion. We show that the rate of spontaneous service adoption is constant, the probability of adoption via social influence is linearly proportional to the fraction of adopting neighbors, and the rate of service termination is time-invariant and independent of the behavior of peers. By implementing the detected diffusion mechanisms into a dynamical agent-based model, we are able to emulate the adoption dynamics of the service in several countries worldwide. This approach enables us to make medium-term predictions of service adoption and disclose dependencies between the dynamics of innovation spreading and the socio-economic development of a country.

The role of endogenous and exogenous mechanisms in the formation of R&D networks [10]. Here we propose a general modeling framework that includes both endogenous and exogenous mechanisms of in link formations in networks with tunable relative importance. The model contains additional ingredients derived from empirical observations, such as the heterogeneous propensity to form alliances and the presence of circles of influence, i.e. clusters of firms in the network. We test our model against the Thomson Reuters SDC Platinum dataset, one of the most complete datasets available nowadays, listing cross-country R&D alliances from 1984 to 2009. Interestingly, by fitting only three macroscopic properties of the network, this framework is able to reproduce a number of microscopic measures characterizing the network topology, including the distributions of degree, local clustering, path length and component size, and the emergence of network clusters. Furthermore, by estimating the link probabilities towards newcomers and established firms from the available data, we find that endogenous mechanisms are predominant over the exogenous ones in the network formation. This quantifies the importance of existing network structures in selecting partners for R&D alliances.

Controlling Contagion Processes in Time-Varying Networks [9]. In this project we derive an analytical framework for the study of control strategies specifically devised for time-varying networks. We consider the removal/immunization of individual nodes according the their activity in the network and develop a block variable mean-field approach that allows the derivation of the equations describing the evolution of the contagion process concurrently to the network dynamic. We derive the critical immunization threshold and assess the effectiveness of the control strategies. Finally, we validate the theoretical picture by simulating numerically the information spreading process and control strategies in both synthetic networks and a large-scale, real-world mobile telephone call dataset.

Data-driven spreading for the detection of weak ties [24]. In this work we propose a new method to infer the strength of social ties by using new data-driven simulation techniques. We qualify links by the importance they play during the propagation of information in the social structure. We apply data-driven spreading processes combined with a river-basin algorithmic method to identify links, which are the responsible to bring the information to large number of nodes. We investigate the correlations of the new importance measure with other conventional characteristics and identify their best combination through a percolation analysis to sophisticate further the assignment of social tie strengths. Finally we explore the role of the identified high importance links in control of globally spreading processes through data-driven SIR model simulations. These results point out that the size of infected population can be reduced considerably by weakening interactions through ties with high importance but zero overlap compared to strategies based on dyadic communications.

Dynamic Contact Network Analysis in Hospital Wards [18]. We analyse a huge and very precise trace of contact data collected during 6 months on the entire population of a rehabilitation hospital. We investigate the graph structure of the average daily contact network. Our main results are to unveil striking properties of this structure in the considered hospital, and to present a methodology that can be used for analyzing any dynamic complex network where nodes are classified into groups.

6.3. Performance analysis and networks protocols

Participants: Anthony Busson [correspondant], Thomas Begin, Isabelle Guérin Lassous.

Modeling and optimization of CSMA/CA in VANET [7]. We propose a simple theoretical model to compute the maximum spatial reuse feasible in a VANET. We focus on the ad hoc mode of the IEEE 802.11p standard. Our model offers simple and closed-form formulas on the maximum number of simultaneous transmitters, and on the distribution of the distance between them. It leads to an accurate upper bound on the maximum capacity. In order to validate our approach, results from the analytical models are compared to simulations performed with the network simulator NS-3. We take into account different traffic distributions (traffic of vehicles), and study the impact of this traffic on capacity. An application of this work is the parameterization of the CSMA/CA mechanism.

Fast and accurate approximate performance analysis of multi-server facilities [4]. Systems with multiple servers are common in many areas and their correct dimensioning is in general a difficult problem under realistic assumptions on the pattern of user arrivals and service time distribution. We present an approximate solution for the underlying $Ph/Ph/c/N$ queueing model. Our approximation decomposes the solution of the Ph/Ph/c/N queue into solutions of simpler $M/Ph/c/N$ and $Ph/M/c/N$ queues. To further mitigate dimensionality issues, for larger numbers of servers and/or service time phases, we use a reduced state approximation to solve the $M/Ph/c/N$ queue. The proposed approach is conceptually simple, easy to implement and produces generally accurate results for the mean number in the system, as well as the loss probability. Typical relative errors for these two quantities are below 5%. A very significant speed advantage compared to the numerical solution of the full $Ph/Ph/c/N$ queue can be gained as the number of phases representing the arrival process and/or the number of servers increases.

Interference and throughput in spectrum sensing cognitive radio networks using point processes .
Spectrum sensing is vital for secondary unlicensed nodes to coexist and avoid interference with the primary licensed users in cognitive wireless networks. In this paper, we develop models for bounding interference levels from secondary network to the primary nodes within a spectrum sensing framework. Instead of classical stochastic approaches where Poisson point processes are used to model transmitters, we consider a more practical model which takes into account the medium access control regulations and where the secondary Poisson process is judiciously thinned in two phases to avoid interference with the secondary as well as the primary nodes. The resulting process will be a modified version of the Mate´rn point process. For this model, we obtain bounds for the complementary cumulative distribution function of interference and present simulation results which show the developed analytical bounds are quite tight. Moreover, we use these bounds to find the operation regions of the secondary network such that the interference constraint is satisfied on receiving primary nodes. We then obtain theoretical results on the primary and secondary throughputs and find the throughput limits under the interference constraint.

Modeling of IEEE 802.11 Multi-hop Wireless Chains with Hidden Nodes [11]. We follow up an existing modeling framework to analytically evaluate the performance of multi-hop flows along a wireless chain of four nodes. The proposed model accounts for a non-perfect physical layer, handles the hidden node problem, and is applicable under workload conditions ranging from flow(s) with low intensity to flow(s) causing the network to saturate. Its solution is easily and quickly obtained and delivers estimates for the expected throughput and for the datagram loss probability of the chain with a good accuracy.

Anticipation of ETX Metric to manage Mobility in Ad Hoc Wireless Networks [19]. When a node is moving in a wireless network, the routing metrics associated to its wireless links may reflect link quality degradations and help the routing process to adapt its routes. Unfortunately, an important delay between the metric estimation and its inclusion in the routing process makes this approach inefficient. In this paper, we introduce an algorithm that predicts metric values a few seconds in advance, in order to compensate the delay involved by the link quality measurement and their dissemination by the routing protocol. We consider classical metrics, in particular ETX (Expected Transmission Count) and ETT (Expected Transmission Time), but we combine their computations

to our prediction algorithm. Extensive simulations show the route enhancement as the Packet Delivery Ratio (PDR) is close to 1 in presence of mobility.

6.4. Graphs & Signal Processing

Participants: Paulo Gonçalves [correspondant], Éric Fleury, Christophe Crespelle.

6.4.1. Signal Processing on Graphs

Semi-Supervised Learning for Graph to Signal Mapping: a Graph Signal Wiener Filter Interpretation [14]. We investigate a graph to signal mapping with the objective of analyzing intricate structural properties of graphs with tools borrowed from signal processing. We successfully use a graph-based semi-supervised learning approach to map nodes of a graph to signal amplitudes such that the resulting time series is smooth and the procedure efficient and scalable. Theoretical analysis of this method reveals that it essentially amounts to a linear graph-shift-invariant filter with the a priori knowledge put into the training set as input. Further analysis shows that we can interpret this filter as a Wiener filter on graphs. We finally build upon this interpretation to improve our results.

6.4.2. Graphs

(Nearly-)tight bounds on the contiguity and linearity of cographs [6]. In this paper we show that the contiguity and linearity of cographs on n vertices are both $O(\log n)$. Moreover, we show that this bound is tight for contiguity as there exists a family of cographs on n vertices whose contiguity is $\Omega(\log n)$. We also provide an $\Omega(\log n/\log\log n)$ lower bound on the maximum linearity of cographs on n vertices. As a by-product of our proofs, we obtain a min-max theorem, which is worth of interest in itself, stating equality between the rank of a tree and the minimum height of one of its path partitions.

6.4.3. Signal processing

Analysis of intrapartum foetal heart rate (FHR), enabling early detection of foetal acidosis to prevent asphyxia and labour adverse outcomes, remains a challenging signal processing task. In this direction, we carried out a series of works to characterize the fetal heart rate variability with specific attributes able to discriminate between healthy fetuses and fetuses presenting a risk of brain injury. Last year, we investigated two different approaches:

Nearest-Neighbor based Wavelet Entropy Rate Measures for Intrapartum Fetal Heart Rate Variability [23]. Firstly, we showed that a k-nearest neighbor procedure yields estimates for entropy rates that are robust and well-suited to FHR variability. Secondly, we experimentally proved that entropy rates measured on multiresolution wavelet coefficients permit to improve classification performance.

Impacts of labour first and second stages on Hurst parameter based intrapartum FHR analysis [22]. In this study, we proposed to quantify the FHR temporal dynamics with a Hurst exponent estimated within a wavelet framework. Analyses performed over a large (3049 records) and well documented database revealed that the evolution of the Hurst exponent during delivery, is significantly different for healthy fetuses and for acidotic fetuses.

6.5. Complex network metrology

Participant: Christophe Crespelle.

Measuring the Degree Distribution of Routers in the Core Internet [15]. Most current models of the internet rely on knowledge of the degree distribution of its core routers, which plays a key role for simulation purposes. In practice, this distribution is usually observed directly on maps known to be partial, biased and erroneous. This raises serious concerns on the true knowledge one may have of this key property. Here, we design an original measurement approach targeting reliable estimation of the degree distribution of core routers, without resorting to any map. It consists in sampling random core routers and precisely estimate their degree thanks to probes sent from many distributed monitors. We run and assess a large-scale measurement following this approach, carefully controlling and

correcting bias and errors encountered in practice. The estimate we obtain is much more reliable than previous knowledge, and it shows that the true degree distribution is very different from all current assumptions.

Measuring Routing Tables in the Internet [21]. The most basic function of an Internet router is to decide, for a given packet, which of its interfaces it will use to forward it to its next hop. To do so, routers maintain a routing table, in which they look up for a prefix of the destination address. The routing table associates an interface of the router to this prefix, and this interface is used to forward the packet. We explore here a new measurement method based upon distributed UDP probing to estimate this routing table for Internet routers.

DIANA Team

5. New Results

5.1. Highlights of the Year

Arnaud Legout and Thierry Parmentelat designed and realized the very first Inria Mooc hosted on the FUN platform. This Mooc is devoted to the study of the Python language, and targets undergrandudate students. The objective of the course is to give students a thorough understanding of the internal mechanisms of language, and lead them to small and realistic applications. This Mooc was a big success: 9166 persons registered to the course, out of them five hundred followed the whole course and more than a hundred finished the project. For more details on this Mooc see https://www.france-universite-numerique-mooc.fr/courses/inria/41001/Trimestre_4_2014/about.

5.2. From network-level measurements to expected QoE: the Skype use case

Contributors: Salim Afra, Chadi Barakat and Damien Saucez. Applications rely on rich multimedia contents and experience of end users is sensitive to network conditions. Consequently, network operators must design their infrastructure to ensure high Quality of Experience (QoE) for their customers. However, applications are usually over-the-top services on which network operators have no control and users have no mean to tune the network when they undergo poor QoE. In this project, called ACQUA for Application for the Prediction of Quality of Experience at Internet Access, we propose a new approach that allows network operators to determine how their network performance will influence QoE and end users to predict the QoE even before launching their applications. We predict the subjective QoE users will undergo based on the knowledge of objective network performance parameters obtained with active measurements (e.g., delay, loss) and machine learning. With the particular case of Skype calls and using a decision tree, we show that our approach achieves 83% of accuracy when estimating QoE from the delay, bandwidth, and loss. Our approach can be seen as a new way of performing measurements at the Internet access, where instead of expressing the expected performance in terms of network-level measurements, the performance of the access is expressed in clear terms related to the expected quality for the main applications of interest to the end user. The strength of the approach is in its capacity of expressing directly the QoE as a function of network-level measurements, which is an enabler for QoE prediction, and in reusing the same network-level measurements as input to different models for the QoE of end user applications. More details on this approach and on our application ACQUA can be found in section 4.5 , in the report summarizing the results [24] and on the application web page http://team.inria.fr/diana/acqua/.

5.3. Understanding of modern web traffic

Contributors: Salim Afra, Chadi Barakat, Byungchul Park and Damien Saucez.

Mobile devices are everywhere nowadays but little is known about the way they differ from traditional non-mobile devices in terms of usage and the characteristics of the web traffic they generate. In this contribution, we propose a first study of the differences that exist between mobile and non-mobile Web traffic seen from the lorgnette of a university campus network. The study is performed at different levels starting from users' behavior to transport protocol configurations. Our main findings are that mobile users often browse websites tailored to their devices. They show a significant adoption of Apps to browse the web and a preference for multimedia content. The different way of conceiving the web for mobiles is reflected at the HTTP and TCP levels with much less HTTP redirections and abrupt TCP connection terminations. Interestingly, mobile traffic carries larger contents and have larger TCP flows than non-mobile traffic. By cross-analysis of protocols and users' behavior, we explain why TCP flows in mobile traffic are larger than those of non-mobiles. Further details on this study can be found in [30].

5.4. Characterizing ICMP Rate Limitation on Routers

Contributors: Chadi Barakat and Ricardo Ravaioli.

In the last decade, path discovery has been extensively covered in the literature. In its simplest form, it generally works by sending probes that expire along the path from a host to a destination. It is also known that network administrators often configure their routers to limit the amount of ICMP replies sent, a common practice typically referred to as ICMP rate limitation. In this contribution we attempt to characterize the responsiveness of routers to expiring ICMP echo-request packets. Our contribution is twofold: first, we provide a detailed analysis of how routers are most commonly configured to respond to expiring packets; next, we show that for the vast majority of routers the measured round-trip time is not affected by the probing rate. This contribution is published in ICC'2015 [21]. It is the result of a collaboration with the SIGNET group at I3S in the context of a PhD thesis funded by the UCN@SOPHIA Labex.

5.5. Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph

Contributors: Maksym Gabielkov and Arnaud Legout.

Twitter is one of the largest social networks using exclusively directed links among accounts. This makes the Twitter social graph much closer to the social graph supporting real life communications than, for instance, Facebook. Therefore, understanding the structure of the Twitter social graph is interesting not only for computer scientists, but also for researchers in other fields, such as sociologists. However, little is known about how the information propagation in Twitter is constrained by its inner structure. We have performed an in-depth study of the macroscopic structure of the Twitter social graph unveiling the highways on which tweets propagate, the specific user activity associated with each component of this macroscopic structure, and the evolution of this macroscopic structure with time for the past 6 years. For this study, we crawled Twitter to retrieve all accounts and all social relationships (follow links) among accounts; the crawl completed in July 2012 with 505 million accounts interconnected by 23 billion links. Then, we proposed a methodology to unveil the macroscopic structure of the Twitter social graph. This macroscopic structure consists of 8 components defined by their connectivity characteristics. Each component group users with a specific usage of Twitter. For instance, we identified components gathering together spammers, or celebrities. Finally, we presented a method to approximate the macroscopic structure of the Twitter social graph in the past, validate this method using old datasets, and discuss the evolution of the macroscopic structure of the Twitter social graph during the past 6 years. This study was published in ACM Sigmetrics 2014 [17].

5.6. When AIMD meets ICN: a bandwidth sharing perspective

Contributors: Chadi Barakat and Damien Saucez.

Information-centric networking (ICN) leverages content demand redundancy and proposes in-network caching to reduce network and servers load and to improve quality of experience. In this contribution, we study the interaction between in-network caching of ICN and Additive Increase Multiplicative Decrease (AIMD) end-to-end congestion control with a focus on how bandwidth is shared, as a function of content popularity and cache provisioning. As caching shortens AIMD feedback loop, the download rate of AIMD is impacted. Supported by an analytical model based on Discriminatory Processor Sharing and real experiments, we observe that popular contents benefit from caching and realize a shorter download time at the expense of unpopular contents, which see their download time inflated by a factor bounded by $1/(1-\rho)$, where ρ is the network load. This bias can be removed by redefining congestion control to be delay independent or by over-provisioning link capacity at the edge so that to compensate for the greediness of popular contents. Further details on this study, which is the result of a collaboration with Politecnico di Bari, can be found in [23].

5.7. On the incentives and incremental deployments of ICN technologies for OTT services

Contributors: Chadi Barakat and Damien Saucez.

With the explosion of broadband Over-The-Top (OTT) services, the Internet is autonomously migrating toward overlay and incrementally deployable content distribution infrastructures. Information-Centric Networking (ICN) technologies are the natural candidates to efficiently distribute popular content to users. However, the strategic incentives in exploiting ICN, for both users and ISPs, are much less understood to date. We hence studied in [15] the strategic incentives for ICN overlay adoption in OTT services based on a game theoretical approach and discussed how OTTs shall shape their prices to motivate ICN overlay usages.

5.8. On ICN Cache Allocation to Content Providers

Contributor: Damien Saucez
Cross-Team Contributors: Mahmoud El Chamie (Maestro)
External contributors: Sahar Hoteit and Stefano Secci from Sorbonne Universités, UPMC Univ Paris 06.

Information Centric Networks (ICNs) allow offloading content distribution from content service providers by means of in-network caching. Despite a rather high maturation in the definition of ICN forwarding techniques, minor attention has been given to the strategic interaction among the multiple ICN stakeholders. We decided to focus on situations involving multiple Content Providers (CPs) and one ICN provider having to give them access to its caches. Intuitively, this situation is prone to high cache contention, in particular at the appealing topology cross-points. To address this problem we propose a resource allocation and pricing framework to support the network provider in the cache allocation to multiple CPs, for situations where CPs have non-overlapping sets of files and untruthful demands need to be avoided. As cache imputations to CPs need to be fair and robust against overclaiming, we evaluated common proportional and max-min fairness (PF, MMF) allocation rules, as well as coalitional game rules, the Nucleolus and the Shapley value. We found that the naive least-recently-used-based ICN approach provides proportional fairness. Moreover, the game-theoretic rules outperform in terms of content access latency the naive ICN approach as well as PF and MMF approaches, while sitting in between PF and MMF in terms of fairness. This paper is under submission [27].

5.9. Demonstrating a unified ICN development and evaluation framework

Contributors: Walid Dabbous, Alina Quereilhac, Damien Saucez and Thierry Turletti.

Information-Centric Networking solutions target world-wide deployment in the Internet. It is hence necessary to have access to a development and evaluation environment which enables both controllable and realistic experimentation to thoroughly understand how ICN solutions would behave in real life deployment. Such solution can be obtained with NEPI that we demonstrated at the ACM Information Centric Networking 2014 conference. In this demonstration, we presented a development and evaluation framework that combines emulation and live prototyping environments to provide ICN designers and implementers the means to build beyond-prototype ICN solutions. This framework is built upon NEPI. We demonstrated the benefits of such integrated approach by showing how complete experimental studies can be carried out with minimum manual intervention and experiment set-up overhead, in both emulation and live environments. More precisely, we demonstrated how to deploy the same experiment in different environment and how NEPI can help to minimise the implementation and operational overhead. This demonstration is summarised in [31].

5.10. Optimizing rules placement in OpenFlow networks: trading routing for better efficiency

Contributors: Chadi Barakat, Xuan Nam Nguyen, Damien Saucez and Thierry Turletti

The idea behind Software Defined Networking (SDN) is to conceive the network as one programmable entity rather than a set of devices to manually configure, and OpenFlow meets this objective. In OpenFlow, a centralized programmable controller installs forwarding rules onto switches to implement policies. However, this flexibility comes at the expense of extra overhead as the number of rules might exceed the memory capacity of switches, which raises the question of how to place most profitable rules on board. Solutions proposed so far strictly impose paths to be followed inside the network. We advocate instead that we can relax routing requirements within the network to concentrate on the final destination to which the traffic should forwarded, not how to route to this destination. In [19] we illustrate the concept, with an optimization problem that gets the maximum amount of traffic delivered according to policies and the actual dimensioning of the network. The traffic that cannot be accommodated is forwarded to the controller that has the capacity to process it further. [19] also demonstrates that our approach permits a better utilization of scarce resources in the network. We extended the work by stating that in many situations (e.g., data-center networks), the exact path followed by packets has not significant impact on performances as long as packets are delivered to their final destination decided by the endpoint policy. It is thus possible to deviate part of the traffic to alternative paths so as to better use network resources without violating the endpoint policy. In [20], we propose a linear optimization model of the rule allocation problem in resource constrained OpenFlow networks with loose routing policies. We show that the general problem is NP-hard and propose a polynomial time heuristic, called OFFICER, that aims at maximizing the amount of carried traffic in under-provisioned networks. Our numerical evaluation on four different topologies show that exploiting various paths allows to increase the amount of traffic supported by the network without significantly increasing the path length.

5.11. A Survey of Software-Defined Networking

Contributors: Bruno Astuto Arouche Nunes, Xuan Nam Nguyen and Thierry Turletti.

We wrote a survey of the emerging field of Software-Defined Networking (SDN). SDN is currently attracting significant attention from both academia and industry. Its field is quite recent, yet growing at a very fast pace. Still, there are important research challenges to be addressed. We look at the history of programmable networks, from early ideas until recent developments. In particular we described the SDN architecture in detail as well as the OpenFlow standard. We provided an overview of current SDN implementations and testing platforms and examined network services and applications that have been developed based on the SDN paradigm. We concluded with a discussion of future directions enabled by SDN ranging from support for heterogeneous networks to Information Centric Networking (ICN). The survey has been published in the IEEE Surveys and Tutorials journal [9]. This paper is among the top downloads on IEEE Exlpore in December 2014. See http://ieeexplore.ieee.org/xpl/browsePopular.jsp?reload=true.

5.12. Software-Defined Networking Enabled Capacity Sharing in User Centric Networks

Contributors: Bruno Astuto Arouche Nunes and Thierry Turletti.

We proposed to use SDN to deploy capacity sharing mechanisms in the context of User Centric Networking (UCN). We consider user-centric networks as a way of considerably mitigating the problem of sharing limited network capacity and resources efficiently and in a fairly manner. UCNs are self-organizing networks where the end user plays an active role in delivering networking functions such as providing Internet access to other users. We propose to leverage the SDN paradigm to enable cooperation between wireless nodes and to provide capacity sharing services in UCNs. Our proposed approach allows coverage of existing network infrastructure (e.g., WiFi or 3GPP) to be extended to other end users or ad hoc networks that would otherwise not be able to have access to network connectivity and services. Moreover, it takes into account current network load and conditions, and QoS requirements of applications. This work has been published in a special issue of Communications Magazine [14].

5.13. Decentralizing SDN's Control Plane

Contributors: Bruno Nunes Astuto and Thierry Turletti.

Motivated by the internets of the future that will likely be considerably larger in size as well as highly heterogeneous and decentralized, we sketched out a framework aiming to enable not only physical, but also logical distribution of the Software-Defined Networking (SDN) control plane. This framework will accomplish network control distribution by defining a hierarchy of controllers that can "match" an internet's organizational– and administrative structure. The main idea is to delegate control between main controllers and secondary controllers in order to accommodate administrative decentralization and autonomy. This work has been presented in a short paper at the IEEE LCN conference [22].

5.14. Extending DCE to emulate Wireless Software Defined Networks.

Participants: Emilio Mancini, Hardik Soni, Thierry Turletti and Walid Dabbous.

Today it is not possible to simulate and evaluate in a realistic way wireless Software Defined Networking solutions. Indeed, the most used SDN emulator tool, Mininet, can only emulate point-to-point physical links using virtual Ethernet pairs (e.g., MAC layer is ignored), and it cannot provide mobility models for wireless nodes. To make the Direct Code Execution module (DCE) able to run Software Defined Networks we started to support OpenFlow NOX controller and Open vSwitch. The actual NOX binary is executed on a simulated ns-3 node. OpenFlow wireless routers are simulated using the Open vSwitch distribution with data-path kernel module support as it is widely used. DCE provides a mechanism to incorporate such a kernel module based application execution. A demonstration has been done at the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems [29].

5.15. On the Performance of the LISP Beta Network

Contributor: Damien Saucez

The future Internet has been a hot topic during the past decade and many approaches towards this future Internet, ranging from incremental evolution to complete clean slate ones, have been proposed. One of the proposition, LISP, advocates for the separation of the identifier and the locator roles of IP addresses to reduce BGP churn and BGP table size. Up to now, however, most studies concerning LISP have been theoretical and, in fact, little is known about the actual LISP deployment performance. In [16], we report the measurement campaigns carried out on the LISP Beta Network. More precisely, we evaluated the performance of the two key components of the infrastructure: the control plane (i.e., the mapping system) and the interworking mechanism (i.e., communication between LISP and non-LISP sites). Our measurements highlight that performance offered by the LISP interworking infrastructure is strongly dependent on BGP routing policies. If we exclude misconfigured nodes, the mapping system typically provides reliable performance and relatively low median mapping resolution delays. Although the bias is not very important, control plane performance favors USA sites as a result of its larger LISP user base but also because European infrastructure appears to be less reliable.

This work resulted in a collaboration with Telecom ParisTech starting in mid-2014 a PhD thesis on the feasibility of large scale measurement of LISP networks with Luigi Iannone as advisor and Damien Saucez as co-advisor.

5.16. Standardization: Contributions to the IETF LISP WG

Contibutor: Damien Saucez

In the context of the LISP WG, we contributed to an Internet-draft called "An Architectural Introduction to the LISP Location-Identity Separation System" [25] that describes the architecture of the Locator/ID Separation Protocol (LISP), making it easier to read the rest of the LISP specifications and providing a basis for discussion about the details of the LISP protocols. This document is used for introductory purposes, more details can be found in RFC6830, the protocol specification. This internet-draft is in RFC queue, for imminent publication as RFC.

In the context of the LISP WG, we contributed to an Internet-draft called "LISP Threats Analysis" [33] that proposes a threat analysis of the Locator/Identifier Separation Protocol (LISP). This internet-draft is under discussion in the Working Group.

In the context of the LISP WG, we contributed to an Internet-draft called "LISP-Security (LISP-SEC)" [28] that specifies LISP-SEC, a set of security mechanisms that provides origin authentication, integrity and anti-replay protection to LISP's EID-to-RLOC mapping data conveyed via mapping lookup process. LISP-SEC also enables verification of authorization on EID-prefix claims in Map-Reply messages. This internet-draft is under discussion in the Working Group.

In the context of the LISP WG, we contributed to an Internet-draft called "LISP Impact" [32]. The Locator/Identifier Separation Protocol (LISP) aims at improving the Internet scalability properties leveraging on three simple principles: address role separation, encapsulation, and mapping. In this internet-draft, based on implementation, deployment, and theoretical studies, we discuss the impact that deployment of LISP can have on both the Internet in general and for the end-users in particular. This internet-draft is adopted as Working Group document on December 2014.

DIONYSOS Project-Team

5. New Results

5.1. Highlights of the Year

Pierre L'Ecuyer received the Award of Merit from the Canadian Operational Research Society, 2014.

We had one best paper award in 2014 on a novel architecture for resilient networks (see 5.8).
BEST PAPER AWARD :
[50] **IEEE International Conference on Innovations for Community Services**. D. LEQUÉRÉ, C. BETOULE, G. THOUENON, Y. HADJADJ-AOUL, A. KSENTINI, R. CLAVIER.

5.2. Quality of Experience

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, Gerardo Rubino, Bruno Sericola, Pantelis Frangoudis, César Viho, Quang Pham Tran Anh.

PSQA. We continue the development of the PSQA technology (Pseudo-Subjective Quality Assessment) in the area of Quality of Experience (QoE). PSQA is today a mature technology allowing to build measuring modules capable of quantifying the quality of a video or an audio sequence, as perceived by the user, when received through an IP network. It provides an accurate and efficiently computed evaluation of quality. Accuracy means that PSQA gives values close to those that can be obtained from a panel of human observers, under a controlled subjective testing experiment, following an appropriate standard (which depends on the type of sequence or application). Efficiency means that our measuring tool can work in real time. Observe that perceived quality is, in general, the main component of QoE when the application or service involves video and audio, or voice. PSQA works by analyzing the networking environment of the communication and some the technical characteristics of the latter. It works without any need to the original sequence (as such, it belongs to the family of *no-reference* techniques). It must be pointed out that a PSQA measuring or monitoring module is network-dependent and application-dependent. Basically, for each specific networking technology, and for any application or service, the module must be built from scratch. But once built, it works automatically and efficiently, allowing if necessary its use in real time, typically for controlling purposes.

QoE and SLA. On the applications side, we focused this year on using QoE estimates to drive service/application-level decisions. As a first use case, we proposed a multi-objective optimization framework for the problem of optimally selecting among a set of available hosting and network connectivity Service-Level Agreements (SLAs) for the migration of enterprise communication services (such as teleconferencing) to the Cloud [59]. Our framework captures the tradeoff between user experience and deployment cost, and offers a service provider the opportunity to weight these two conflicting criteria based on its preferences. Our approach is generic and can be applied to various application settings by appropriately selecting application-specific user experience models. For example, for enterprise voice teleconferencing we used the E-model for estimating user experience under a specific selection of hosting and network SLAs and a specific amount of resources (virtual machines) to deploy.

QoE and collaborative projects. We then considered QoE-aware content delivery, targeting in particular an environment where web and multimedia content is disseminated by over-the-top (OTT) providers, but assuming a level of cooperation between the content provider and the ISP (a trend which has started to become commonplace) [46]. We built on the outcome of our prior work [0], where we designed and implemented a network load estimation methodology and tool which operates by observing the delay behavior of the Precision Time Protocol (PTP) for network clock synchronization. After quantitatively establishing the link between network load and user experience, we proposed an architecture for OTT content delivery where user

[0] P.A. Frangoudis, A. Ksentini, Y. Hadjadj-Aoul, and G. Boime, "PTPv2-based network load estimation," Proc. IEEE ISPCS 2013. (This work was carried out in the context of the FUI project IPChronos, see 6.10 .)

requests are redirected to the data centers expected to offer optimal QoE, taking into account, among others, information about network load in the media path offered by our load estimation service (LES) in real time. In the same context, we developed a demonstrator where the LES is integrated as an additional network probe with the QoE monitoring architecture developed in the Celtic QuEEN project (see 7.2.1.1). Using a simple video QoE model which takes into account network load and video information (quality/resolution, bitrate), we implemented [0] an adaptation scheme for DASH video delivery which switches among video qualities based on QoE estimates received by the QuEEN software agent.

QoE and PTPv2. In [46], we make the case for an alternative use of the PTPv2 protocol: Adopting a learning approach, we observe its delay behavior during the protocol message exchange, derive models of its dependence on network load and build a real-time load estimation service. Then, as an application scenario of this service, we turn our attention to the provision of Over-the-Top (OTT) services. In such an environment, and assuming a level of cooperation between the ISP and the OTT provider, we demonstrate how our service can be used for estimating the QoE for web applications. To this end, we establish quantitatively the link between network load and user experience using a state-of-the-art web QoE monitoring framework, and show how our PTPv2-based load estimation scheme can integrated in an OTT service architecture and be utilized for load-aware, QoE-optimized content delivery decisions.

QoE and reneging. We consider in [45] an important Quality of Experience (QoE) indicator in mobile networks that is reneging of users due to impatience. We specifically consider a cell under heavy load conditions and compute the reneging probability by using a fluid limit analysis. 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 for reneging. We specifically propose several flavors of this pricing 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 the communications.

QoE-aware OLSR for Video Streaming over Wireless Multihop Networks. Multi-hop environments can impact significantly ad-hoc network performance. In [57], we propose a routing algorithm based on optimized link state routing (OLSR), aimed at guaranteeing the quality of experience (QoE) of users in these types of networks. PSQA (see above in this same section) is used to estimate a mean opinion score (MOS), and then this MOS value is exploited by the source for selecting the appropriate path in the network. Moreover, an event- triggered based on the MOS value is used to provide more relevant information in selecting the best path by the source. The performance of this proposed mechanism was validated through intensive simulation under different scenarios. The results in [57] show that the proposed scheme outperforms other OLSR-based routing protocols particularly in a heavy load and high mobility scenario.

QoE-Aware Routing for Video Streaming over VANETs. In-vehicle multimedia applications are gaining interest since recent years. However, the high loss rate caused by high mobility in vehicular networks (VANETs) imposes several challenges in multimedia transmission. Moreover, in the context of multimedia, the quality of service (QoS)-based approaches assess the quality of streaming services through network-oriented metrics while the concept of quality of experience (QoE) is built upon the perception of users. In [58], a QoE-based routing protocol for video streaming over VANETs is proposed. By taking the mean opinion score (MOS) into account for path selection, good performance levels can be achieved, as shown by our simulation results.

5.3. Analytic models

Participants: Bruno Sericola, Gerardo Rubino, Raymond Marie.

New book about Dependability Theory. Dependability metrics are omnipresent in every engineering field, from simple ones through to more complex measures combining performance and dependability aspects of systems. The new book [69] written in the team, entitled “Markov Chains and Dependability Theory” and published in 2014 by Cambridge University Press (see also http://www.amazon.fr/Markov-Chains-Dependability-Theory-Gerardo/dp/1107007577/), presents the mathematical basis of the analysis of

[0] Our video adaptation scheme is implemented in the VLC open-source media player.

these metrics. The modelling context corresponds to the most used framework, Markov models. The book describes both basic results and specialised techniques. The authors first present discrete and continuous time Markov chains before focusing on dependability measures, which necessitate the study of Markov chains on a subset of states representing different user satisfaction levels for the modelled system. Topics covered include Markovian state lumping, analysis of sojourns on subset of states of Markov chains, analysis of most dependability metrics, fundamentals of performability analysis, and bounding and simulation techniques designed to evaluate dependability measures. As stated in its abstract, the book is of interest to graduate students and researchers in all areas of engineering where the concepts of lifetime, repair duration, availability, reliability and risk are important.

Fluid models. In [77] we study congestion periods in a finite fluid buffer when the 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.

Industrial Logistic Aspects. Motivated by the consideration of clauses of penalty, we worked again on the determination of the probability distributions of the delays of unavailability of systems on the operational sites. By considering in particular a given type of spare, we show the important role played by the possible waiting time of the change during the occurrence of a breakdown. In particular we verify that the cumulative probability distribution of the delay of unavailability possesses a relatively low tail diminution as well as a high square of cœfficient of variation. Upper and lower bounds are highlighted in the simplest case. These results allow to calculate the risk inferred by the use of clauses of penalty; for example, by proposing an expression of the expectation of the cost of penalty imposed by unit of time if any unavailability exceeding a certain threshold is penalized [62]. If the possible waiting time of the change is the obsession of the specialists of the maintenance, the consideration of stock shortages in supply chains is often underestimated when these events are rare events. A related work consisted in showing that a low probability of break can be associated with a high coefficient of variation can have a very significant consequence [54].

We also studied the extension of our analytical method of calculation of the operational availability of a fleet of consequent systems deployed on a site and maintained by exchanges on the site of subsets (the LRU for *line repaired unit*) in the specific case where a policy of cannibalization is implemented. We propose an approximated method which is particularly adapted to the case of systems with strong operational availability because in this case the error inferred by the approximation remains low. The developed method consists in determining the expectation of the number of blocked systems due to the lack of change, in the presence of a policy of cannibalization. This expectation is directly associated with a loss of operational availability. At present, in the presence of a policy of cannibalization, the proposed solution concerns only the systems constituted by a series of LRU but the policy of cannibalization can be applied to all or part of the types of LRU [63].

5.4. Performance Evaluation

Participants: Pierre L'Ecuyer, Bruno Sericola, Romaric Ludinard.

Network Monitoring and Fault Detection. Monitoring a system consists in collecting and analyzing relevant information provided by the monitored devices, so as to be continuously aware of the system state (situational awareness). However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. We propose in [39], [61] to address this issue by pushing the monitoring task at the edge of the network. We present a peer-to-peer based architecture, which enables nodes to adaptively and efficiently self-organize according to their "health" indicators. By exploiting both temporal and spatial

correlations that exist between a device and its vicinity, our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator. We show that the end-to-end detection process, *i.e.*, from the local detection to the management operator reporting, requires a logarithmic number of messages in the size of the network.

Robustness Analysis of Large Scale Distributed Systems. In the continuation of [81] which proposed an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, we analyze in [13], the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. When the number of Markov chains goes to infinity, we analyze the asymptotic behavior of the system for an arbitrary probability mass function governing the competition. We give conditions for the existence of the asymptotic distribution and we show how these results apply to cluster-based distributed systems when the competition between the Markov chains is handled by using a geometric distribution.

Detection of distributed deny of service attacks A Deny of Service (DoS) attack tries to progressively take down an Internet resource by flooding it 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, to inform a monitoring server that aggregates all the monitored information to accurately detect icebergs. 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 (called 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 [71].

Randomized Message-Passing Test-and-Set. In [74], we present 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, but 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.

Call centers. We develop research activities around the analysis and design of call centers, from a performance perspective. In [56], we focus on the scheduling problem (which task must be done by which worker at each period of time). We show that a Constraint Programming model can be used to solve large instances of this type of optimization work. In [21], we study call routing policies for call centers with multiple call types and multiple agent groups, focusing on the case of small and medium size centers, whose behavior may differ from those obtained in heavy-traffic regimes, and for which non-work-conserving policies can perform better. We

propose a routing policy based on weights, expressed as linear functions of the call waiting times and agent idle times, or number of idle agents, following a simulation-based optimization approach.

5.5. Network Economics

Participants: Bruno Tuffin, Pierre L'Ecuyer.

The general field of network economics, analyzing the relationships between all acts of the digital economy, has been an important subject for years in the team. The whole problem of network economics, from theory to practice, describing all issues and challenges, is described in our book [67].

Among the topics we have particularly focused on, the network neutrality debate was a major concern in 2014. In the position paper [79], Bruno Tuffin and his co-author Patrick Maillé discuss for a large audience the issues and challenges of network neutrality in response to the European parliament text voted in April 2014. A related (and often forgotten) issue, the recently raised search neutrality debate questions the ranking methods implemented by search engines: when a search is performed, do they (or should they) display the web pages ordered according to the quality-of-experience (relevance) of the content? In [22], we analyze that question in a setting when content is offered for free, content providers making revenue through advertising. For content providers, determining the amount of advertising to add to their content is a crucial strategic decision. Modeling the trade-off between the revenue per visit and the attractiveness, we investigate the interactions among competing content providers as a non-cooperative game, and consider the equilibrium situations to compare the different ranking policies. Our results indicate that when the search engine is not involved with any high-quality content provider, then it is in its best interest to implement a neutral ranking, which also maximizes user perceived quality-of-experience and favors innovation. On the other hand, if the search engine controls some high-quality content, then favoring it in its ranking and adding more advertisement yields a larger revenue. This is not necessarily at the expense of user perceived quality, but drastically reduces the advertising revenues of the other content providers, hence reducing their chances to innovate.

But while ISPs and search engines are almost the only Internet actors being pointed out as potentially non neutral, we investigate the economic impact and strategies of Content Delivery Networks (CDNs), Internet actors that reduce the capacity needs in the backbone network and improve the quality perceived by users. The growing importance of Content Delivery Network (CDN) in the value chain of content delivery raises concerns about the neutrality of these players. We consider in [52] the so-called push and pull models where the traffic is paid by the sender or the receiver, respectively, as well as the situation where the CDN is (vertically) integrated to, i.e., owned by, an Internet Service Provider (ISP). We then discuss the implication of CDNs into the network neutrality debate, another issue forgotten by researchers and regulators. We also propose in [53] a model to analyze the impact of revenue-oriented CDN management policies on the fairness of the competition among two content providers that use CDN services to deliver contents. We show that there exists a unique optimal revenue maximizing policy for a CDN actor –the dimensioning and allocation of its storage capacity– that depends on prices for service/transport/storage, and on the distribution of content popularity. Using data from the analysis of traces from two major content providers (YouTube Live and justin.tv), we remark that a CDN remains a relatively neutral actor even when one of the content providers it serves tries to monopolize the CDN storage space by implementing an aggressive policy to harm its competitors.

Finally, when a customer searches for a keyword at a classified ads website, at an online retailer, or at a search engine (SE), the platform has exponentially many choices in how to sort the output to the query. The two extremes are (a) to consider a ranking based on relevance only, which attracts more customers in the long run because of perceived quality, and (b) to consider a ranking based on the expected revenue to be generated by immediate conversions, which maximizes short-term revenue. Typically, these two objectives are not perfectly positively correlated and hence the main question is what middle ground between them should be chosen. We introduce in [78] stochastic models and propose effective solution methods that can be used to optimize the ranking considering long-term revenues. A key feature of our model is that customers are quality-sensitive and are attracted to the platform or driven away depending on the average relevance of the output. The proposed methods are of crucial importance in e-business and encompass: (i) classified ad websites which can favor paid ads by ranking them higher, (ii) online retailers which can rank products they sell according to buyers'

interests and/or the margins these products have, (iii) SEs which can position the content that they serve higher in the output page than third-party content to keep users in their platforms for longer and earn more. This goes in detriment of just offering rankings based on relevance only and is directly linked to the current search neutrality debate.

5.6. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino, Pierre L'Ecuyer.

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types. A review of Monte Carlo, Quasi-Monte Carlo and pseudo-random generation can be found in [66]. In [27], we examine some properties of the points produced by certain classes of long-period linear multiple recursive random number generators. These generators have their parameters selected in special ways to make the implementation faster. We show that as a result, the points produced by these generators have a poor lattice structure, and a poor initialization of the state can have long-lasting impact, because of the limited diffusion capacity of the recurrence.

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. In [20], we present several state-of-the-art Monte Carlo methods for simulating and estimating rare events. Among variance reduction methods, the most prominent ones for this purpose are Importance Sampling (IS) and Multilevel Splitting, also known as Subset Simulation. Some recent results on both aspects are described, motivated by theoretical issues as well as by applied problems.

A non-negligible part of our activity on the application of rare event simulation was about the evaluation of static network reliability models, with links subject to failures. Exact evaluation of static network reliability parameters belongs to the NP-hard family and Monte Carlo simulation is therefore a relevant tool to provide their estimations. In [34], we propose an adaptive parameterized method to approximate the zero-variance change of measure. The method uses two rough approximations of the unreliability function, conditional on the states of any subset of links being fixed. One of these approximations, based on mincuts, under-estimates the true unknown unreliability, whereas the other one, based on minpaths, over-estimates it. Our proposed change of measure takes a convex linear combination of the two, estimates the optimal (graph-dependent) coefficient in this combination from pilot runs, and uses the resulting conditional unreliability approximation at each step of a dynamic importance sampling algorithm. This new scheme is more general and more flexible than a previously-proposed zero-variance approximation one, based on mincuts only, and which was shown to be robust asymptotically when unreliabilities of individual links decrease toward zero. Our numerical examples show that the new scheme is often more efficient when low unreliability comes from a large number of possible paths connecting the considered nodes rather than from small failure probabilities of the links. Another paper, reference [18], focuses on another technique, known as Recursive Variance Reduction (RVR) estimator 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: 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. Continuing the analysis of existing method, we have described in [44] a necessary and sufficient condition for a well known technique called Fishman's method to verify BRE and have realized a deep analysis of the technique.

But in the literature and the previously described static network reliability models 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 [75], 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. When the network (or the number of shocks) becomes too large, PMC and turnip eventually fail, but GS works nicely for very large networks, with over 5000 shocks in our examples. [65] focuses on the application of our zero-variance approximation IS estimator to this same type of model.

Another family of models of interest in the group are the highly reliable Markovian systems, made of components subject to failures and repairs. We describe in [60] how importance sampling can be applied to efficiently estimate the average interval availability of those models. We provide a methodology for approximating the zero-variance change of measure. The method is illustrated to be very efficient on a small example, compared with standard importance sampling strategies developed in the literature.

Finally, in Quasi-Monte Carlo (QMC), the error when estimating an integral uses a deterministic sequence (instead of a random one) called a low discrepancy sequence and having the property to spread quickly over the integration domain. The estimation error is bounded by the product of a quantity depending on the discrepancy of the sequence and the variation of the integrand. But this bound is proved to be useless in practice. By combining MC and QMC methods, we can benefit from the advantages of both approaches: error estimation from MC and convergence speed from QMC. Randomized quasi-Monte Carlo (RQMC) is another class of methods for reducing the noise of simulation estimators, by sampling more evenly than with standard MC. In [37], we analyze the convergence rate of the *array-RQMC* technique, a randomized QMC method we have previously designed and devoted to the simulation of Markov chains.

In [19], we propose a method for estimating performability metrics built upon non-binary network states, determined by the hop distances between distinguished nodes. In other words, we explore the analysis of a generalization of network reliability, particularly relevant for instance in telecommunications. The estimation is performed by a Monte Carlo simulation method where the sampling space is reduced using edge sets known as d-pathsets and d-cutsets. Numerical experiments over two mesh-like networks are presented. They show significant efficiency improvements relative to the crude Monte Carlo method, in particular as link failures become rare events, which is usually the case in most real communication networks.

5.7. Wireless Networks

Participants: Osama Arouk, Btissam Er-Rahmadi, Adlen Ksentini, Yassine Hadjadj-Aoul, Quang Pham Tran Anh, Hyunhee Park, César Viho.

We continue our activities around wireless and mobile networks, where we focus particularly on 4G/5G networks as well as on a new mobile architecture known as mobile cloud.

LTE improvements. In [35], we investigated, at both the core network (EPC) and Radio Access Network (RAN), the impact of caching the shared content among users. We reviewed the different locations were data could be cached and their impacts on user QoS/QoE. In [33], we proposed several new mechanisms to handle the gateway relocation in the context of highly decentralized mobile network. To evaluate these mechanisms, we proposed an analytical model based on Markov Chains, whereby we captured the randomness of user mobility and its impact on the user QoS in terms of the probability to be connected to the optimal gateway, the drop rate, etc. In [32], we devised an agile admission control mechanism that anticipates QoS/QoE degradation and proactively defines policies for admitting UEs handing-in from the macro network to the

small cell network. It also enables IP flow mobility between small cells and macro networks. We provided an analytical model to the admission control mechanism based on Markov Decision Processes (MDP). The ultimate objective of the proposed model is to derive the optimal policy (i.e., reject or accept flows in the macro or the small cell) which maximizes users' QoE under different load scenarios (low and high load user traffic). Another work regarding small cells in LTE was proposed in [76], where we used the small cell principle to extend the mobile network coverage in emerging countries that not include a wired infrastructure. The proposed framework aims to backhaul the small cell with the less costly connection, while ensuring minimal QoS to users. In this vein, we formulated this problem through an Integer Linear Program (ILP), and solve it for small network sizes. For large instances of the network size, we proposed two new heuristics. In [30], we investigated network decentralization in conjunction with the Selective IP Traffic Offload (SIPTO) approaches to handle the mobile 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. Two of our solutions are based on Network Address Translation (NAT) named simple-NATing and twice-NATing, while a third one employs simple tunneling and a forth proposal adopts multiple Access Point Names (APNs). We also proposed methods enabling User Equipment (UEs), both in idle and active modes and while being on the move, to always have efficient Packet Data Network (PDN) connections. A qualitative analysis and a simulation study compared the different approaches with respect to cost, complexity, service continuity and network performance, demonstrating the significance of the proposed schemes for multimedia applications.

M2M. We addressed another type of traffic that appeared these last years, namely Machine to Machine (M2M) communication or Machine Type Communication (MTC). Such traffic is known by its intensity and its impact on increasing congestion in both parts of 4G networks, the Radio Access Network (RAN) and the core network. The main spirit of the proposed solutions is to proactively anticipate system overload by reducing the amount of MTC signaling messages exchanged in normal network operations. In [49] we introduced a solution that operates at the core network. We proposed that the Mobility Management Entity (MME), or an alike core network node, computes the device trigger rate that alleviates congestion, and communicates this value to the MTC-Interworking Function (MTC-IWF) element that enforces MTC traffic control, via admission control or data aggregation, on the device trigger request rate received from the different MTC servers.

As mentioned earlier, the MTC would impact not only the EPC part, but also the RAN. Group paging is currently considered as one of the most efficient mechanisms proposed to alleviate the problem of the RAN overload. In [42] we introduced a new solution to improve the performance of the current group paging method and overcome its disadvantages. The proposed solution is intended for MTC devices in connected mode state, in which they have an RRC context without being synchronized with the network. In [41] we devised a novel algorithm which estimates the network status (the number of active devices), thus better controlling the RAN access. Unlike most existing methods that consider only one channel, the proposed solution uses the statistics of all the channels in order to estimate the number of arrivals (UE and MTC devices) in each RA (Random Access) slot.

Most of the above-proposed solutions are basically incremental ones. In [31], we devised a complete new architectural vision to support MTC in mobile networks. This vision relies on the marriage of mobile networks and the cloud, specifically based on Network Function Virtualization (NFV). The proposed solution simplifies the network attach procedure for MTC devices by creating only one NFV MTC function that groups all the usual procedures. By doing so, the proposed solution is able to create and scale instances of NFV MTC functions on demand and in an elastic manner to cope with any sudden increase in traffic generated by MTC devices.

Wireless Sensor Networks (WSN). WSNs are complex systems that are mainly limited by the battery life of the nodes in order to have an adequate performance. In most cases, it is possible to have a re-deployment of new nodes in order to prolong the systems lifetime. This leads to a situation where some nodes have a low energy level while other nodes (the majority of nodes a few instants after the re-deployment procedure) have high energy levels. In these environments, it is clear that ancient nodes, those with low energy levels, have to contend for the shared medium against the majority of high energy nodes. As such, the remaining

battery life of low energy nodes would be rapidly consumed. In [64], we propose to extend the battery life of low energy nodes by means of assigning prioritized access to the shared channel to those nodes. The goal is to content among a low population of such nodes, while delaying the contention access of high energy nodes which can support higher number of collisions before energy depletion. This is done by studying two different transmission strategies referred to as "hard" and "soft" transmission probabilities. Results show that a soft transmission strategy achieves better results in terms of reduced energy consumption than both the conventional protocol or a hard transmission assignment.

The communication between nodes is the greedy factor to the energy consumption. One important mechanism to reduce the energy consumption is the in-network data aggregation. This mechanism removes repeated and unnecessary data readings and thus cuts on the energy used in communications. In [14] we reviewed the state of art on this topic. Then, we proposed a classification of the available solutions according to the way the aggregation is done. In [15], we addressed the reliable minimum data aggregation scheduling problem in wireless sensor networks under multi-channel frequency use. The proposed solution ensures the collection reliability and reduces the latency in disseminating aggregated data to the base-station over multi-frequency radio links. Another mechanism to improve energy efficiency is to optimize link scheduling when using TDMA-based techniques and data fragmentation when using slotted CSMA/CA access methods. In this line, we proposed a protocol, named DLSP, with the objective of achieving both low energy consumption and low latency in Wireless Sensor Networks. DLSP takes advantage of the spatial reuse of interference-free time slots by means of conflicts graphs. Unlike the previous studies that often consider saturated nodes, we propose to relax the saturation assumption in order to maintain good performance when some of the nodes have no data to send citemouloud:hal-01101396. In [55], we noticed that the standardized slotted CSMA/CA may lead to a wastage of the bandwidth utilization and an additional transmission delay. This drawback is mainly caused by Deferred Transmission in the CSMA/CA algorithm at the end of the superframe, when there is not sufficient time to complete the frame transmission. Thus, we proposed to fragment a data frame into a short frame and attempt its transmission in the current frame and transmit the remaining frame in the next superframe. The data fragmentation mechanism was modeled using a Markov chain. A non-saturated traffic and acknowledgement transmission are considered in our analysis.

High data rate WiFi networks. The IEEE 802.11ac Task Group (TGac) is actively working on an amendment that allows WLAN to reach a maximum aggregate network throughput up to 7 Gbps on bands below 6 GHz. In particular, the standard envisions a maximum Medium Access Control (MAC) throughput of at least 500 Mbps for a single user, and at least 1 Gbps in case of multiple users. In [36] we proposed an analysis of the IEEE 802.11ac TXOP Sharing mechanism, which was recently introduced by the 802.11ac group, by providing a Markov chain-based model. Based on the proposed Markov chain, we provided an analytical model of the achievable throughput for each AC. Accordingly, we can analyze the impact of the TXOP Sharing on the throughput of each AC, hence highlighting the improvement achieved in terms of bandwidth utilization and channel access fairness among the different ACs.

Mobile cloud. One of the 5G-architecture visions considers the usage of clouds to build mobile networks and help in decentralizing mobile networks on demand, elastically, and in the most cost-efficient way. This concept of carrier cloud becomes of vital importance knowing that several cloud providers are distributing their cloud/network, globally deploying more regional data centers, to meet their ever-increasing business demands. As an important enabler of the carrier cloud concept, network function virtualization (NFV) is gaining great momentum among industries. NFV aims for decoupling the software part from the hardware part of a carrier network node, traditionally referring to a dedicated hardware, single service and single-tenant box, that is using virtual hardware abstraction. Network functions become thus a mere code, runnable on a particular, preferably any, operating system and on top of a dedicated hardware platform. The ultimate objective is to run network functions as software in standard virtual machines (VMs) on top of a virtualization platform in a general-purpose multi-service multi-tenant node (e.g., Carrier Grade Blade Server) put into the cloud. In [26], we presented a LISP-based implementation of the Follow Me Cloud (FMC) concept, whereby mobile services hosted in federated clouds follow mobile users as they move and according to their needs. This implementation clearly demonstrates the feasibility of the FMC concept. On the other hand, service migration in FMC may be an expensive operation given the incurred cost in terms of signaling messages and

data transferred between DCs. Indeed, decision on service migration defines therefore a tradeoff between cost and user perceived quality. In [48] we addressed this tradeoff by modeling the service migration procedure using a Markov Decision Process (MDP). The aim was to formulate a decision policy that determines whether to migrate a service or not when the concerned User Equipment (UE) is at a certain distance from the source DC.

In order to meet the general needs of mobile operators, efficient mobile cloud must give high importance to the placement/instantiation of mobile network functions (such as data anchor gateways) in the federated cloud. In [43] we argued the need of using service/application type and requirements as metrics for efficiently: (i) create virtual instance of the Packet Data Network Gateway (PDN-GW); (ii) select the virtual PDN-GW for UEs with specific application type. After modeling this procedure though a nonlinear Optimization Problem (OP) and proving it as a NP-hard problem, we proposed three solutions to solve this issue.

Wireless Local Area Networks. User-centric networking has emerged as a disruptive new communication paradigm. We particularly focused on its expressions in *wireless* networking and the challenges it brings about [23]. In this context, by means of testbed experiments and simple analytic models, we quantified the upper bounds on VoIP capacity of a purely user-centric secure VoIP communications scheme that we designed, identifying the major quality degradation factors. Our results have shown that typical user Wi-Fi equipment can sustain a satisfactory number of concurrent secure VoIP sessions with acceptable QoE and, at the same time, protection from malicious user activity can be offered to access providers, while a level of roaming privacy can be guaranteed [24]. We then studied the role of users in wireless network management tasks. In particular, we proposed a scheme where monitoring the topology of Wi-Fi deployments is crowdsourced to roaming users, who submit reports on wireless coverage in their vicinity [25]. Topology information can then be used as input to reconfiguration mechanisms, such as channel assignment schemes. Users cannot be assumed trustworthy, though. They can engage in fraudulent reporting, which, unless specific countermeasures are in place, can severely impact one's view of the network topology. To this end, we designed and implemented an architecture for accurate Wi-Fi topology discovery, devising a reputation-based mechanism to tackle realistic and simple to implement attacks. We have shown analytically and via simulation that, even in the presence of large numbers of attackers, our user-centric scheme significantly outperforms pure infrastructure-based approaches, where monitoring is carried out only by trusted Access Points.

In another line of research, we focused on efficiently integrating wireless users in an Information-Centric Network (ICN) architecture. In ICN, multicast content delivery is the norm. At the same time, wireless multicast is problematic. To address this issue, we took advantage of the content awareness inherent in ICN and proposed a relay-based approach for local wireless multicasting: ICN information *scoping* mechanisms assist in expressing content semantics and, in turn, encoding the heterogeneous performance requirements of different content/application types. Under this premise, we proposed a multiobjective optimization approach for relay selection and multicast transmission rate assignment which allows to optimize for reliability, delivery time, or energy cost on a per content basis [47].

Energy saving. Another part of our activities in wireless network are related to energy saving. Indeed, one of the biggest problem today in the wireless world is that wireless devices are battery-driven, which reduce their operating lifetime. The experimental measurements we have achieved in [16] and [17] revealed that operating system overhead causes a drop in performance and energy consumption properties as compared to the GPP in case of certain low video qualities. We propose, thus, a new approach for energy-aware processor switching (GPP or DSP) which takes into consideration video quality. We show the pertinence of our solution in the context of adaptive video decoding and implement it on an embedded Linux operating system.

Adaptive Beam Scheduling for Scalable Video Multicast in Wireless Networks. Design of efficient multicast for a scalable video coding (SVC) streaming combined with directional beamforming is a challenging issue. In [29], we propose a QoE-aware directional beam scheduling (QBS) scheme which optimizes overall quality of experience (QoE) for multirate multicast of SVC, with beamforming in wireless networks. We optimally schedule different SVC layers to different beams and rate modulations. We provide a mixed integer linear programing (MILP) formulation of the problem, and then propose a heuristic algorithm. Extensive sim-

ulation results demonstrate that QBS can increase the overall QoE and can satisfy a minimum expected QoE for all users.

5.8. Future networks and architectures

Participants: Damien Le Quéré, Adlen Ksentini, Yassine Hadjadj-Aoul, Jean-Michel Sanner.

LOCARN. LOCARN (i.e. Low Opex & Capex Architecture for Resilient Networks) is a flat, dynamic and very simple packet architecture that focuses on plug-and-play guidance to provide flexibility and resiliency on the transport of client data traffics. To that end, the counterpart of the solution is a significant overhead due to the generation of control plane packets. In [50], we proved that in typical meshed operators transport networks applications, (i.e. infrastructures having high data-rates and high resiliency requirements), the LOCARN overhead is acceptable up to thousands of communications. In [51], we introduced two proposals that permit to increase the amount of simultaneous communications while maintaining the good properties of the initial design.

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 an automated hierarchical controller-based architecture handling the whole control chain.

DYOGENE Project-Team

6. New Results

6.1. Highlights of the Year

- F. Baccelli received 2014 IEEE Communications Society Stephen O. Rice Prize in the Field of Communications Theory:

  http://www.comsoc.org/about/memberprograms/comsoc-awards/rice.
- F. Baccelli received 2014 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems:

  http://www.comsoc.org/about/memberprograms/comsoc-awards/abraham.
- F. Baccelli received ACM Sigmetrics Achievement Award 2014:

  http://www.sigmetrics.org/achievementaward-2014.shtml.
- F. Simatos received 2014 ACM SIGMETRICS Rising Star Researcher Award:

  http://www.sigmetrics.org/risingstar-2014.shtml.
- P. Brémaud published a book "Fourier Analysis and Stochastic Processes". Series: Universitext. Springer, Sept. 2014 - 385 pages.
- PhD student C. Rovetta received best tool paper award at Valuetools 2014 for the paper [18].

6.2. On Spatial Point Processes with Uniform Births and Deaths by Random Connection

With I. Norros (VTT Finland) and F. Mathieu (Bell Labs France), F. Baccelli has continued the line of thought on the geometry of Peer-to-Peer systems that was initiated in their Infocom 13 paper. This type of dynamics leads to a class of spatial birth and death process of the Euclidean space where the birth rate is constant and the death rate of a given point is the shot noise created at its location by the other points of the current configuration for some response function f. An equivalent view point is that each pair of points of the configuration establishes a random connection at an exponential time determined by f, which results in the death of one of the two points. The research concentrated on space-motion invariant processes of this type. Under some natural conditions on f, one can construct the unique time-stationary regime of this class of point processes by a coupling argument. The birth and death structure can then be used to establish a hierarchy of balance integral relations between the factorial moment measures. One can also show that the time-stationary point process exhibits a certain kind of repulsion between its points that is called f-repulsion.

These results were published in [29].

6.3. A Stochastic Geometry Framework for Analyzing Pairwise-Cooperative Cellular Networks

With A. Giovanidis, IMT, F. Baccelli has studied a cooperation model where the positions of base stations follow a Poisson point process distribution and where Voronoi cells define the planar areas associated with them. For the service of each user, either one or two base stations are involved. If two, these cooperate by exchange of user data and reduced channel information (channel phase, second neighbour interference) with conferencing over some backhaul link. The total user transmission power is split between them and a common message is encoded, which is coherently transmitted by the stations. The decision for a user to choose service with or without cooperation is directed by a family of geometric policies. The suggested policies further control the shape of coverage contours in favor of cell-edge areas. Analytic expressions based on stochastic geometry are derived for the coverage probability in the network. Their numerical evaluation shows benefits from cooperation, which are enhanced when Dirty Paper Coding is applied to eliminate the second neighbour interference.

These results were published in [7].

6.4. Analysis of a Proportionally Fair and Locally Adaptive spatial Aloha in Poisson Networks

With C. Singh (IIT), F. Baccelli and B. Blaszczyszyn worked on combining adaptive protocol design, utility maximization and stochastic geometry. The focus was on a spatial adaptation of Aloha within the framework of ad hoc networks. Quasi-static networks are considered, in which mobiles learn the local topology and incorporate this information to adapt their medium access probability (MAP) selection to their local environment. The cases where nodes cooperate in a distributed way to maximize the global throughput or to achieve either proportional fair or max-min fair medium access were considered. The proportionally fair sharing case leads to closed-form performance expressions in two extreme cases: (1) the case without topology information, where the analysis boils down to a parametric optimization problem leveraging stochastic geometry; (2) the case with full network topology information, which was recently solved using shot-noise techniques. It was shown that there exists a continuum of adaptive controls between these two extremes, based on local stopping sets, which can also be analyzed in closed form. These control schemes are implementable, in contrast to the full information case which is not. As local information increases, the performance levels of these schemes are shown to get arbitrarily close to those of the full information scheme. The analytical results are combined with discrete event simulation to provide a detailed evaluation of the performance of this class of medium access controls.

These results were published in [16].

6.5. Quality of Real-Time Streaming in Wireless Cellular Networks - Stochastic Modeling and Analysis

We present a new stochastic service model with capacity sharing and interruptions, appropriate for the evaluation of the quality of real-time streaming (e.g. mobile TV) in wireless cellular networks [2]. It takes into account multi-class Markovian process of call arrivals (to capture different radio channel conditions, requested streaming bit-rates and call-durations) and allows for a general resource allocation policy saying which users are temporarily denied the requested fixed streaming bit-rates (put in outage) due to resource constraints. We develop general expressions for the performance characteristics of this model, including the mean outage duration and the mean number of outage incidents for a typical user of a given class, involving only the steady-state of the traffic demand. We propose also a natural class of least-effort-served-first resource allocation policies, which cope with optimality and fairness issues known in wireless networks, and whose performance metric! s can be easily calculated using Fourier analysis of Poisson variables. We specify and use our model to analyze the quality of real time streaming in 3GPP Long Term Evolution (LTE) cellular networks. Our results can be used for the dimensioning of these networks.

6.6. On Comparison of Clustering Properties of Point Processes

In [3], we propose a new comparison tool for spatial homogeneity of point processes, based on the joint examination of void probabilities and factorial moment measures. We prove that determinantal and permanental processes, as well as, more generally, negatively and positively associated point processes are comparable in this sense to the Poisson point process of the same mean measure. We provide some motivating results on percolation and coverage processes, and preview further ones on other stochastic geometric models, such as minimal spanning forests, Lilypond growth models, and random simplicial complexes, showing that the new tool is relevant for a systemic approach to the study of macroscopic properties of non-Poisson point processes. This new comparison is also implied by the directionally convex ordering of point processes, which has already been shown to be relevant to the comparison of the spatial homogeneity of point processes. For this latter ordering, using a notion of lattice perturbation, we provide a large monotone spectrum of comparable point processes, ranging from periodic grids to Cox processes, and encompassing Poisson point processes as well. They are intended to serve as a platform for further theoretical and numerical studies of clustering, as

well as simple models of random point patterns to be used in applications where neither complete regularity nor the total independence property are realistic assumptions.

6.7. SINR in Wireless Networks and the Two-Parameter Poisson-Dirichlet Process

Stochastic geometry models of wireless networks based on Poisson point processes are increasingly being developed with a focus on studying various signal-to-interference-plus-noise ratio (SINR) values. In [9], we show that the SINR values experienced by a typical user with respect to different base stations of a Poissonian cellular network are related to a specific instance of the so-called two-parameter Poisson-Dirichlet process. This process has many interesting properties as well as applications in various fields. We give examples of several results proved for this process that are of immediate or potential interest in the development of analytic tools for cellular networks. Some of them simplify or are akin to certain results that are being developed in the network literature. By doing this we hope to motivate further research and use of Poisson-Dirichlet processes in this new setting.

6.8. How User Throughput Depends on the Traffic Demand in Large Cellular Networks

In [17], we assume a space-time Poisson process of call arrivals on the infinite plane, independently marked by data volumes and served by a cellular network modeled by an infinite ergodic point process of base stations. Each point of this point process represents the location of a base station that applies a processor sharing policy to serve users arriving in its vicinity, modeled by the Voronoi cell, possibly perturbed by some random signal propagation effects. User service rates depend on their signal-to-interference-and-noise ratios with respect to the serving station. Little's law allows to express the mean user throughput in any region of this network model as the ratio of the mean traffic demand to the steady-state mean number of users in this region. Using ergodic arguments and the Palm theoretic formalism, we define a global mean user throughput in the cellular network and prove that it is equal to the ratio of mean traffic demand to the mean number of users in the steady st! ate of the "typical cell" of the network. Here, both means account for double averaging: over time and network geometry, and can be related to the per-surface traffic demand, base-station density and the spatial distribution of the signal-to-interference-and-noise ratio. This latter accounts for network irregularities, shadowing and cell dependence via some cell-load equations. Inspired by the analysis of the typical cell, we propose also a simpler, approximate, but fully analytic approach, called the mean cell approach. The key quantity explicitly calculated in this approach is the cell load. In analogy to the load factor of the (classical) M/G/1 processor sharing queue, it characterizes the stability condition, mean number of users and the mean user throughput. We validate our approach comparing analytical and simulation results for Poisson network model to real-network measurements.

6.9. Pioneers of Influence Propagation in Social Networks

In [20], we present a diffusion model developed by enriching the generalized random graph (a.k.a. configuration model), motivated by the phenomenon of viral marketing in social networks. The main results on this model are rigorously proved in [3], and in this paper we focus on applications. Specifically, we consider random networks having Poisson and Power Law degree distributions where the nodes are assumed to have varying attitudes towards influence propagation, which we encode in the model by their transmitter degrees. We link a condition involving total degree and transmitter degree distributions to the effectiveness of a marketing campaign. This suggests a novel approach to decision-making by a firm in the context of viral marketing which does not depend on the detailed information of the network structure.

6.10. QoS and Network Performance Estimation in Heterogeneous Cellular Networks Validated by Real-Field Measurements

Mobile network operators observe a significant disparity of quality of service (QoS) and network performance metrics, such as the mean user throughput, the mean number of users and the cell load, over different network base stations. The principal reason being the fact that real networks are never perfectly hexagonal, base stations are subject to different radio conditions, and may have different engineering parameters. In [21], we propose a model that takes into account these network irregularities in a probabilistic manner, in particular assuming Poisson spatial location of base stations, lognormal shadowing and random transmission powers. Performance of base stations is modeled by spatial processor sharing queues, which are made dependent of each other via a system of load equations. In order to validate our approach, we estimate all the model parameters from the data collected in a commercial network, solve it and compare the spatial variability of the QoS and performance metrics! in the model to the real network performance metrics. Considering two scenarios: downtown of a big city and a mid-size city, we show that our model predicts well the network performance.

6.11. Clustering Comparison of Point Processes with Applications to Random Geometric Models

In [27], we review some examples, methods, and recent results involving comparison of clustering properties of point processes. Our approach is founded on some basic observations allowing us to consider void probabilities and moment measures as two complementary tools for capturing clustering phenomena in point processes. As might be expected, smaller values of these characteristics indicate less clustering. Also, various global and local functionals of random geometric models driven by point processes admit more or less explicit bounds involving void probabilities and moment measures, thus aiding the study of impact of clustering of the underlying point process. When stronger tools are needed, directional convex ordering of point processes happens to be an appropriate choice, as well as the notion of (positive or negative) association, when comparison to the Poisson point process is considered. We explain the relations between these tools and provide examples of point processes admitting them. Furthermore, we sketch some recent results obtained using the aforementioned comparison tools, regarding percolation and coverage properties of the germ-grain model, the SINR model, subgraph counts in random geometric graphs, and more generally, U-statistics of point processes. We also mention some results on Betti numbers for Čech and Vietoris-Rips random complexes generated by stationary point processes. A general observation is that many of the results derived previously for the Poisson point process generalise to some "sub-Poisson" processes, defined as those clustering less than the Poisson process in the sense of void probabilities and moment measures, negative association or dcx-ordering.

6.12. Sublinear-Time Algorithms for Monomer-Dimer Systems on Bounded Degree Graphs

For a graph G, let $Z(G,\lambda)$ be the partition function of the monomer-dimer system defined by $\sum_k m_k(G)\lambda^k$, where $m_k(G)$ is the number of matchings of size k in G. In [11], we consider graphs of bounded degree and develop a sublinear-time algorithm for estimating $\log Z(G,\lambda)$ at an arbitrary value $\lambda > 0$ within additive error ϵn with high probability. The query complexity of our algorithm does not depend on the size of G and is polynomial in $1/\epsilon$, and we also provide a lower bound quadratic in $1/\epsilon$ for this problem. This is the first analysis of a sublinear-time approximation algorithm for a $\#P$-complete problem. Our approach is based on the correlation decay of the Gibbs distribution associated with $Z(G,\lambda)$. We show that our algorithm approximates the probability for a vertex to be covered by a matching, sampled according to this Gibbs distribution, in a near-optimal sublinear time. We extend our results to approximate the average size and the entropy of such a matching within an additive error with high probability, where again the query complexity is polynomial in $1/\epsilon$ and the lower bound is quadratic in $1/\epsilon$. Our algorithms are simple to implement and of practical use when dealing with massive datasets. Our results extend to other systems where the correlation decay is known to hold as for the independent set problem up to the critical activity.

6.13. How Clustering Affects Epidemic in Random Networks

Motivated by the analysis of social networks, we study a model of random networks that has both a given degree distribution and a tunable clustering coefficient. We consider two types of growth processes on these graphs: diffusion and symmetric threshold model. The diffusion process is inspired from epidemic models. It is characterized by an infection probability, each neighbor transmitting the epidemic independently. In the symmetric threshold process, the interactions are still local but the propagation rule is governed by a threshold (that might vary among the different nodes). An interesting example of symmetric threshold process is the contagion process, which is inspired by a simple coordination game played on the network. Both types of processes have been used to model spread of new ideas, technologies, viruses or worms and results have been obtained for random graphs with no clustering. In [6], we are able to analyze the impact of clustering on the growth processes. While clustering inhibits the diffusion process, its impact for the contagion process is more subtle and depends on the connectivity of the graph: in a low connectivity regime, clustering also inhibits the contagion, while in a high connectivity regime, clustering favors the appearance of global cascades but reduces their size. For both diffusion and symmetric threshold models, we characterize conditions under which global cascades are possible and compute their size explicitly, as a function of the degree distribution and the clustering coefficient. Our results are applied to regular or power-law graphs with exponential cutoff and shed new light on the impact of clustering.

6.14. Edge Label Inference in Generalized Stochastic Block Model: From Spectral Theory to Impossibility Results

The classical setting of community detection consists of networks exhibiting a clustered structure. To more accurately model real systems we consider a class of networks (i) whose edges may carry labels and (ii) which may lack a clustered structure. Specifically we assume that nodes possess latent attributes drawn from a general compact space and edges between two nodes are randomly generated and labeled according to some unknown distribution as a function of their latent attributes. Our goal is then to infer the edge label distributions from a partially observed network. In [22], we propose a computationally efficient spectral algorithm and show it allows for asymptotically correct inference when the average node degree could be as low as logarithmic in the total number of nodes. Conversely, if the average node degree is below a specific constant threshold, we show that no algorithm can achieve better inference than guessing without using the observations. As a byproduct of our analysis, we show that our model provides a general procedure to construct random graph models with a spectrum asymptotic to a pre-specified eigenvalue distribution such as a power-law distribution.

6.15. Balanced Graph Edge Partition

Balanced edge partition has emerged as a new approach to partition an input graph data for the purpose of scaling out parallel computations, which is of interest for several modern data analytics computation platforms, including platforms for iterative computations, machine learning problems, and graph databases. This new approach stands in a stark contrast to the traditional approach of balanced vertex partition, where for given number of partitions, the problem is to minimize the number of edges cut subject to balancing the vertex cardinality of partitions.

In [19], we first characterize the expected costs of vertex and edge partitions with and without aggregation of messages, for the commonly deployed policy of placing a vertex or an edge uniformly at random to one of the partitions. We then obtain the first approximation algorithms for the balanced edge-partition problem which for the case of no aggregation matches the best known approximation ratio for the balanced vertex-partition problem, and show that this remains to hold for the case with aggregation up to factor that is equal to the maximum in-degree of a vertex. We report results of an extensive empirical evaluation on a set of real-world graphs, which quantifies the benefits of edge- vs. vertex-partition, and demonstrates efficiency of natural greedy online assignments for the balanced edge-partition problem with and with no aggregation.

6.16. Streaming, Memory-limited Algorithms for Community Detection

In [23], we consider sparse networks consisting of a finite number of non-overlapping communities, i.e. disjoint clusters, so that there is higher density within clusters than across clusters. Both the intra- and inter-cluster edge densities vanish when the size of the graph grows large, making the cluster reconstruction problem nosier and hence difficult to solve. We are interested in scenarios where the network size is very large, so that the adjacency matrix of the graph is hard to manipulate and store. The data stream model in which columns of the adjacency matrix are revealed sequentially constitutes a natural framework in this setting. For this model, we develop two novel clustering algorithms that extract the clusters asymptotically accurately. The first algorithm is *offline*, as it needs to store and keep the assignments of nodes to clusters, and requires a memory that scales linearly with the network size. The second algorithm is *online*, as it may classify a node when the corresponding column is revealed and then discard this information. This algorithm requires a memory growing sub-linearly with the network size. To construct these efficient streaming memory-limited clustering algorithms, we first address the problem of clustering with partial information, where only a small proportion of the columns of the adjacency matrix is observed and develop, for this setting, a new spectral algorithm which is of independent interest.

6.17. State Space Collapse for Critical Multistage Epidemics

We study a multistage epidemic model which generalizes the SIR model and where infected individuals go through $K > 0$ stages of the epidemic before being removed. An infected individual in stage k may infect a susceptible individual, who directly goes to stage k of the epidemic; or it may go to the next stage $k + 1$ of the epidemic. For this model, we identify the critical regime in which we establish diffusion approximations. Surprisingly, the limiting diffusion exhibits an unusual form of state space collapse which we analyze in detail.

6.18. Perfect Sampling for Closed Queueing Networks

In [4], we investigate coupling from the past (CFTP) algorithms for closed queueing networks. The stationary distribution has a product form only in a very limited number of particular cases when queue capacity is finite, and numerical algorithms are intractable due to the cardinality of the state space. Moreover, closed networks do not exhibit any monotonic property enabling efficient CFTP. We derive a bounding chain for the CFTP algorithm for closed queueing networks. This bounding chain is based on a compact representation of sets of states that enables exact sampling from the stationary distribution without considering all initial conditions in the CFTP. The coupling time of the bounding chain is almost surely finite, and numerical experiments show that it is close to the coupling time of the exact chain.

In [18], we present Clones, a Matlab toolbox for exact sampling from the stationary distribution of a closed queueing net-work with finite capacities. This toolbox is based on recent results using a compact representation of sets of states that enables exact sampling from the stationary distribu-tion without considering all initial conditions in the coupling from the past (CFTP) scheme. This representation reduces the complexity of the one-step transition in the CFTP al-gorithm to O(KM 2), where K is the number of queues and M the total number of customers; while the cardinality of the state space is exponential in the number of queues. In this paper, we focus on the algorithmic and implementation issues. We propose a new representation, that leads to one-step transition complexity of the CFTP algorithm that is in O(KM). We provide a detailed description of our matrix-based implementation. The toolbox can be downloaded at http://www.di.ens.fr/~rovetta/Clones.

6.19. Individual Risk in Mean-Field Control Models for Decentralized Control, with Application to Automated Demand Response

Flexibility of energy consumption can be harnessed for the purposes of ancillary services in a large power grid. In prior work by the authors a randomized control architecture is introduced for individual loads for this purpose. In examples it is shown that the control architecture can be designed so that control of the loads is easy at the grid level: Tracking of a balancing authority reference signal is possible, while ensuring

that the quality of service (QoS) for each load is acceptable on average. The analysis was based on a mean field limit (as the number of loads approaches infinity), combined with an LTI-system approximation of the aggregate nonlinear model. In [15], we examine in depth the issue of individual risk in these systems. The main contributions of the paper are of two kinds: Risk is modeled and quantified: (i) The average performance is not an adequate measure of success. It is found empirically that a histogram of QoS is approximately Gaussian, and consequently each load will eventually receive poor service. (ii) The variance can be estimated from a refinement of the LTI model that includes a white-noise disturbance; variance is a function of the randomized policy, as well as the power spectral density of the reference signal. Additional local control can eliminate risk: (iii) The histogram of QoS is truncated through this local control, so that strict bounds on service quality are guaranteed. (iv) This has insignificant impact on the grid-level performance, beyond a modest reduction in capacity of ancillary service.

6.20. Passive Dynamics in Mean Field Control

Mean-field models are a popular tool in a variety of fields. They provide an understanding of the impact of interactions among a large number of particles or people or other "self-interested agents", and are an increasingly popular tool in distributed control. In [14], we consider a particular randomized distributed control architecture introduced in our own recent work. In numerical results it was found that the associated mean-field model had attractive properties for purposes of control. In particular, when viewed as an input-output system, its linearization was found to be minimum phase. In this paper we take a closer look at the control model. The results are summarized as follows: (i) The Markov Decision Process framework of Todorov is extended to continuous time models, in which the "control cost" is based on relative entropy. This is the basis of the construction of a family of controlled Markovian generators. (ii) A decentralized control architecture is proposed in which each agent evolves as a controlled Markov process. A central authority broadcasts a common control signal to each agent. The central authority chooses this signal based on an aggregate scalar output of the Markovian agents. (iii) Provided the control-free system is a reversible Markov process, the following identity holds for the linearization,

$$\text{Real}(G(j\omega)) = \text{PSD}_Y(\omega) \geq 0 \qquad \omega \in \mathbb{R}\,,,$$

where the right hand side denotes the power spectral density for the output of any one of the individual (control-free) Markov processes.

6.21. Optimization of Dynamic Matching Models

The bipartite matching model was born in the work of Gale and Shapley, who proposed the stable marriage problem in the 1960s. In [36], we consider a dynamic setting, modeled as a multi-class queueing network or MDP model. The goal is to compute a policy for the matching model that is optimal in the average cost sense. Computation of an optimal policy is not possible in general, but we obtain insight by considering relaxations. The main technical result is a form of "heavy traffic" asymptotic optimality. For a parameterized family of models in which the network load approaches capacity, a variant of the MaxWeight policy is approximately optimal, with bounded regret, even though the average cost grows without bound. Numerical results demonstrate that the policies introduced in this paper typically have much lower cost when compared to polices considered in prior work.

6.22. Stochastic Bounds with a Low Rank Decomposition

In [5], we investigate how we can bound a discrete time Markov chain (DTMC) by a stochastic matrix with a low rank decomposition. We show how the complexity of the analysis for steady-state and transient distributions can be simplified when we take into account the decomposition. Finally, we show how we can obtain a monotone stochastic upper bound with a low rank decomposition.

6.23. Generalizations of Bounds on the Index of Convergence to Weighted Digraphs

Sequences of maximum-weight walks of a growing length in weighted digraphs have many applications in manufacturing and transportation systems, as they encode important performance parameters. It is well-known that they eventually enter a periodic regime if the digraph is strongly connected. The length of their transient phase depends, in general, both on the size of digraph and on the magnitude of the weights. In this paper, we show that certain bounds on the transients of unweighted digraphs, such as the bounds of Wielandt, Dulmage-Mendelsohn, Schwarz, Kim, and Gregory-Kirkland-Pullman, remain true for critical nodes in weighted digraphs.

This work was done by Thomas Nowak together with Glenn Merlet from Aix-Marseille Unversité, Hans Schneider from the University of Wisconsin at Madison, and Sergeĭ Sergeev from the University of Birmingham. It was presented at the 53th IEEE Conference on Decision and Control and appeared in the journal Discrete Applied Mathematics.

6.24. Approximate Consensus in Highly Dynamic Networks: The Role of Averaging Algorithms

In this paper, we investigate the approximate consensus problem in highly dynamic networks in which topology may change continually and unpredictably. We prove that in both synchronous and partially synchronous systems, approximate consensus is solvable if and only if the communication graph in each round has a rooted spanning tree, i.e., there is a coordinator at each time. The striking point in this result is that the coordinator is not required to be unique and can change arbitrarily from round to round. Interestingly, the class of averaging algorithms which are memoryless and require no process identities entirely captures the solvability issue of approximate consensus in that the problem is solvable if and only if it can be solved using any averaging algorithm. Concerning the time complexity of averaging algorithms, we show that approximate consensus can be achieved with precision of ε in a coordinated network model in $O(n^{n+1}\log 1/\varepsilon)$ synchronous rounds, and in $O((\Delta n)^{n\Delta+1}\log 1/\varepsilon)$ rounds when the maximum round delay for a message to be delivered is Δ. We investigate various network models in which this exponential bound in the number of nodes reduces to a polynomial bound, and we prove that a general upper bound on the time complexity of averaging algorithms has to be exponential. We apply our results to networked systems with a fixed topology and classical benign fault models, and deduce both known and new results for approximate consensus in these systems. In particular, we show that for solving approximate consensus, a complete network can tolerate up to $2n-3$ arbitrarily located link faults at every round, in contrast with the impossibility result established by Santoro and Widmayer (STACS '89) showing that exact consensus is not solvable with $n-1$ link faults per round originating from the same node.

This work was done by Thomas Nowak together with Bernadette Charron-Bost from the CNRS and Matthias Függer from Vienna University of Technology. It is currently under submission.

6.25. Towards Binary Circuit Models That Faithfully Capture Physical Solvability

In contrast to analog models, binary circuit models are high-level abstractions that play an important role in assess-ing the correctness and performance characteristics of digital circuit designs: (i) modern circuit design relies on fast digital timing simulation tools and, hence, on binary-valued circuit models that faithfully model signal propagation, even throughout a complex design, and (ii) binary circuit models provide a level of abstraction that is amenable to formal correctness proofs. A mandatory feature of any such model is the ability to trace glitches and other short pulses precisely as they occur in physical circuits, as their presence may affect a circuit's correctness and its performance characteristics. Unfortunately, it was recently proved [Függer et al., ASYNC'13] that none of the existing binary-valued circuit models proposed so far, including the two most commonly used pure and inertial delay channels and any other bounded single-history channel, is realistic

in the following sense: For the simple Short-Pulse Filtration (SPF) problem, which is related to a circuit's ability to suppress a single glitch, they showed that every bounded single-history channel either contradicts the unsolvability of SPF in bounded time or the solvability of SPF in unbounded time in physical circuits, i.e., no existing model correctly captures physical solvability with respect to glitch propagation. We propose a binary circuit model, based on so-called in-volution channels, which do not suffer from this deficiency. In sharp contrast to what is possible with all the existing models, they allow to solve the SPF problem precisely when this is possible in physical circuits. To the best of our knowledge, our involution channel model is hence the very first binary circuit model that realistically models glitch propagation, which makes it a promising candidate for developing more accurate tools for simulation and formal verification of digital circuits.

This work was done by Thomas Nowak together with Matthias Függer, Robert Najvirt, and Ulrich Schmid from Vienna University of Technolgy. It will be presented at the conference DATE 2105.

6.26. Weak CSR Expansions and Transience Bounds in Max-Plus Algebra

This paper aims to unify and extend existing techniques for deriving upper bounds on the transient of max-plus matrix powers. To this aim, we introduce the concept of weak CSR expansions: $A^t = CS^tR \oplus B^t$. We observe that most of the known bounds (implicitly) take the maximum of (i) a bound for the weak CSR expansion to hold, which does not depend on the values of the entries of the matrix but only on its pattern, and (ii) a bound for the CStR term to dominate. To improve and analyze (i), we consider various cycle replacement techniques and show that some of the known bounds for indices and exponents of digraphs apply here. We also show how to make use of various parameters of digraphs. To improve and analyze (ii), we introduce three different kinds of weak CSR expansions. As a result, we obtain a collection of bounds, in general incomparable to one another, but better than the bounds found in the literature.

This work was done by Thomas Nowak together with Glenn Merlet from Aix-Marseille Unversité and Sergeĭ Sergeev from the University of Birmingham. It appeared in the journal Linear Algebra and its Applications.

6.27. An Overview of Transience Bounds in Max-Plus Algebra

This book chapter surveys and discusses upper bounds on the length of the transient phase of max-plus linear systems and sequences of max-plus matrix powers. In particular, It explains how to extend a result by Nachtigall to yield a new approach for proving such bounds and states an asymptotic tightness result by using an example given by Hartmann and Arguelles.

This work was done by Thomas Nowak together with Bernadette Charron-Bost from the CNRS. It appeared in the book "Tropical and Idempotent Mathematics and Applications" in the AMS's book series Contemporary Mathematics.

FUN Project-Team

5. New Results

5.1. Highlights of the Year

- Opening of the 256 M3 sensor nodes of the Lille's FIT IoT Lab platform.
- We have designed a novel single-based localization method, UNS, for accurate localization of mobile devices that only needs a small aperture array unlike all previous works. UNS is currently under patenting.
- We have provided a set of recognized contributions in the area of Smart Cities, re-thinking their architecture and break vertical silos between every network and application.

5.2. Routing in FUN

Participants: Valeria Loscri, Nathalie Mitton, Riccardo Petrolo.

According to a wide range of studies, IT should become a key facilitator in establishing primary education, reducing mortality and supporting commercial initiatives in Least Developed Countries (LDCs). The main barrier to the development of IT services in these regions is not only the lack of communication facilities, but also the lack of consistent information systems, security procedures, economic and legal support, as well as political commitment. In [3], [10], we propose the vision of an infrastructure-less data platform well suited for the development of innovative IT services in LDCs. We propose a participatory approach, where each individual implements a small subset of a complete information system thanks to highly secure, portable and low-cost personal devices as well as opportunistic networking, without the need of any form of infrastructure. We review the technical challenges that are specific to this approach. Relying on such an infrastructure, wireless routing must be opportunistic and take advantages of the availability of every infrastructure point when in range. Two different approaches depending on the available devices are presented in [20] and [2]. When partial positions of nodes are available, the system can take advantage of such knowledge to enhance the routing performance. This is what has been investigated in[12] where coordinates are used in an opportunistic fashion when available.

5.3. Self-organization

Participants: Natale Guzzo, Valeria Loscri, Nathalie Mitton.

Self-organization encompasses several mechanisms. This year, the FUN research group has contributed to specific aspects; topology importance and clustering.

5.3.1. Impact of the topology

Wireless Sensor Networks (WSN) are composed of constrained devices and deployed in unattended and hostile environments. Most papers presenting solutions for WSN evaluate their work over random topologies to highlight some of their "good" performances. They rarely study these behaviors over more than one topology. Yet, the topology used can greatly impact the routing performances. [13] presents a study of the impact of the network topology on algorithm performance in WSNs and illustrate it with the geographic routing. Geographic routing relies on node coordinates to route data packets from source to destination. We measure the impact of different network topologies from realistic ones to regular and very popular ones through extensive simulation and experimentation campaigns. We show that different topologies can lead to a difference of up to 25% on delivery ratio and average route length and more than 100% on energy costs.

5.3.2. Clustering

Clustering in wireless sensor networks is an efficient way to structure and organize the network. It aims to identify a subset of nodes within the network and bind it a leader (i.e. cluster-head). This latter becomes in charge of specific additional tasks like gathering data from all nodes in its cluster and sending them by using a longer range communication to a sink or a Base Station (BS) which may be far away from the monitoring area. Many algorithms proposed in the literature compute the routing process by clustering the network and by designing new election mechanisms in which the cluster-heads are chosen taking account of the remaining energy, the communication cost and the density of nodes. However, they do not consider the connectivity to the BS, and assume that all the nodes or only few prefixed nodes are able to directly communicate with it. We believe that this assumption is not suitable for many applications of WSN and to tackle this problem we propose CESAR [14], a multi-hop and energy-efficient routing protocol for large-scale WSN which includes a new cluster-head selection mechanism aware of the battery level and the connectivity to the BS. Furthermore, our solution employs an innovative hybrid approach to combine both clustering and on-demand techniques in order to provide an adaptive behavior for different dynamic topologies. Simulation results show that our solution outperforms in terms of energy consumption and data delivery other known routing algorithms in the literature. Note that CESAR is currently the object of two pending patents.

5.4. Controlled mobility based services

Participants: Emilio Compagnone, Valeria Loscri, Karen Miranda, Nathalie Mitton, Tahiry Razafindralambo, Dimitrios Zormpas, Jean Razafimandimby Anjalalaina.

Sensors have more and more functionality in terms of capture techniques, communication capabilities, processing capabilities and energy harvesting. Another interesting feature available on sensors is mobility. The FUN research group tries to exploit the controlled mobility of sensors to solve some known issues in wireless sensors networks regarding deployment or routing but also raises some new challenges regarding coverage optimization and energy harvesting.

5.4.1. Coverage

Wireless sensors are used to gather information from a field of interest. In order to capture all the events in this field, the sensors must be properly placed. When the sensors have motion capabilities such as robots, the deployment can be optimized. The use of controlled mobility raises some new challenges and opportunities in the field of wireless sensor networks. Milan Erdelj and Karen Miranda in [33] presents the advances in context. They provide a detailed literature review regarding the techniques behind controlled mobility in order to deploy or redeploy sensors. When the wireless sensors are mobile, it is possible to optimize the capture of information regarding their time and space evolution. This allows the sensors to focus on different zones of interest depending on the evolution of the observed events. Valeria Loscri, Enrico Natalizio and Nathalie Mitton present a performance evaluation of different algorithms for zone of interest coverage in [18]. Their work particularly focuses on providing a set of distributed version of a combined particle swarm optimization and virtual forces algorithm. The proposed algorithms and their evaluation show an high reactivity to changing events and targets. Energy is an important constraint in wireless sensor networks and message exchange is a functionality that drains huge amount of energy. Dimitrios Zorbas and Christos Douligeris in [30] present a low-overhead localized algorithm for the target coverage problem in wireless sensor networks. To tackle this problem they propose two variations of a localized algorithm with low communication complexity in term of message exchange. The results show a great improvement in terms of communication cost while achieving an adequate network lifetime.

5.4.2. Connectivity and performance

Information gathered by sensors are to be processed in a remote location. The transportation from the point where the raw data is generated (the sensor) and the data processing unit (sink or other infrastructure) relies on routing techniques. Routing is a fundamental functionality of a wireless sensors network. Nicolas Gouvy, Nathalie Mitton and David Simplot-Ryl in their book chapter [34] provide a review of the routing techniques

described in the literature. They highlight the challenges, main issues and future work direction in this domain and provide some important assumption and characteristics that should be kept in mind when designing routing protocols for wireless sensor networks. When route between a source and the destination of data does not exist or cannot be established, using a mobile router is a possible solution. Christos Katsikiotis, Dimitrios Zorbas and Periklis Chatzimisios in [15] propose an algorithm that restores connectivity by the use of mobile wireless router after a routing failure. They provide a fast mechanism to heal the network and restore connectivity between the network partitions. In their solution, a mobile wireless router finds the end points that should be re-connect and place itself in the correct position to restore the connectivity. Their solution shows a fast restoration process based on the implementation done on a real robotic platform.

5.4.3. Energy suppliance

Energy is an important constraint in static wireless sensor networks and even more important when sensors are mobile. However, when sensors have motion capabilities, they can use this ability to move toward a recharging point in order to increase the network operation. Dimitrios Zorbas and Tahiry Razafindralambo in [31] use the motion capability of sensors to provide an algorithm that allow the sensor to go to a recharging point while minimize the impact of their movement on the network operation such as portioning or data gathering. They provide theoretical bounds on the realisation of such operation and evaluate the average behaviour of their algorithm based on extensive simulations. Both results show a big improvement in terms of network lifetime extension compared to the case where no replacement is performed and to the case where rerouting is considered.

5.4.4. Video-based applications

Video Surveillance and Target Detection represent key components for many organizations in terms of safety and security protocols. The value of Video Surveillance has become more sophisticated and very accurate, by leveraging specific sensors able to detect motion, heat, etc. In [17], Valeria Loscri, Michele Magno and Rosario Surace show how the nodes of a sensor network can learn which is their best position based on a certain number of WebCams that need to be "woken-up" when a suspicious event is detected. The main purpose is to reduce power consumption, especially in the case of Video Surveillance, when the most of the time the power is wasted by doing nothing. On the other hand, Target Detection, namely determining whether or not a target object exists in a video frame, has grown significantly with the recent advances in embedded computing and sensors which have opened the possibility to realize smaller and low-cost autonomous systems. In [16], Valeria Loscri, Nathalie Mitton and Emilio Compagnone show the feasibility of low-cost embedded system for detection of objects based either on the shape or on the color.

5.5. Security

Participants: Valeria Loscri, Nathalie Mitton.

Security has been always a critical issue both for the users and providers of wireless communication systems. The definition of novel paradigms and innovative communication systems, such as the Internet of Things (IoT) and the nanocommunication systems, exacerbated the criticality of security and privacy factors. These latter aspects are faced in [23] and [5]. In [23], Riahi et al. face with the security issues related to the IoT paradigm, by taking into consideration that this paradigm enable daily objects to become active participants of everyday activities. They envisage the main challenges and propose solutions to address them. In [5], Valeria Loscri et al. analyze the innovative aspects that characterize the molecular communication paradigm, by proposing innovative and revolutionary methods that take into consideration the very limited available resources (i.e. we work at molecular level and then we cannot leverage on high processing and computing capabilities) and the very high criticality of the potential applications of similar systems (e.g. in-vivo applications).

5.6. RFID

Participants: Ibrahim Amadou, Nathalie Mitton.

Due to the dedicated short range communication feature of passive radio frequency identification (RFID) and the closest proximity operation of both tags and readers in a large-scale dynamic RFID system, when nearby readers simultaneously try to communicate with tags located within their interrogation range, serious interference problems may occur. Such interferences may cause signal collisions that lead to the reading throughput barrier and degrade the system performance. Although many efforts have been done to maximize the throughput by proposing protocols such as NFRA or more recently GDRA, which is compliant with the EPCglobal and ETSI EN 302 208 standards. However, the above protocols are based on unrealistic assumptions or require additional components with more control packet and perform worse in terms of collisions and latency, etc. In [9], we explore the use of some well-known Carrier Sense Multiple Access (CSMA) backoff algorithms to improve the existing CSMA-based reader-to-reader anti-collision protocol in dense RFID networks. Moreover, the proposals are compliant with the existing standards. We conduct extensive simulations and compare their performance with the well-known state-of-the-art protocols to show their performance under various criteria. We find that the proposals improvement are highly suitable for maximizing the throughput, efficiency and for minimizing both the collisions and coverage latency in dense RFID Systems.

5.7. VANET

Participant: Nathalie Mitton.

VANET (Vehicular Networks) is an arising kind of network which features specific functionalities and requirements especially in terms of delay.

[26] analyzes the information delivery delay for roadside unit deployment in an intermittently connected vehicular network. A mathematical model is developed to describe the relationship between the average information delivery delay and the distance between two neighbor RSUs (Road Side Unit) deployed along a road. The derived 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 delay requirement of some time-critical applications, this model can be used to estimate the maximum distance allowed between two neighbor RSUs, which can provide a reference basis for the deployment of RSUs in such scenarios.

Abstract–Broadcasting is an effective routing paradigm for data dissemination in vehicular ad hoc networks (VANETs). One concern that arises with broadcasting is the broadcast storm problem, which would cause node contentions and data collisions, and thus degrade the transmission efficiency of a network. [27] proposes a Dynamic trAnsmission delaY based broadcast (DAYcast) protocol for a VANET. To alleviate the effect of the broadcast storm and improve the transmission efficiency of the network, DAYcast only allows the effective neighbors of a source vehicle to broadcast a received data packet and the selection of the effective neighbors are based on the position information on the one-hop neighbors of the source vehicle. Meanwhile, it allows each effective neighbor to wait a certain transmission delay before it broadcasts a received packet. The transmission delay of an effective neighbor depends on the distance between the neighbor and the source vehicle, and the number of effective neighbors of the source vehicle. Simulation results show that DAYcast can effectively improve the network performance in terms of network reachability and the successful delivery ratio as compared with existing weighted p-persistence broadcasting (WPB) and slotted 1-persistence broadcasting (SPB).

5.8. Smart cities architecture

Participants: Valeria Loscri, Nathalie Mitton, Riccardo Petrolo, Nicola Zema.

Smart City represents one of the most promising and prominent Internet of Things (IoT) applications. In the last few years, indeed, smart city concept has played an important role in academic and industry fields, with the development and deployment of various middleware platforms. However, this expansion has followed distinct approaches creating, therefore, a fragmented scenario, in which different IoT ecosystems are not able to communicate between them. To fill this gap, there is a need to re-visit the smart city IoT semantic and offer a global common approach. In order to allow cities to share data across systems and coordinate processes across domains, it is essential to break these silos. A way to achieve the purpose is sensor virtualization, discovery and data restitution. This last year, the FUN team has lead several investigations in this direction.

We have looked at the heterogeneity of devices and network technologies under a different perspective by not perceiving it as a limitation but as a potential to increase the connectivity in a smart city [1]. We propose a new generation of network nodes, called stem nodes, based on the innovative idea of 'stemness', which pushes forward the well-known self-configuration and self-management concepts towards the idea of node mutation and evolution. We also deployed prototypes that demonstrate the stem-node architecture and basic operations in different hardware platforms of common communication devices (an Alix-based router, a laptop and a smartphone)

In [7], we illustrate semantic interoperability solutions for IoT systems. Based on these solutions, we describe how the FP7 VITAL project aims to bridge numerous silo IoT deployments in smart cities through repurposing and reusing sensors and data streams across multiple applications without carelessly compromising citizens' security and privacy. This approach holds the promise of increasing the Return-On-Investment (ROI), which is associated with the usually costly smart city infrastructures, through expanding the number and scope of potential applications.

To this purpose, [21] browses the semantic annotation of the sensors in the cloud, and innovative services can be implemented and considered by bridging Clouds and Internet of Things. Things-like semantic will be considered to perform the aggregation of heterogeneous resources by defining the Clouds of Things paradigm. We survey the smart city vision, providing information on the main requirements and highlighting the benefits of integrating different IoT ecosystems within the cloud under this new CoT vision. This paper also discusses relevant challenges in this research area.

Going further, we also presented [22] a first implementation of this federation: a federation of FIT IoT-LAB within OpenIoT. OpenIoT is a middleware that enables the collection of data streams from multiple heterogeneous geographically dispersed data sources, as well as their semantic unification and streaming with a cloud infrastructure. Future Internet of Things IoT-LAB (FIT IoT-LAB) provides a very large scale infrastructure facility suitable for testing small wireless sensor devices and heterogeneous communicating objects. The integration proposed represents a way to reduce the gap existing in the IoT fragmentation, and, moreover, allows users to develop smart city applications by interacting directly with sensors at different layers. We illustrate it trough a basic temperature monitoring application to show its efficiency.

So, once all city network and infrastructure are set at the same level thanks to the above mentioned approaches, they can go further and offer additional services. An example of them is navigation[11] as also described in "Localization" section. Another example is to make use of the urban bikes [19]. Indeed, besides the growing enthusiast provoked by bicycles in smart and green cities and the benefit for health they bring, there still exists some reluctance in using bikes because of safety, road state, weather, etc. To counter-balance these feelings, there is a need to better understand bicycle users habits, path, road utilization rate in order to improve the bicycle path quality. In this perspective,in this paper, we propose to deploy a set of mobile sensors on bicycles to gather this different data and to exploit them to make the bike easier and make people want to ride bicycles more often. Such a network will also be useful for several entities like city authorities for road maintenance and deployment, doctors and environment authorities, etc. Based on such a framework,we propose a first basis model that helps to dimension the network infrastructure and the kind of data to be real time gathered from bikes. More specifically, we present a theoretical model that computes the quantity of data a bike will be able to send along a travel and the quantity of data a base station should be able to absorb.We have based our study on real data to provide first numerical results and be able to draw some preliminary conclusions and open new research directions.

5.9. Localization

Participants: Ibrahim Amadou, Roudy Dagher, Nathalie Mitton, Roberto Quilez, Nicola Zema.

Navigate in or based on a wireless sensor network present many advantages but it is still an open issue. We have focused on two particular cases in which navigation or WSN-based localization is needed [32]. The former aspect considers that sensors need to be visited on-demand by a mobile sink to offload data. This mobile sink thus needs to locate the data source. The second aspect feature a mobile entity that is needed to be localized.

In a event-based WSN, where is necessary a prompt response in terms of data processing and offloading, a set of mobile flying sinks could be a good option for the role of autonomous data collectors. For those reasons in [28], we propose a distributed algorithm to independently and autonomously drive a mobile sink through the nodes of a WSN and we show its preferability over more classical routing approaches especially in the presence of a localized generation of large amount of information. Our result shows that, in the case of fairly complete coverage of the area where the nodes lie, it is possible to promptly notify a mobile sink about the presence of data to offload, drive it to the interested area and achieve interesting performances. [29] enhanced the previous approach by relaxing some GPS-use assumptions. We show that, under fairly common circumstances, it is possible to set the trajectory of the mobile sink and fulfill the offloading requests without the needs of additional equipment installed on nodes. We show how our system is preferable over more classical routing solutions especially in the presence of localized generation of large amounts of information.

[11] proposes Ubiquitous Navigation System (UNS), a WSN-based navigation system, which takes benefit from a WSN mesh deployment to provide a local navigation service. The positioning part of the system uses Angle of Arrival (AoA) measurements to estimate the vehicle position on the map. Based on a realistic network scenario, extracted from a city map using Google Maps, we study the performance of Triangulation using AoA in a smart urban environment that exhibits topology related constraints. Simulations results show that such constraints lead to particular spatial distribution of the anchor nodes that affects both positioning accuracy and beacon packets reception rate. We also propose and evaluate the use of the network communication range as a technique to mitigate the effect of geometric dilution of precision (GDOP). The simulation results show that this technique successfully detected GDOP-affected positions and thus significantly enhanced the positioning accuracy. One of the biggest strengths of UNS is that it relies on a single anchor unlike literature approaches. The different underlying studies are detailed in [38] in which we study the ambiguity of source localization using signal processing of large aperture antenna arrays under spherical wave propagation. This novel localization approach has been recently proposed, providing an estimate of the source position by means of two methods: geometrical and analytical. The former finds the source position as the estimate of circular loci, the latter as a solution of a linear system of equations. Although this method is proved to work for a general array geometry, we show that it suffers from ambiguities for a particular class of array geometries. Namely, in 2D, we prove that when the array geometry is linear or circular, there exist two possible solutions where only one corresponds to the actual position of the source. We also prove a relation of symmetry between the solutions with respect to the array geometry. This relation is very useful to assist the disambiguation process for discounting one of the estimates. By extension to 3D, planar (resp. spherical) arrays exhibit the same behavior i.e they provide two symmetrical estimates of the source position when the latter is not on the array plane (resp. sphere).
Note that UNS is currently a pending patent.

GANG Project-Team

5. New Results

5.1. Highlights of the Year

Pierre Fraigniaud has received the Prize for Innovation in Distributed Computing 2014.

5.2. Graph and Combinatorial Algorithms

5.2.1. Collision-Free Network Exploration

In the collision-free exploration model considered in [16], a set of mobile agents is placed at different nodes of a n-node network. The agents synchronously move along the network edges in a collision-free way, i.e., in no round may two agents occupy the same node. In each round, an agent may choose to stay at its currently occupied node or to move to one of its neighbors. An agent has no knowledge of the number and initial positions of other agents. We are looking for the shortest possible time required to complete the collision-free *network exploration*, i.e., to reach a configuration in which each agent is guaranteed to have visited all network nodes and has returned to its starting location.

In this work, we first considered the scenario when each mobile agent knows the map of the network, as well as its own initial position. We established a connection between the number of rounds required for collision-free exploration and the degree of the minimum-degree spanning tree of the graph. We provided tight (up to a constant factor) lower and upper bounds on the collision-free exploration time in general graphs, and the exact value of this parameter for trees. For our second scenario, in which the network is unknown to the agents, we proposed collision-free exploration strategies running in $O(n^2)$ rounds for tree networks and in $O(n^5 \log n)$ rounds for general networks.

5.2.2. Properties of Graph Search Procedures

In [4], we study the last vertex discovered by a graph search such as BFS or DFS. End-vertices of a given graph search may have some nice properties (as for example it is well known that the last vertex of Lexicographic Breadth First Search (LBFS) in a chordal graph is simplicial). Therefore it is interesting to consider if these vertices can be recognized in polynomial time or not. A graph search is a mechanism for systematically visiting the vertices of a graph. At each step of a graph search, the key point is the choice of the next vertex to be explored. Graph searches only differ by this selection mechanism during which a tie-break rule is used. In this paper we study how the choice of the tie-break rule can determine the complexity of the end-vertex problem for BFS or DFS. In particular we prove a counter-intuitive NP-completeness result for Breadth First Search, answering a question of D.G. Corneil, E. Köhler and J-M Lanlignel.

5.2.3. 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. The aim of [3] is twofold—to put order in the multitude of conjectures that relate to this concept (some first presented here), and to prove partial results on one of the central conjectures settled by Ryser, Brualdi and Stein.

5.2.4. Common Intervals and Application to Genome Comparison

In [6], we show how to identify generalized common and conserved nested intervals. This is a bio-informatics papers, explaining how to compute more relaxed variants of common or of conserved intervals of two permutations, which has applications in genome comparison. It also presents some properties of the family of intervals, useful for storing them.

5.2.5. Graph Decomposition

In [10], we present a general framework for computing a large family of graph decomposition, the H-join. It generalizes some well know tools like modular decomposition or split decomposition. The paper explains how to compute it in polynomial time. A new canonical decomposition for sesquiprime graphs is also presented.

5.2.6. Combinatorial Optimization

Normal cone and subdifferential have been generalized through various continuous functions; in [8], we focus on a non separable Q-subdifferential version. Necessary and sufficient optimality conditions for unconstrained nonconvex problems are revisited accordingly. For inequality constrained problems, Q-subdifferential and the lagrangian multipliers, enhanced as continuous functions instead of scalars, allow us to derive new necessary and sufficient optimality conditions. In the same way, the Legendre-Fenchel conjugate is generalized into Q-conjugate and global optimality conditions are derived by Q-conjugate as well, leading to a tighter inequality.

5.3. Distributed Computing

5.3.1. Rendezvous

5.3.1.1. Rendezvous of Anonymous Agents in Trees

In [5], we study the so-called *rendezvous problem* in the mobile agent setting in graph environments. In the studied model, two identical (anonymous) mobile agents start from arbitrary nodes of an unknown tree and have to meet at some node. Agents move in synchronous rounds: in each round an agent can either stay at the current node or move to one of its neighbors. We consider deterministic algorithms for this rendezvous task. The main result of our research is a tight trade-off between the optimal time of completing rendezvous and the size of memory of the agents. For agents with k memory bits, we show that optimal rendezvous time is $\Theta(n + n^2/k)$ in n-node trees. More precisely, if $k \geq c \log n$, for some constant c, we design agents accomplishing rendezvous in arbitrary trees of size n (unknown to the agents) in time $O(n + n^2/k)$, starting with arbitrary delay. We also show that no pair of agents can accomplish rendezvous in time $o(n + n^2/k)$, even in the class of lines of known length and even with simultaneous start. Finally, we prove that at least logarithmic memory is necessary for rendezvous, even for agents starting simultaneously in a n-node line.

5.3.1.2. Rendezvous of Distance-Aware Mobile Agents in Unknown Graphs

In [17], we study the problem of rendezvous of two mobile agents starting at distinct locations in an unknown graph. The agents have distinct labels and walk in synchronous steps. However, the graph is unlabeled and the agents have no means of marking the nodes of the graph and cannot communicate with or see each other until they meet at a node. When the graph is very large, we would like the time to rendezvous to be independent of the graph size and to depend only on the initial distance between the agents and some local parameters such as the degree of the vertices, and the size of the agent's label. It is well known that even for simple graphs of degree Δ, the rendezvous time can be exponential in Δ in the worst case. In this study, we introduce a new version of the rendezvous problem where the agents are equipped with a device that measures its distance to the other agent after every step. We show that these *distance-aware* agents are able to rendezvous in any unknown graph, in time polynomial in all the local parameters such the degree of the nodes, the initial distance D and the size of the smaller of the two agent labels $l = \min(l_1, l_2)$. Our algorithm has a time complexity of $O(\Delta(D + \log l))$ and we show an almost matching lower bound of $\Omega(\Delta(D + \log l / \log \Delta))$ on the time complexity of any rendezvous algorithm in our scenario. Further, this lower bound extends existing lower bounds for the general rendezvous problem without distance awareness.

5.3.1.3. Rendezvous of Heterogeneous Mobile Agents in Edge-Weighted Networks

In [22], we study the deterministic rendezvous problem in which a pair of heterogeneous agents, differing in the time required to traverse particular edges of the graph, need to meet on an edge or node of the graph. Each of the agents knows the complete topology of the undirected graph and the initial positions of both of the 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. In this scenario, we study the time required by the agents

to meet, compared to the time T_{OPT} in the offline scenario in which the agents have complete knowledge of each others capabilities. When no additional assumptions are made, we show that rendezvous can be achieved after time $O(nT_{\mathrm{OPT}})$ in a n-node graph, and that this time is essentially the best possible in some cases. However, the rendezvous time can be reduced to $\Theta(T_{\mathrm{OPT}})$ 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 $\Theta(T_{\mathrm{OPT}})$.

5.3.1.4. Rendezvous with Different Speeds

In [32] we introduce the study of the rendezvous problem in the context of agents having different speeds, and present tight and almost tight bounds for this problem, restricted to a ring topology.

5.3.2. Fair Synchronization

A non-blocking implementation of a concurrent object is an implementation that does not prevent concurrent accesses to the internal representation of the object, while guaranteeing the deadlock-freedom progress condition without using locks. Considering a failure free context, G. Taubenfeld has introduced (DISC 2013) a simple modular approach, captured under a new problem called the *fair synchronization* problem, to transform a non-blocking implementation into a starvation-free implementation satisfying a strong fairness requirement.

This approach is illustrated in [19] with the implementation of a concurrent stack. The spirit of the paper is mainly pedagogical. Its aim is not to introduce new concepts or algorithms, but to show that a powerful, simple, and modular transformation can provide concurrent objects with strong fairness properties.

In [20], we extend this approach in several directions. It first generalizes the fair synchronization problem to read/write asynchronous systems where any number of processes may crash. Then, it introduces a new failure detector and uses it to solve the fair synchronization problem when processes may crash. This failure detector, denoted QP (Quasi Perfect), is very close to, but strictly weaker than, the perfect failure detector. Last but not least, the paper shows that the proposed failure detector QP is optimal in the sense that the information on failures it provides to the processes can be extracted from any algorithm solving the fair synchronization problem in the presence of any number of process crash failures.

5.3.3. Wait Free with Advice

In [7], we motivate and propose a new way of thinking about failure detectors which allows us to define, quite surprisingly, what it means to solve a distributed task *wait-freeusing 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* 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.

5.3.4. Adaptive Register Allocation

In [18], we give an adaptive algorithm in which processes use multi-writer multi-reader registers to acquire exclusive write access to their own single-writer, multi-reader registers. It is the first such algorithm that uses a number of registers linear in the number of participating processes. Previous adaptive algorithms require at least $\Theta(n^{3/2})$ registers.

5.3.5. Leader Election

Considering the case of homonyms processes (some processes may share the same identifier) on a ring [21], we give a necessary and sufficient condition on the number of identifiers to enable leader election. We prove that if l is the number of identifiers then message-terminating election is possible if and only if l is greater than the greatest proper divisor of the ring size even if the processes do not know the ring size. If the ring size is known, we propose a process-terminating algorithm exchanging $O(n \log(n))$ messages that is optimal.

5.3.6. Concurrency and Fault-tolerance

In [15], we study the connections between self-stabilization and proof-labeling schemes. It follows from the definition of *silent* self-stabilization, and from the definition of *proof-labeling* scheme, that if there exists a silent self-stabilizing algorithm using ℓ-bit registers for solving a task T, then there exists a proof-labeling scheme for T using registers of at most ℓ bits. The first result in this paper is the converse to this statement. We show that if there exists a proof-labeling scheme for a task T, using ℓ-bit registers, then there exists a silent self-stabilizing algorithm using registers of at most $O(\ell + \log n)$ bits for solving T, where n is the number of processes in the system. Therefore, as far as memory space is concerned, the design of silent self-stabilizing algorithms essentially boils down to the design of compact proof-labeling schemes. The second result in this paper addresses time complexity. We show that, for every task T with k-bits output size in n-node networks, there exists a silent self-stabilizing algorithm solving T in $O(n)$ rounds, using registers of $O(n^2 + kn)$ bits. Therefore, as far as running time is concerned, *every* task has a silent self-stabilizing algorithm converging in a linear number of rounds.

In [27], we study the connections between, on the one hand, asynchrony and concurrency, and, on the other hand, the quality of the expected solution of a distributed algorithm. The state machine approach is a well-known technique for building distributed services requiring high performance and high availability, by replicating servers, and by coordinating client interactions with server replicas using consensus. Indulgent consensus algorithms exist for realistic eventually partially synchronous models, that never violate safety and guarantee liveness once the system becomes synchronous. Unavoidably, these algorithms may never terminate, even when no processor crashes, if the system never becomes synchronous. We propose a mechanism similar to state machine replication, called *RC-simulation*, that can always make progress, even if the system is never synchronous. Using RC-simulation, the quality of the service will adjust to the current level of asynchrony of the network — degrading when the system is very asynchronous, and improving when the system becomes more synchronous. RC-simulation generalizes the state machine approach in the following sense: when the system is asynchronous, the system behaves as if $k + 1$ threads were running concurrently, where k is a function of the asynchrony. In order to illustrate how the RC-simulation can be used, we describe a long-lived renaming implementation. By reducing the concurrency down to the asynchrony of the system, RC-simulation enables to obtain renaming quality that adapts linearly to the asynchrony.

5.3.7. Quantum Computing

In [1], we provide illustrative examples of distributed computing problems for which it is possible to design tight lower bounds for *quantum* algorithms without having to manipulate concepts from quantum mechanics, at all. As a case study, we address the following class of 2-player problems. Alice (resp., Bob) receives a boolean x (resp., y) as input, and must return a boolean a (resp., b) as output. A *game* between Alice and Bob is defined by a pair (δ, f) of boolean functions. The objective of Alice and Bob playing game (δ, f) is, for every pair (x, y) of inputs, to output values a and b, respectively, satisfying $\delta(a, b) = f(x, y)$, in *absence of any communication* between the two players, but in *presence of shared resources*. The ability of the two players to solve the game then depends on the type of resources they share. It is known that, for the so-called CHSH game, i.e., for the game $a \oplus b = x \wedge y$, the ability for the players to use entangled quantum bits (qubits) helps. We show that, apart from the CHSH game, quantum correlations do not help, in the sense that, for every game not equivalent to the CHSH game, there exists a classical protocol (using shared randomness) whose probability of success is at least as large as the one of any protocol using quantum resources. This result holds for both worst case and average case analysis. It is achieved by considering a model stronger than quantum correlations, the *non-signaling model*, which subsumes quantum mechanics, but is far easier to handle.

5.3.8. Distributed Decision and Verification

5.3.8.1. Randomization

In [12], we study the power of randomization in the context of locality by analyzing the ability to "boost" the success probability of deciding a distributed language. The main outcome of this analysis is that the distributed computing setting contrasts significantly with the sequential one as far as randomization is concerned. Indeed, we prove that in some cases, the ability to increase the success probability for deciding distributed languages is rather limited.

5.3.8.2. Model Variants

In a series of papers [14], [28], we analyze distributed decision in the context of various models for distributed computing.

In [28], we carry on the effort to bridging runtime verification with distributed computability, studying necessary conditions for monitoring failure prone asynchronous distributed systems. It has been recently proved that there are correctness properties that require a large number of opinions to be monitored, an opinion being of the form true, false, perhaps, probably true, probably no, etc. The main outcome of this paper is to show that this large number of opinions is not an artifact induced by the existence of artificial constructions. Instead, monitoring an important class of properties, requiring processes to produce at most k different values does require such a large number of opinions. Specifically, our main result is a proof that it is impossible to monitor k-set-agreement in an n-process system with fewer than $\min\{2k, n\} + 1$ opinions. We also provide an algorithm to monitor k-set-agreement with $\min\{2k, n\} + 1$ opinions, showing that the lower bound is tight.

Finally, in [14], we tackle *local distributed testing* of graph properties. This framework is well suited to contexts in which data dispersed among the nodes of a network can be collected by some central authority (like in, e.g., sensor networks). In local distributed testing, each node can provide the central authority with just a few information about what it perceives from its neighboring environment, and, based on the collected information, the central authority is aiming at deciding whether or not the network satisfies some property. We analyze in depth the prominent example of checking *cycle-freeness*, and establish tight bounds on the amount of information to be transferred by each node to the central authority for deciding cycle-freeness. In particular, we show that distributively testing cycle-freeness requires at least $\lceil \log d \rceil - 1$ bits of information per node in graphs with maximum degree d, even for connected graphs. Our proof is based on a novel version of the seminal result by Naor and Stockmeyer (1995) enabling to reduce the study of certain kinds of algorithms to order-invariant algorithms, and on an appropriate use of the known fact that every free group can be linearly ordered.

5.3.9. Voting Systems

In [44], [38], we consider a general framework for voting systems with arbitrary types of ballots such as orders of preference, grades, etc. We investigate their manipulability: in what states of the population may a coalition of electors, by casting an insincere ballot, secure a result that is better from their point of view?

We show that, for a large class of voting systems, a simple modification allows to reduce manipulability. This modification is *Condorcification*: when there is a Condorcet winner, designate her; otherwise, use the original rule.

When electors are independent, for any non-ordinal voting system (i.e. requiring information that is not included in the orders of preferences, for example grades), we prove that there exists an ordinal voting system whose manipulability rate is at most as high and which meets some other desirable properties. Furthermore, this result is also true when voters are not independent but the culture is *decomposable*, a weaker condition that we define.

Combining both results, we conclude that when searching for a voting system whose manipulability is minimal (in a large class of systems), one can restrict to voting systems that are ordinal and meet the Condorcet criterion.

In [35], we examine the geometrical properties of the space of expected utilities over a finite set of options, which is commonly used to model the preferences of an agent. We focus on the case where options are assumed to be symmetrical a priori, which is a classical neutrality assumption when studying voting systems. 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. Whereas Impartial Culture is widely used in Social Choice literature but limited to ordinal preference, our theoretical result allows to extend it canonically to cardinal preferences.

In [25], we study the manipulability of voting systems in a real-life experiment: electing the best paper in the conference Algotel 2012. Based on real ballots, we provide a quantitative study of the manipulability, as a function of the voting system used. We show that, even in a situation where all voting systems give the same winner by sincere voting, choosing the voting system is critical, because it has a huge impact on manipulability. In particular, one voting system fare way be better than the others: Instant-Runoff Voting.

5.4. Network Algorithms and Analysis

5.4.1. Bounds on the Cover Time in the Rotor-Router Model

In [23] and [33], we consider the *rotor-router mechanism*, which provides a deterministic alternative to the random walk in undirected graphs. In this model, a set of k identical walkers is deployed in parallel, starting from a chosen subset of nodes, and moving around the graph in synchronous steps. During the process, each node maintains a cyclic ordering of its outgoing arcs, and successively propagates walkers which visit it along its outgoing arcs in round-robin fashion, according to the fixed ordering. We consider the *cover time* of such a system, i.e., the number of steps after which each node has been visited by at least one walk, regardless of the starting locations of the walks. In the case of $k = 1$, Yanovski et al. (2003) and Bampas et al. (2009) showed that a single walk achieves a cover time of exactly $\Theta(mD)$ for any n-node graph with m edges and diameter D, and that the walker explores increasingly large Eulerian subgraphs before eventually stabilizes to a traversal of an Eulerian circuit on the set of all directed edges of the graph.

In [23], we provide tight bounds on the cover time of k parallel rotor walks in a graph. We show that this cover time is at most $\Theta(mD/\log k)$ and at least $\Theta(mD/k)$ for any graph, which corresponds to a speedup of between $\Theta(\log k)$ and $\Theta(k)$ with respect to the cover time of a single walk. Both of these extremal values of speedup are achieved for some graph classes. Our results hold for up to a polynomially large number of walks, $k = O(poly(n))$.

In [33], we perform a case study of cover time of the rotor-router, showing how the cover time depends on k for many important graph classes. We determine the precise asymptotic value of the rotor-router cover time for all values of k for degree-restricted expanders, random graphs, and constant-dimensional tori. For hypercubes, we also resolve the question precisely, except for values of k much larger than n. Our results can be compared to those obtained by Elsässer and Sauerwald (2009) in an analogous study of the cover time of k independent parallel random walks in a graph; for the rotor-router, we obtain tight bounds in a slightly broader spectrum of cases. Our proofs take advantage of a relation which we develop, linking the cover time of the rotor-router to the mixing time of the random walk and the local divergence of a discrete diffusion process on the considered graph.

5.4.2. Web Ranking and Aliveness

In [29] and [30], we investigate how to efficiently retrieve large portions of alive pages from an old crawl using orderings we called LiveRanks. Our work establishes the possibility of efficiently recovering a significant portion of the alive pages of an old snapshot and advocates for the use of an adaptive sample-based PageRank for obtaining an efficient LiveRank. Additionally, application field is not limited to Web graphs. It can be straightforwardly adapted to any online data with similar linkage enabling crawling, like P2P networks or online social networks.

5.4.3. Wireless Positioning

In [31], we consider how to construct a low-cost and efficient positioning system. We have proposed a new method called Two-Step Movement (2SM) to estimate the position of Mobile Terminal (MT). By exploiting useful information given by the position change of the device or user movement, this method can minimize the number of Reference Points (RP) required (*i.e.,* only one) in a localization system or navigation service and reduce system implementation cost. Analytical result shows that the user position can be derived, under noisy environment, with an estimation error about 10% of the distance between the RP and MT, or even less.

5.4.4. Content Centric Networking

Today's Internet usage is mostly centered around location-independent services. Because the Internet architecture is host-centric, content or service requests still have to be translated into locations, or the IP address of their hosts. This translation is realized through different technologies, e.g. DNS and HTTP redirection, which are currently implemented at the Application Layer. (ICN) proposes to evolve the current Internet infrastructure by extending the networking layer with name-based primitives.

In [45], we target the design and implementation of a content router, which is a network entity that implements *name-based forwarding*, or it can forward packets based on the content name they are addressed to. This work makes three major contributions. First, we propose an algorithm for name-based longest prefix match whose main novelty is the *prefix Bloom filter*, a Bloom filter variant that exploits the hierarchical nature of content prefixes. Second, a content router design that is compatible with both today's networking protocols and with widely used network equipments. Third, two innovative features that increase the scalability of a content router both in term of forwarding-information-base size and forwarding speed.

In the demonstration [34] held in the ICN conference, we demonstrate a high speed Information-Centric Network in a mobile backhaul setting. In particular, we emulate an information aware data plane and we highlight the significant benefits it provides in terms of both user experience and network provider cost in the backhaul setting. Our setup consists of high-speed ICN devices employed in a down-scaled realistic representation of a mobile backhaul topology, fed with traffic workloads characterized from Orange's mobile network. We compare numerical results activating and de-activating the ICN feature at run-time, showing the main differences between the two approaches. All the devices are implemented in a real high-speed multi-core equipment, and they are connected by means of internal port connections. Traffic is injected using a Traffic Generator which is implemented in the same architecture.

5.4.5. Information Dissemination

5.4.5.1. Dissemination with Noise or Limited Memory

In [26], we introduce the study of basic distributed computing problems in the context of noise in communication. We establish tight and almost tight bounds for the rumor spreading problem as well as for the majority-consensus problem.

In [11], we theoretically study a general model of information sharing within animal groups. We take an algorithmic perspective to identify efficient communication schemes that are, nevertheless, economic in terms of communication, memory and individual internal computation. We present a simple and natural algorithm in which each agent compresses all information it has gathered into a single parameter that represents its confidence in its behavior. Confidence is communicated between agents by means of active signaling. We motivate this model by novel and existing empirical evidences for confidence sharing in animal groups. We rigorously show that this algorithm competes extremely well with the best possible algorithm that operates without any computational constraints. We also show that this algorithm is minimal, in the sense that further reduction in communication may significantly reduce performances. Our proofs rely on the Cramér-Rao bound and on our definition of a Fisher Channel Capacity. We use these concepts to quantify information flows within the group which are then used to obtain lower bounds on collective performance.

5.4.5.2. Gossip and Rumor Spreading with Flooding

In [2], we address the flooding problem in dynamic graphs, where flooding is the basic mechanism in which every node becoming aware of an information at step t forwards this information to all its neighbors at all forthcoming steps $t' > t$. In particular, we show that a technique developed in a previous paper, for analyzing flooding in a Markovian sequence of Erdös-Rényi graphs, is robust enough to be used also in different contexts. We establish this by analyzing flooding in a sequence of graphs drawn independently at random according to a model of random graphs with given expected degree sequence. In the prominent case of power-law degree distributions, we prove that flooding takes almost surely $O(\log n)$ steps even if, almost surely, none of the graphs in the sequence is connected. In the general case of graphs with an arbitrary degree sequence, we prove several upper bounds on the flooding time, which depend on specific properties of the degree sequence.

5.4.6. Small-world Networks

In [9], we study decentralized routing in small-world networks that combine a wide variation in node degrees with a notion of spatial embedding. Specifically, we consider a variant of J. Kleinberg's grid-based small-world model in which (1) the number of long-range edges of each node is not fixed, but is drawn from a power-law probability distribution with exponent parameter $\alpha \geq 0$ and constant mean, and (2) the long-range edges are considered to be bidirectional for the purposes of routing. This model is motivated by empirical observations indicating that several real networks have degrees that follow a power-law distribution. The measured power-law exponent α for these networks is often in the range between 2 and 3. For the small-world model we consider, we show that when $2 < \alpha < 3$ the standard greedy routing algorithm, in which a node forwards the message to its neighbor that is closest to the target in the grid, finishes in an expected number of $O(\log^{\alpha-1} n \cdot \log\log n)$ steps, for any source–target pair. This is asymptotically smaller than the $O(\log^2 n)$ steps needed in Kleinberg's original model with the same average degree, and approaches $O(\log n)$ as α approaches 2. Further, we show that when $0 \leq \alpha < 2$ or $\alpha \geq 3$ the expected number of steps is $O(\log^2 n)$, while for $\alpha = 2$ it is $O(\log^{4/3} n)$. We complement these results with lower bounds that match the upper bounds within at most a $\log\log n$ factor.

5.4.7. Voting Systems and Path Selection in Networks

In [24], we apply our theoretical and experimental results on voting systems to a network use case: choosing a path in a network. In our model, nodes have an economical reward or cost for each possible path and they vote to elect the path. We show that the choice of the voting system has an important impact on the manipulability and the economical efficiency of this system. From both points of view, Instant-Runoff Voting gives the best results.

HIPERCOM2 Team

6. New Results

6.1. Highlights of the Year

- Hipercom 2 took part to the Inria-Industry meeting focusing on Telecommunications organized by Inria at Rocquencourt in November 2014. We presented a demonstration of the OCARI wireless sensor network.
- Hipercom 2 organized an Inria-DGA day "Software Defined Network (SDN) & MANET" at Paris in October 2014.

6.2. New Results about Wireless Sensor Networks

6.2.1. Node activity scheduling and routing in Wireless Sensor Networks

Participants: Cédric Adjih, Ichrak Amdouni, Pascale Minet.

The need to maximize network lifetime in wireless ad hoc networks and especially in wireless sensor networks requires the use of energy efficient algorithms and protocols. Motivated by the fact that a node consumes the least energy when its radio is in sleep state, we achieve energy efficiency by scheduling nodes activity. Nodes are assigned time slots during which they can transmit and they can turn off their radio when they are neither transmitting nor receiving. Compared to classical TDMA-based medium access scheme, spatial bandwidth use is optimized: non interfering nodes are able to share the same time slots, collisions are avoided and overhearing and interferences are reduced.

In 2014, we study the issue of delay optimization and energy efficiency in grid wireless sensor networks (WSNs). We focus on STDMA (Spatial Reuse TDMA)) scheduling, where a predefined cycle is repeated, and where each node has fixed transmission opportunities during specific slots (defined by colors). We assume a STDMA algorithm that takes advantage of the regularity of grid topology to also provide a spatially periodic coloring ("tiling" of the same color pattern). In this setting, the key challenges are: 1) minimizing the average routing delay by ordering the slots in the cycle 2) being energy efficient. Our work follows two directions: first, the baseline performance is evaluated when nothing specific is done and the colors are randomly ordered in the STDMA cycle. Then, we propose a solution, ORCHID that deliberately constructs an efficient STDMA schedule. It proceeds in two steps. In the first step, ORCHID starts form a colored grid and builds a hierarchical routing based on these colors. In the second step, ORCHID builds a color ordering, by considering jointly both routing and scheduling so as to ensure that any node will reach a sink in a single STDMA cycle. We study the performance of these solutions by means of simulations and modeling. Results show the excellent performance of ORCHID in terms of delays and energy compared to a shortest path routing that uses the delay as a heuristic. We also present the adaptation of ORCHID to general networks under the SINR interference model.

6.2.2. Time slot and channel assignment in multichannel Wireless Sensor Networks

Participants: Pascale Minet, Ridha Soua, Erwan Livolant.

Applying WSNs in industrial environment requires fast and reliable data gathering (or data convergecast). If packets are forwarded individually to the sink, it is called raw data convergecast. We resort to the multichannel paradigm to enhance the data gathering delay, the robustness against interferences and the throughput. Since some applications require deterministic and bounded convergecast delays, we target conflict free joint time slot and channel assignment solutions that minimize the schedule length. Such solutions allow nodes to save energy by sleeping in any slot where they are not involved in transmissions.

After a comprehensive survey on multichannel assignment protocols in wireless sensor networks, we study raw convergecast in multichannel wireless sensor networks (WSNs) where the sink may be equipped with multiple radio interfaces. We propose $Wave$,a simple, efficient and traffic-aware distributed joint channel and time slot assignment for raw convergecast. Our target is to minimize the data gathering delays and ensure that all packets transmitted in a cycle are delivered to the sink in this cycle, assuming no packet loss at the physical layer. We evaluate the number of slots needed to complete the convergecast by simulation and compare it to the optimal schedule and to a centralized solution. Simulations results indicate that our heuristic is not far from the optimal bound for raw convergecast. Unlike most previously published papers, $Wave$ does not suppose that all interfering links have been removed by channel allocation. In addition, $Wave$ is able to easily adapt to traffic changes. $Wave$ could be used to provide the schedule applied in the $802.15.4e$ TSCH based networks.

6.2.3. Optimized WSN Deployment

Participants: Ines Khoufi, Pascale Minet, Erwan Livolant.

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 a temporary industrial worksite, damage assessment after a disaster.... Many of these applications require the full coverage of the area considered. 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.

In the literature, many studies assume that this area is rectangular and adopt the classical deployment in triangular lattice that has been proved optimal. In real life, things are more complex. For instance, in an industrial worksite, the area to cover has an irregular shape with many edges and is not necessarily convex. Moreover, few papers take obstacles into account. Those that do assume that obstacles are constituted by a juxtaposition of rectangles that seems an unrealistic assumption. In real deployments, the shape of obstacles may be irregular. 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. That is why we focus on the deployment of wireless sensor nodes in an arbitrary realistic area with an irregular shape, and with the presence of obstacles that may be opaque. Moreover, we propose a method that tends to minimize the number of sensor nodes needed to fully cover such an area.

Mobile robots can be used to deploy static wireless sensor nodes to achieve the coverage and connectivity requirements of the applications considered. Many solutions have been provided in the literature to compute the set of locations where the sensor nodes should be placed. We show how this set of locations can be used by a mobile robot to optimize its tour to deploy the sensor nodes to their right locations. In order to reduce both the energy consumed by the robot, its exposure time to a hostile environment, as well as the time at which the wireless network becomes operational, the optimal tour of the robot 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 Robot Deploying Sensor nodes problem, in short RDS. We first show how this problem differs from the well-known traveling salesman problem. We then propose an integer linear program formulation of the RDS 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.

6.2.4. 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 formalization.

6.2.5. 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.

6.2.6. 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.

6.2.7. Opportunistic routing cross-layer schemes for low duty-cycle wireless sensor networks

Participants: Mohamed Zayani, Paul Muhlethaler.

This is a joint work with Nadjib Aitsaadi from University of Paris 12.
The opportunistic aspect of routing is suitable with such networks where the topology is dynamic and protocols based on topological information become inefficient. Previous work initiated by Paul Muhlethaler and Nadjib Aitsaadi consisted in a geographical receiver-oriented scheme based on RI-MAC protocol (Receiver-Initiated MAC). This scheme is revised and a new contribution proposes to address the same problem with a sender-oriented approach. After scrutinising different protocols belonging to this classification, the B-MAC protocol is chosen to build a new opportunistic cross-layer scheme. Our choice is motivated by the ability of this protocol to provide to a sender the closest neighbor to the destination (typically a sink). In other words, such a scheme enables us to obtain shorter paths in terms of hops which would increase the efficiency of information delivery. In counterparts, as it relies on long preambles (property of B-MAC) to solicit all the neighborhood, it needs

larger delays and energy consumption (1% of active time). Nevertheless, this proposal remains interesting as the studied networks are dedicated to infrequent event detection and are not real time-oriented.

When we use BMAC with opportunistic routing, one main advantage is that there is no transmission when there is no event detected in the network in contrast to RI-MAC where beacons of awaking nodes are periodically sent. However, when an event occurs in the area monitored, the end-to-end delay to deliver the alert packet to the sink is much greater with BMAC than with RI-MAC. This may pose problem to some real-time applications. We have propose a scheme where, instead of sending a long preamble to gather all the neighbor nodes, the packet is directly sent. The acknowledgement of the packet allows tthe sender to know whether (or not) the progression towards the destination is sufficient. If it is not the case the packet is sent again. More neighbor node will be awaken and the progression towards the destination will be improved. The selection of the relay terminates when the progression towards the destination is above a given threshold. Actually this relaying scheme encompasses two levels of opportunism. The first level consists in selecting only the awake nodes, the second level consists in selecting the best nodes among the awake nodes. We can show that doing so only slightly increase the number of hops to reach the sink whereas the delay per hop is largely reduced. Thus the end-to-end is very significantly reduced and we still have the property that there is no transmission when there is no event detected in the network.

6.3. Cognitive Radio Networks

6.3.1. Multichannel time slot assignment in Cognitive Radio Sensor Networks

Participants: Ons Mabrouk, Pascale Minet, Ridha Soua, Ichrak Amdouni.

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). In 2013 we focused on the MultiChannel Time Slot Assignment problem (MC-TSA) in CRSN and proposed an Opportunistic centralized TIme slot assignment in COgnitive Radio sensor networks (OTICOR). This latter 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 show that, even in the presence of several PUs, the average throughput granted to SUs remains important. We also show how to get the best performances of OTICOR when the channel occupancy by PUs is known.

In 2014, we proposed 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. To achieve our goal, we propose an optimized version of our previous scheduling algorithm Opportunistic centralized TIme slot assignment in COgnitive Radio sensor networks (OTICOR). This algorithm 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 shown through simulations that using the control period for data transmission minimizes the schedule length and maximizes the throughput.

6.4. Mobile ad hoc and mesh networks

6.4.1. Development and implementation of a network coding module for NS3

Participants: Cédric Adjih, Ichrak Amdouni, Hana Baccouch.

DragonNet is a complete modular solution of network coding. This solution is responsible of coding, decoding, maintaining necessary information and the associated signaling. It is designed to be extensible. A variant of DragonNet was specified for wireless sensor networks and implemented.

As a follow-up to the ADT MOBSIM (and the previous module EyWifi), DragonNet was also integrated as a module for the NS-3 simulation tool.

6.4.2. Optimized Broadcast Scheme for Mobile Ad hoc Networks

Participants: Nadjib Achir, Paul Muhlethaler.

The main objective is to select the most appropriate relay nodes according to a given cost function. Basically, after receiving a broadcast packet each potential relay node computes a binary code according to a given cost function. Then, each node starts a sequence of transmit/listen intervals following this code. In other words, each 0 corresponds to a listening interval and each 1 to a transmit interval. During this active acknowledgment signaling period, each receiver applies the following rule: if it detects a signal during any of its listening intervals, it quits the selection process, since a better relay has also captured the packet. Finally, we split the transmission range into several sectors and we propose that all the nodes within the same sector use the same CDMA orthogonal spreading codes to transmit their signals. The CDMA codes used in two different sectors are orthogonal, which guarantees that the packet is broadcast in all possible directions. The obtained results demonstrate that our approach outperforms the classical flooding by increasing the delivery ratio and decreasing the number of required relays and thus the energy-cost.

6.5. Learning for an efficient and dynamic management of network resources and services

6.5.1. Learning in wireless sensor networks

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet, Selma Boumerdassi.

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 short-term, average-term and long-term 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.

In 2014, we applied on-line learning strategies to predict the quality of a wireless link in a WSN, based on the LQI metric 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. The model we propose learns on-line how to adapt to dynamic changes of the environment to compute efficient predictions. It presents a very good reactivity and adaptability. The simulations using traces collected in a real WSN based on the IEEE 802.15.4 standard have shown that the past time-windows which are effective for the prediction should have medium durations, about

200-400ms. The time windows durations less than 200ms do not give a good prediction, while durations larger than 400ms are efficient only in low variations environment. We note that these results strongly depend on the real traces, but the great advantage of the model is that it is self-adaptive to input traces profile. In this context, because of data normalization, the impact of loss functions is limited: entropy and square loss functions seem to give better and more stable predictions. Also, the experts prediction method should be adapted to traces profile. For low variation environment values, the average on past time windows is a good approximation. For high variation environment, a method predicting smoothed values close to minimum real values is more appropriate. Hence, the predicted values will be stabilized around the low values, avoiding estimations varying too much. Simulation results also show that for both types of experts (AMW and SES), the best expert depends on the phase considered. This is the reason why a forecaster is needed. Furthermore, the predictions of the EWA forecaster using SES experts are shown to be reactive and accurate. This combination minimizes the cumulated loss regarding the real LQI values, compared with any other combination such as EWA-AMW, BE-AMW and BE-SES, given by decreasing performance order.

6.5.2. Prediction and energy efficiency for datacenters

Participants: Dana Marinca, Nesrine Ben Hassine, Pascale Minet, Selma Boumerdassi.

The exponential development of Information and Communication Technologies (ICT) have led to an over consumption of services and data shared in networks. From computing in companies to unified communications through social networks and Internet of Things, the use of ICT a reach the highest level ever. The complexity involved by these different services reveals the limits of computing in companies and leads a majority of organisms to partially or completely host the management of there information system in data centers. The latter are larger and larger and are composed of buildings containing powerful computing equipments and air-conditionning systems. 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.

6.6. Vehicular Ad hoc NETworks (VANETs)

6.6.1. Congestion Control in VANETs

Participants: Paul Muhlethaler, Anis Laouiti.

We have reviewed the schemes of Congestion Control in VANETs for safety messages. The solutions proposed are: to adapt the generation rate, to adapt the transmission power or to adapt the carrier sense threshold. Some mechanisms employ different states depending on the channel load. Some other schemes use recursive adaptation of their parameters (e.g. LIMERIC). According to a few studies the recursive adaptation system provide a better adaptation of the VANET to the channel load. We will study how the transmission rate and the carrier sense threshold (or transmission power) can be best adapted in order to send CAM: Car Awareness Messages with the highest rate and to the furthest vehicles while maintaining the total load below a given threshold. We will also study the better combination of transmission rate and the carrier sense threshold for the CAM.

6.6.2. TDMA schemes for VANETs

Participants: Mohamed Hadded, Paul Muhlethaler, Anis Laouiti.

This is a joint work with Leila Saidane and Rachid Zabrouba from ENSI (Tunisia).

Vehicular Ad-hoc NETworks (VANETs) help improving traffic safety and efficiency. Each vehicle can exchange information to inform other vehicles about the current status of the traffic flow or a dangerous situation such as an accident. Road safety and traffic management applications require a reliable communication scheme with minimal transmission collisions, which thus increases the need for an efficient Medium Access Control (MAC) protocol. However, the MAC in a vehicular network is a challenging task due to the high speed of the nodes, frequent changes in topology, the lack of an infrastructure, and various QoS requirements. Recently several Time Division Multiple Access (TDMA)-based medium access control protocols have been proposed for vehicular ad hoc networks in an attempt to ensure that all the vehicles have enough time to send safety messages without collisions and reducing end-to-end delay and packet loss rate. We have identified the reasons for using the collision-free medium access control paradigm in VANETs. We have then presented a novel topology-based classification and we provide an overview of TDMA-based MAC protocols that have recently been proposed for VANETs. We have focus on the characteristics of these protocols as well as their benefits and limitations. Finally we have given a qualitative comparison, and we have discussed some open issues that need to be tackled in future studies to improve the performance of TDMA-based MAC protocols for vehicle to vehicle V2V communication.

INFINE Team

6. New Results

6.1. Highlights of the Year

- We proved a conjecture made in 2011 about the feasibility of non-trivial community detection just above a threshold below which it was known that only trivial detection could be done, see [13]. This was published in ACM STOC'14 and well-received, as the proof required the invention of new techniques to control the spectral properties of random matrices.
- The official opening of IoT-LAB of all sites through the "Workshop Internet Of Things/Equipex FIT IoT-LAB" held in Grenoble (on 6 and 6 november 2014), has been a major event for our team: it concludes several years of preparation of the IoT-LAB site located in Rocquencourt, currently managed by C. Adjih, E. Baccelli and I. Amdouni, which was itself opened the same month https://www.iot-lab.info/opening-of-the-paris-rocquencourt-site/.

6.2. Panorama

All the INFINE research activities encompass both theoretical or protocol designing research (to seek for conceptual advances or optimizations) and applied research (to validate and/or experiment the proposed concepts against real networking scenarios). The target applications range from Internet-based applications to mobile wireless networks. Taking an information- and user-centric perspective, we envision networks as means to convey relevant information to users, while adapting to customary practices (in terms of context, interests, or content demands) of such users. INFINE is thus organized along three main axes, namely Online Social Networks, Resource and Traffic Management, and Spontaneous Wireless Networks.

6.3. Online Social Networks (OSN)

Community detection; bandit algorithms; privacy preservation; reward mechanisms

6.3.1. Community detection

Participants: Laurent Massoulié, Marc Lelarge, Jiaming Xu.

We have progressed in the design of spectral methods for community detection and in the corresponding analysis (see above and references [3], [13], [22]).

6.3.2. Bandit algorithms for active learning of content type at low spam cost

Participants: Laurent Massoulié, Mesrob Ohanessian, Alexandre Proutière.

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. Work currently under submission.

6.4. Resource and Traffic Management

Traffic offloading; infrastructure deployment; opportunistic routing; traffic modeling; intermittently connected networks.

6.4.1. From Routing 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. Different parts of this work has been published in the international conferences IEEE SECON 2014 [14], IEEE WCNC 2014 [18] and IEEE WMNC 2014 [17], and in the international Student workshop IEEE Infocom [15]. An extended version of this paper is under submission in a trasaction. This version includes new characterization results of the used trace and new analysis of space-traffic correlation.

6.4.2. 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 [14] and to improve the quality of communication service provided. Hence, the definition of a *usage pattern* can allow telecommunication operators to better forseen 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 has been published in the international conference IEEE PERCOM 2014 [16]. An extended version of this paper is under submission in a trasaction and a technical report is available in [26]. This version includes new characterization results of the used trace, including analysis correlating age and gender to traffic demands, as well as new profiling results.

6.4.3. 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 in under submission in an international conference.

6.4.4. 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. Some parts of this work were previously published in the international conference ACM GECCO 2012 and in the Elsevier Computer Networks journal.

6.4.5. 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. This work was published in the international conference IEEE ISCC 2014 [10], in the international Student workshop of IEEE Infocom 2014 [9], and at the internation workshop Internet of Things Communications and Technologies (IoT 2013) held in conjunction with IEEE WiMob 2013.

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. This work was published as a short paper at the international conference ACM MSWiM 2014 [8]

6.4.6. 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. An initial version of this work was published at the international conference ACM MobiHoc 2013, and an extended version is under submission in an international trasaction.

6.4.7. 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. An initial version of this work was published in the international conference ACM MSWiM 2013 (selected as one of the five better papers of this venue) and an extented version bringing new analysis (e.g., the contact duration-related analysis performed by the internship Lavanya Addepalli and the PhD student Guangshuo Chen) was accepted to be published in 2015 at the Performance Evaluation Elsevier Journal [2].

6.5. Spontaneous Wireless Networks and Internet of Things

internet of things; wireless sensor networks; dissemination; resource management

6.5.1. Network Coding in Large Scale IoT Networks

Participants: Cedric Adjih, Ichrak Amdouni, Hana Baccouch, Antonia Masucci.

We had designed a generic broadcast protocol, called DragonNet, based on network coding and designed for constrained networks such as wireless sensor networks and internet of things. It minimizes the assumptions made of the networks. A variant of this protocol was implemented and run on IoT-LAB: some results were initially presented at IRTF, and a live demo was presented in MASS in october 2014.

6.5.2. Information-Centric Networking in the Internet of Things

Participants: Emmanuel Baccelli, Oliver Hahm, Matthias Waehlisch, Thomas Schmidt, Christian Mehlis.

Within this activity, we explored the feasibility, advantages, and challenges of an ICN-based approach in the Internet of Things. We report on the first NDN experiments in a life-size IoT deployment, spread over tens of rooms on several floors of a building. Based on the insights gained with these experiments, we have analysed the shortcomings of CCN applied to IoT. Several interoperable CCN enhancements are then proposed and evaluated. We significantly decreased control traffic (i.e., interest messages) and leverage data path and caching to match IoT requirements in terms of energy and bandwidth constraints. Our optimizations increase content availability in case of IoT nodes with intermittent activity. Within this activity, we also provided the first experimental comparison of CCN with the common IoT standards 6LoWPAN/RPL/UDP.

MADYNES Project-Team

6. New Results

6.1. Highlights of the Year

The following points of 2014 deserves to be highlighted:

- One new permanent member joined the MADYNES team: Jérôme François as Inria researcher.
- An IBM Faculty Award has been received by a team member (Rémi Badonnel, TELECOM Nancy) for his work on security and cloud computing.

BEST PAPER AWARD :

[21] **8th IFIP WG 6.6 International Conference on Autonomous Infrastructure, Management, and Security, AIMS 2014**. A. MAYZAUD, A. SEHGAL, R. BADONNEL, I. CHRISMENT, J. SCHÖNWÄLDER.

6.2. Monitoring

6.2.1. P2P network monitoring

Participants: Thibault Cholez [contact], Isabelle Chrisment, Olivier Festor.

Finishing a work started several years ago with our colleagues from the team Complex Network [0] at the LIP6, we published a final result on the comparison of paedophile activity in different P2P systems [5]. We designed a methodology for comparing KAD and eDonkey, two P2P systems among the most prominent ones and with different anonymity levels. We have detected paedophile-related queries with a previously validated tool and we proposed, for the first time, a large-scale comparison of paedophile activity in two different P2P systems.

We are also glad to have contributed to a book chapter in french on the uses and misuses of digital identities on the Internet [33]. It summarizes several years of work of the team, fighting against the Sybil attack in P2P networks in order to improve their security and quality of service.

6.2.2. Anonymous networks monitoring

Participants: Juan Pablo Timpanaro, Isabelle Chrisment [contact], Olivier Festor.

Anonymous networks have emerged to protect the privacy of network users. Large scale monitoring on these systems allows us to understand how they behave and which type of data is shared among users.

In 2014, we continued our research about the I2P anonymous network [0]. This network is optimized for anonymous web hosting and anonymous file-sharing. I2P's file-sharing community is highly active with users deploying their file-sharing applications on top of the network. I2P uses a variation of Onion routing, thus assuring the unlinkability between a user and its file-sharing application. In [26] we took the first step towards the linkability of users and applications in the I2P network. We conducted a group-based characterization, where we determine to what extent a group of users is responsible for the overall I2P's file-sharing activity. We used Pearson's coefficient to correlate users from two cities and the most used anonymous file-sharing application.

6.2.3. Smartphone usage monitoring

Participants: Vassili Rivron [contact], Mohammad Irfan Khan, Simon Charneau [Inria], Isabelle Chrisment.

Over the last few years the number of smartphone applications has increased enormously. In 2014, we passively collected smartphones usage logs in the wild by inviting the crowd to participate in the PRACTIC [0] contest and install our crowdsensing application to contribute anonymous smartphone usage logs, voluntarily and in the most natural settings (their own phone, own pricing plan) .

[0] http://www.complexnetworks.fr/
[0] http://i2p2.de
[0] http://beta.apisense.fr/practic

Complementary to sensing we also collected contextual information (social, demographic, professional) and information about users'perception via survey questionnaires built in the application or on the web.

This experiment used a crowd sensing platform called APISENSE ®[0] and developed by the Inria Spirals Team. It was carried out in the context of building a country-wide Internet observation platform in France, called Metroscope [0].

6.3. Security

6.3.1. Security Automation

Participants: Rémi Badonnel [contact], Martin Barrere, Gaëtan Hurel, Abdelkader Lahmadi, Olivier Festor.

The main research challenge addressed in this work is focused on enabling configuration security automation in dynamic networks and services.

A first part of our work in the year 2014 was centered on a strategy for remediating known vulnerabilities, formalizing the correction decision problem as a satisfiability or SAT problem [10]. From a proactive perspective, it should be able to decide which potential states could be dangerous. By specifying our vulnerability knowledge source (OVAL repository) as a propositional logical formula, we have fixed system properties that we cannot change and free those variables for which changes are available. We have introduced the X2CCDF language, built on top of XCCDF and OVAL, that allows us to express the impact of these changes over target systems. These descriptions can be used for analyzing the security impact of changes without actually changing the system. When this information is not available, we have considered the NETCONF protocol and its notion of candidate state where changes can be applied, analyzed and rolled back if necessary.

A second part of our work has been dedicated to the orchestration of security functions in the context of mobile smart environments [19]. 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 inducing local resource consumption. In that context, we have investigated a new approach for outsourcing mobile security functions as cloud-based services for smartphones and tablets [32]. The outsourced functions are dynamically activated, configured and orchestrated using software-defined networking and virtualization techniques. We consider the use of security compositions in order to dynamically fit the security requirements of mobile devices according to their current contexts. This approach is based on different traversal schemes (sequential, conditional, and concurrent). The solution has been prototyped based on the mininet software-defined networking emulator, jointly with mobile devices using the android operating system.

6.3.2. SDN-based security

Participants: Jérôme François [contact], Lautaro Dolberg [University of Luxembourg], Olivier Festor, Thomas Engel [University of Luxembourg].

By decoupling the data and control plane, Software-Defined Networking allows a fine grained network management. Protocols like OpenFlow allow multiple actions like traffic forwarding or blocking but also modifications or monitoring with the extensive use of counters. Hence, many approaches have emerged the last year to enable some security functions like firewalls, flow monitoring and traffic redirection to middleboxes. These different scenarios have been evaluated in a survey paper [17] in cooperation with the university of Luxembourg.

Furthermore, we also proposed to leverage SDN, especially OpenFlow, for forensics purpose [18]. Indeed, through a recursive analysis on network path and flow tables in OpenFlow, it is possible to reconstruct the paths traversing by an anomaly.

[0] http:///www.apisense.com/
[0] http://metroscope.eu/

6.3.3. Phishing Detection

Participants: Jérôme François [contact], Samuel Marchal [University of Luxembourg], Radu State [University of Luxembourg], Thomas Engel [University of Luxembourg].

This work is a joint work with the University of Luxembourg.

The language used for phishing is a particular language aiming at attracting victims. To achieve that the attackers uses specific words related to well known brand names and reassuring words. Our method to detect such abnormal domain names relies on word decomposition and semantic analysis. As an example, we can learn if having both *microsoft* and *protected* in domain is significative of a malicious domain. Actually, not all words can be represented during the learning and we use semantic similarities to also extend this knowledge (for example, we can *derive* safe from *protected*).

Our recent work [20] was focusing on extending this domain-based analysis to the full analysis of an url. We have also observed that most of false positives or negatives we obtained with previous methods are biased by natural language corpus while the *Internet vocabulary* is different.

Hence, we extracted from Google and Yahoo statistics about search queries. Our observation highlights that the relation between the different parts of the URL (the domain and the path) is a discriminative feature for malicious URL identification.

Finally, a more in-depth feature analysis is provided in [8], which also proposes leveraging streaming data analytics by instantiating our method on Storm.

6.3.4. Flows and logs analysis

Participants: Jérôme François [contact], Abdelkader Lahmadi.

Machine generated-log data is a fundamental part of information technology systems. They are usually generated at every component of distributed information systems including routers, security products, web proxies, DHCP servers, VPN servers, or any end-points like mobile devices or connected things, etc. They often contain high volumes of interesting information and are among the first data source to be analyzed for the detection of abnormal activities due to running attacks or malicious running applications. A better understanding of these attacks and malicious applications requires the elaboration of efficient and novel methods and techniques able to analyze these logs.

In [16], we carried an empirical analysis of the logs generated by the logging system available in Android environments. The logs are mainly related to the execution of the different components of applications and services running on an Android device. We have analyzed the logs using self organizing maps where our goal is to establish behavioral fingerprints of Android applications. The developed methodology allows us the better understand Android Apps regarding their granted permissions and performed actions.

During the year 2014, we have also maintained an IETF draft [50] to make 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.

6.3.5. Sensor networks monitoring

Participants: Rémi Badonnel, Isabelle Chrisment, Olivier Festor, Abdelkader Lahmadi [contact], Anthéa Mayzaud.

Low Power and Lossy Networks (LLNs) are made of interconnected wireless devices with limited resources in terms of energy, computing and communication. The communication channels are low-bandwidth, high loss rate and volatile wireless links subject to failure over time.

This year, our work on security-oriented monitoring [28] has focused on quantifying the effects of version number manipulation attacks within RPL networks [21]. Through simulations it was discovered that control overhead can increase by up to 18 times, thereby impacting energy consumption and channel availability. This in turn can reduce the delivery ratio of packets by up to 30% and nearly double the end-to-end delay in a network. A strong correlation between the position of the attacker and the effect on the network was also observed.

In that context, we have designed a mitigation strategy based on an adaptive threshold to cover a large variety of DODAG inconsistency attacks [25] in a lightweight manner. Currently RPL attempts to counteract such attacks by using a fixed threshold. During experimentations it becomes clear that the adaptive threshold is able to reduce the control message overhead, compared to fixed threshold, by up to 13% in short lived and 55% in long-lived networks. This leads to large reductions, i.e., between 10%-40%, in energy consumption.

In addition, we have investigated a distributed passive monitoring architecture for RPL-based advanced measurement infrastructure networks.

6.3.6. *Intrusion Detection System in Wireless Sensor*

Participants: Emmanuel Nataf [contact], Hubert Kenfack Ngankam.

This work is based on a previous work about the definition of an ontology to classify intrusion attacks in a wireless sensors network. A first implementation of this ontology focuses on the black hole and the sink hole intrusion where some malicious sensor node either do not forward data to a central point of collect or try to be elected as the best next hop toward the central point.

We look at discover malicious nodes by an analysis of the network topology obtained by data gathered from the network itself. At regular interval, we built a snapshot view of the network topology and compare it with the previous one in order to detect anomalies such as a whole sub network that disappear or an under-optimal network topology.

Simulation results are good and we will continue on this way.

6.3.7. *SCADA systems security*

Participants: Abdelkader Lahmadi [contact], Younes Abid.

SCADA systems are facing several attacks and threats which are growing in number and complexity. A key challenge in this context is the simulation and the assessment of the impact and the propagation of these attacks on SCADA system components over time. During the year 2014, we have developed a novel methodology [38] based on stochastic modeling to simulate the impact of attacks on SCADA systems. The system is modeled as a network of interacting markov chains and the impact of an attack is simulated using the influence model. In this model, the state of each node of the system is either influence by its own Markov chain or by the state of its neighboring nodes. We have modeled and analyzed a SCADA system with 200 control nodes and several servers. We have modeled different attacks (intrusion, DoS, malware) where attack nodes are introduced in the interacting SCADA network to influence control node behaviors. For each attack, we have simulated and assessed over time the availability of the overall system regarding the number of failed nodes.

6.3.8. *Management of HTTPS traffic*

Participants: Thibault Cholez [contact], Isabelle Chrisment, Shbair Wazen, Jérôme François.

Surveys show that websites are more and more being served over HTTPS. They highlight an increase of 48% of sites using TLS over the past year (2013),

We 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 show that SNI has two weaknesses, regarding (1) backward compatibility and (2) multiple services using a single certificate. We demonstrated thanks to a web browser plug-in called *Escape* that we designed and implemented, how these weaknesses can be practically used to bypass firewalls and monitoring systems relying on SNI. The results show positive evaluation (firewall's rules successfully bypassed) for all tested websites. This work will be published in the experience session of the IFIP/IEEE International Symposium on Integrated Network Management (IFIP/IEEE IM'15).

We also started a new work on the precise identification of websites accessed through HTTPS in the context of network forensic investigation. We use a new set of features in conjunction with machine learning techniques to achieve a high accuracy.

6.4. Routing

6.4.1. Routing in Wireless Sensor Networks

Participants: Emmanuel Nataf [contact], Patrick-Olivier Kamgueu.

We deployed a wireless sensors network in the laboratory during two time period of 3 months. The first was with the legacy routing (based on expected transmission time metric) and the second was with our routing process based on a composition of several metrics (i.e. energy, transmission time and delay) by the use of fuzzy logic. We have compared these experiments by packet loss ratio and energy consumption. In all case, our routing leads to a better network [48].

6.4.2. Operator calculus based routing in Wireless Sensor Networks

Participants: Evangelia Tsiontsiou, Bernardetta Addis, Ye-Qiong Song [contact].

For supporting different QoS requirements, routing in WSN must simultaneously consider several criteria (e.g., minimizing energy consumption, hop counts or delay, packet loss probability, etc.). When multiple routing metrics are considered, the problem becomes a multi-constrained optimal path problem (MCOP), which is known as NP-complete.

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. 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 and emulated for TelosB motes using Cooja. We compared its performance against tree and directional flooding routing and show the advantages of our technique. Our ongoing work consists in its comparison with RPL to show its effective contribution to handle simultaneously several IETF ROLL routing metrics.

This work is under development as part of Lorraine AME Satelor project.

6.4.3. 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 [Polytecnique de Montreal, Canada], Sara Mattia [IASI, CNR, Roma, Italy], Brunide Sansò [Polytecnique 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. We also present heuristic results to compare the optimal operational planning with online energy management operation ([3])

Two very important issues that may be affected by green networking techniques are resilience to node and link failures, and robustness to traffic variations. We thus extended the optimization models. To guarantee network survivability we consider two different schemes, dedicated and shared protection, which assign a backup path to each traffic demand and some spare capacity on the links along the path. Robustness to traffic variations is provided by tuning the capacity margin on active links in order to accommodate load variations of different magnitude. Both exact and heuristic methods are proposed. Experimentations carried out on realistic networks operated with flow-based routing protocols (like MPLS) allow us to quantitatively analyze the trade-off between energy cost and level of protection and robustness. Results show that significant savings, up to 30%, may be achieved even when both survivability and robustness are fully guaranteed [4].

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. Preliminary results on a simplified model ([29]) are very encouraging.

6.4.4. Energy-aware joint management of networks and Cloud infrastructures

Participants: Bernardetta Addis [contact], Danilo Ardagna [DEIB, Politecnico di Milano, Italy], Giuliana Carello [DEIB, Politecnico di Milano, Italy], Antonio Capone [DEIB, Politecnico di Milano, Italy].

Fueled by the massive adoption of Cloud services, overall service centers and networks account for 2–4% of global CO_2 emissions and it is expected they can reach up to 10% in 5–10 years.

The geographical distribution of the computing facilities offers many opportunities for optimizing energy consumption and costs by means of a clever distribution of the computational workload exploiting different availability of renewable energy sources, but also different time zones and hourly energy pricing. Energy and cost savings can be pursued by dynamically allocating computing resources to applications at a global level, while communication networks allow to assign flexibly load requests and to move data. We propose an optimization framework able to jointly manage the use of brown and green energy in an integrated system and to guarantee quality requirements. We propose an efficient and accurate problem formulation that can be solved for real-size instances in few minutes to optimality. Numerical results, on a set of randomly generated instances and a case study representative of a large Cloud provider, show that the availability of green energy have a big impact on optimal energy management policies and that the contribution of the network is far from being negligible ([2]).

6.4.5. Content centric wireless sensor networks

Participants: Abdelkader Lahmadi [contact], Younes Abid, Olivier Festor.

During this year, we have instantiated a novel named data aggregation method [9] dedicated to wireless sensor networks . The method relies on an adaptation of the CCNx protocol implementation that we have developed in a previous work. Our method extends the CCNx protocol with in-network processing functions to aggregate named data efficiently. We have implemented and tested our solution with the Contiki operating system which is an operating system for resources-constrained embedded systems and wireless sensor networks. Our simulation and measurement results using the Cooja simulator and physical nodes show that our solution has a small overhead in terms of exchanged messages and provides acceptable data retrieval delays.

6.5. Quality-of-Service

6.5.1. ICN cache management

Participants: Olivier Festor [contact], César Bernardini, Thomas Silverston.

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. As a central component of ICN is in-network caching, the ely used as a micro-blogging service. At the same time, Online Social Networks (OSN) carry extremely valuable information about users and their relationships. We argue that this knowledge can help to drastically improve the efficiency of ICN.

We therefore propose SACS, a caching strategy designed for the CCN architecture that includes social information [11]. CCN is to date the most widely adopted ICN architecture by the research and industrial community. The underlying idea in such strategy is that a small number of users counts a huge amount of social relationships, dominates the activity and receives most attention from other users. We call such users Influential users, and we argue that they produce content that is more likely to be consumed by others, and in consequence their content must be favored and replicated in priority. Our novel caching strategy is therefore prioritizing content from Influential users of the social network. To validate our strategy, we first propose a model of social network over the CCN architecture [30]. Our model has been designed based on the measurement of Pinterest, a web-based OSN system. Extensive simulations of the strategy have been performed, as well as a real implementation on CCNx and deployment over the PlanetLab testbed. Our results with SACS are significant and increase drastically the caching performance of ICN architecture. content

Efficient management of caches is a key success factor in Concent-Centric Networks where multiple (up to every single node in the network) entities act as caches of the shared content in the network. We pursued our investigations towards a common evaluation framework for cache strategies in Content-centric networks and towards the definition of novel cache strategies, exploiting context information available at the service level of today's internet.

6.5.2. *Self-adaptive MAC protocol for both QoS and energy efficiency*

Participants: Kévin Roussel, Shuguo Zhuo, Ye-Qiong Song [contact].

WSN research focus has progressively been moved from the energy issue to the QoS issue. Typical example is the MAC protocol design, which cares about not only low duty-cycle at light traffic, but also high throughput with self-adaptation to dynamic traffic bursts.

The two MAC protocols that we have previously designed namely S-CoSenS and iQueue-MAC, have been successfully implemented on SMT32W108 SoC chips. Two 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. Secondly, both S-CoSenS and iQueue-MAC have been implemented on RIOT OS. An additional contribution is related to the RIOT OS development itself since we have improved the robustness of the existing ports of RIOT OS on MSP430-based motes, making it a suitable software platform for tiny motes and devices. More generally, through this part of work, we have shown that RIOT OS is also suitable for implementing high-performance MAC protocols, thanks to its real-time features (especially hardware timers management). Part of this work has been supported by ANR-NFSC Quasimodo and PIA LAR projects.

6.5.3. *End-to-end delay modelling 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 [15] we have presented a new approach for extracting empirical Markov chain models from network protocol traces by means of Process Mining techniques. An empirical Markov chain model was obtained for the IEEE 802.15.4 beacon-enabled mode protocol allowing us to estimate the e2e delay for a multi-hop scenario. This approach has also been successfully applied to the case of ContikiMAC [14].

6.5.4. *Dynamic resource allocation in network virtualization*

Participants: Mohamed Said Seddiki, Mounir Frikha [SupCom, Tunis, Tunisie], Ye-Qiong Song [contact].

The objective of this research topic is to develop different resource allocation mechanisms in Network Virtualization, for creating multiple virtual networks (VNs) from a single physical network. It is accomplished by logical segmentation of the network nodes and their physical links.

This year we have focused on implementing and evaluating the used of SDN for managing the QoS in broadband access networks. Unfortunately, application-based QoS on a home network gateway faces significant constraints, as commodity home routers are not typically powerful enough to perform application classification, and many home users are not savvy enough to configure QoS parameters. In [24] we designed FlowQoS, an SDN-based approach where users can specify upstream and downstream bandwidth allocations for different applications at a high level, offloading application identification to an SDN controller that dynamically installs traffic shaping rules for application flows We designed a custom DNS-based classifier to identify different applications that run over common web ports; a second classifier performs lightweight packet inspection to classify non-HTTP traffic flows. We implemented FlowQoS on OpenWrt and demonstrated that it can improve the performance of both adaptive video streaming and VoIP in the presence of active competing traffic.

This work has been carried out as part of a co-supervised PhD thesis between University of Lorraine and SupCom Tunis.

6.5.5. Task and message scheduling in distributed real-time systems

Participants: Florian Greff, Laurent Ciarletta, Ye-Qiong Song [contact].

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. In [37], [36], [34], and [35], we provided an overview of both message scheduling techniques in networks and joint task and message scheduling approaches in closed-loop distributed control systems (networked control systems). Fault-tolerance is another critical issue that one must take into account. In collaboration with an industrial partner, we started a study on the real-time dependability of UAV multi-criticity system interconnected by an embedded mesh network. The future work aims at developing a robust mesh network routing protocol and studying the schedulability under constraints of multi-criticality and graceful degradation during mode change.

6.6. Multi-modeling and co-simulation tools for the evaluation and development of Smart* and other Pervasive Computing systems

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Yannick Presse, Emmanuel Nataf, Benjamin Segault.

Vincent Chevrier (Maia team, LORIA) is a collaborator and the correspondant for the MS4SG project, Benjamin Camus, Victorien Elvinger and Christine Bourjot (Maia team, LORIA) are collaborators for the AA4MM. 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 [51] 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 projet which involves MAIA, Madynes and EDF R&D on smart-grid simulation, we developed a proof of concepts for a smart-appartment case[12].

In 2014 we worked on the following research topics:

- Assessment and evaluation of complex systems.

  This work, centered on the problem of controlling complex systems proposed a control architecture within Tomas Navarrete's work [22], [23]. This "equation-free" approach uses a multi-agent model to evaluate the global impact of local control actions before applying the most pertinent set of actions. Based on a partial perception of the system state, we determine which actions to execute in order to avoid or favor certain global states of the system.

  Associated to our architecture, an experimental platform has been developed to confront the basic ideas or the architecture within the context of simulated "free-riding" phenomenon in peer to peer file exchange networks. We have demonstrated that our approach allows us to drive the system to a state where most peers share files, despite given initial conditions that are supposed to drive the system to a state where no peer shares.

- Cyber Physical Systems [13]

  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 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 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 has given us the opportunity 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[12].

In 2015, we will continue working on the hybrid protocols and on the UAV platform, and apply our co-simulation work to Smart Grids and other Smart*.

MAESTRO Project-Team

6. New Results

6.1. Highlights of the Year

E. Altman has received the "Isaacs' Award" granted by the International Society on Dynamic Games in recognition for his research on dynamic game theory.

M. El Chamie got the Best Session Presentation Award at the IEEE American Control Conference ACC 2014 for the paper "Newton's method for constrained norm minimization and its application to weighted graph problems," co-authored with G. Neglia.

THANES is a new French-Brazilian joint-team between MAESTRO and researchers from Univ. Federal do Rio de Janeiro (Brazil) and Carneggie Mellon Univ. (USA). The team investigates network science problems with a particular focus on Online Social Networks.

BEST PAPERS AWARDS :

[43] **6th IEEE INFOCOM International Workshop on Network Science for Communication Networks (NetSciCom)**. K. AVRACHENKOV, P. BASU, G. NEGLIA, B. RIBEIRO, D. TOWSLEY.

[70] **4th IEEE Online Conference on Green Communications (GreenComm)**. C. ROTTONDI, G. NEGLIA, G. VERTICALE.

6.2. Network Science

Participants: Eitan Altman, Konstantin Avrachenkov, Mahmoud El Chamie, Julien Gaillard, Arun Kadavankandy, Jithin Kazhuthuveettil Sreedharan, Hlib Mykhailenko, Philippe Nain, Giovanni Neglia, Yonathan Portilla, Alexandre Reiffers, Vikas Singh, Marina Sokol.

6.2.1. Epidemic models of propagation of content

Epidemic models have received significant attention in the past few decades to study the propagation of viruses, worms and ideas in computer and social networks. In the case of viruses, the goal is to understand how the topology of the network and the properties of its nodes impact the spread of the epidemics. In [38], E. Altman, A. Avritzer and L. Pfleger de Aguiar (Siemens Corporation, Princeton, USA), R. El-Azouzi (Univ. of Avignon), and D. S. Menasche (Federal Univ. of Rio de Janeiro, Brazil) propose rejuvenation as a way to cope with epidemics. Reformatting a computer may solve the problem of virus contamination (but it might be a costly operation) while less dramatic actions may render the computer operational again (even in the presence of the virus). In this work they evaluate the performance gain of such measures as well as sampling for early detection of viruses while these incubate. During incubation, contaminated terminals are infectious and yet, if not detected to be so, they cannot be isolated and treated.

In [60], Y. Hayel (Univ. of Avignon), S. Trajanovski and P. Van Mieghem (Delft Univ. of Technology, The Netherlands), E. Altman, and H. Wang (Delft Institute of Applied Mathematics, The Netherlands), compare solutions involving vaccination to those that involve healing from a selfish point of view of an individual networked user. A game theoretical model is presented and the obtained equilibrium is computed for various types of topologies including the fully connected one, the bipartite graph and a community structure. A novel use of potential games is presented to compute the equilibria.

In [61], L. Maggi and F. De Pellegrini (CREATE-NET, Italy), A. Reiffers, J. J. Herings (Maastricht Univ., The Netherlands) and E. Altman, study a viral diffusion of a content in a multi-community environment. Exploiting time scale separation, the authors are able to reduce the dimensionality of the problem and to compute its limiting behavior in closed form. They further study regulation and cooperative approaches for sharing the cost for fighting the spread of the infection among the communities.

Social networks can have asymmetric relationships. In the online social network Twitter, a follower receives tweets from a followed person but the followed person is not obliged to subscribe to the channel of the follower. Thus, it is natural to consider the dissemination of information in directed networks. In [44], K. Avrachenkov in collaboration with B. Prabhu (LAAS-CNRS), K. De Turck and D. Fiems (Ghent Univ., Belgium) use the mean-field approach to derive differential equations that describe the dissemination of information in a social network with asymmetric relationships. In particular, their model reflects the impact of the degree distribution on the information propagation process. They further show that for an important subclass of their model, the differential equations can be solved analytically.

6.2.2. Bio-Inspired Models for Characterizing YouTube Viewcount

Bio-inspired models have long been advocated for the dissemination of content in the Internet. How good are such models and how representative are they? In [69], C. Richier, R. El-Azouzi, T. Jimenez, G. Linares (all with Univ. of Avignon), E. Altman and Y. Portilla propose six different epidemic models. These are classified according to various criteria: (i) the size of the target population, which may be constant, or linearly increasing or infinite, (ii) the virality of the content: it is said to be viral if nodes that receive the content participate in retransmitting it (by sharing or embedding). They then collected data on the viewcounts of videos in youtube and examined how well they fit their models. They showed that their six models cover 90% of the videos with an average mean square error of less than 5%. They further studied the capability of using these models to predict the evolution of the viewcount.

6.2.3. Network centrality measures

Finding quickly top-k lists of nodes with the largest degrees in large complex networks is a basic problem of recommendation systems. If the adjacency list of the network is known (not often the case in complex networks), a deterministic algorithm to solve this problem requires an average complexity of $O(n)$, where n is the number of nodes in the network. Even this modest complexity can be excessive for large complex networks. In [18], K. Avrachenkov and M. Sokol in collaboration with N. Litvak (Twente Univ., The Netherlands) and D. Towsley (Univ. of Massachusetts, Amherst, USA) propose to use a random-walk-based method. They show theoretically and by numerical experiments that for large networks, the random-walk method finds good-quality top lists of nodes with high probability and with computational savings of orders of magnitude. They also propose stopping criteria for the random-walk method that requires very little knowledge about the structure of the network.

In [46], K. Avrachenkov in collaboration with N. Litvak (Twente Univ., the Netherlands) and L. Ostroumova and E. Suyargulova (both from Yandex, Russia) address the problem of quick detection of high-degree entities in large online social networks. The practical importance of this problem is attested by a large number of companies that continuously collect and update statistics about popular entities, usually using the degree of an entity as an approximation of its popularity. They suggest a simple, efficient, and easy to implement two-stage randomized algorithm that provides highly accurate solutions for this problem. For instance, their algorithm needs only one thousand API requests in order to find the top-100 most followed users in Twitter, a network with approximately a billion of registered users, with more than 90% precision. Their algorithm significantly outperforms existing methods and serves many different purposes, such as finding the most popular users or the most popular interest groups in social networks. They show that the complexity of the algorithm is sublinear in the network size, and that high efficiency is achieved in networks with high variability among the entities, expressed through heavy-tailed distributions.

Personalized PageRank is an algorithm to classify the importance of web pages on a user-dependent basis. In [48], K. Avrachenkov and M. Sokol in collaboration with R. van der Hofstad (EURANDOM, The Netherlands) introduce two generalizations of Personalized PageRank with node-dependent restart. The first generalization is based on the proportion of visits to nodes before the restart, whereas the second generalization is based on the proportion of time a node is visited just before the restart. In the original case of constant restart probability, the two measures coincide. They discuss interesting particular cases of restart probabilities and restart distributions. They show that both generalizations of Personalized PageRank have an elegant expression

connecting the so-called direct and reverse Personalized PageRanks that yield a symmetry property of these Personalized PageRanks.

Along with K. Avrachenkov and N. M. Markovich (Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia), J. K. Sreedharan investigated distribution and dependence of extremes in network sampling processes [47]. This is one of the first studies associating extremal value theory to sampling of large networks. The work showed that for any general stationary samples from the graph (function of node samples) meeting two mixing conditions, the knowledge of bivariate distribution or bivariate copula is sufficient to derive many of its extremal properties. The work proved the usage of a single parameter to find many relevant extremes in networks like order statistics, first hitting time, mean cluster size etc. In particular, correlation in degrees of adjacent nodes are modelled and different random walks, such as PageRank, are studied in detail. This work has been done in the context of Inria Alcatel-Lucent Bell Labs joint laboratory's ADR "Network Science" (see §7.1.2).

6.2.4. Influence maximization in complex networks

Efficient marketing or awareness-raising campaigns seek to recruit a small number, w, of influential individuals—where w is the campaign budget—that are able to cover the largest possible target audience through their social connections. In [43] K. Avrachenkov and G. Neglia in collaboration with P. Basu (BBN Technologies, US), B. Ribeiro (CMU, US) and D. Towsley (Univ. of Massachusetts, Amherst, USA) assume that the topology is gradually discovered thanks to recruited individuals disclosing their social connections. They analyze the performance of a variety of online myopic algorithms (i.e. that do not have a priori information on the topology) currently used to sample and search large networks. They also propose a new greedy online algorithm, Maximum Expected Uncovered Degree (MEUD). Their proposed algorithm greedily maximizes the expected size of the cover, but it requires the degree distribution to be known. For a class of random power law networks they show that MEUD simplifies into a straightforward procedure, denoted as MOD because it requires only the knowledge of the Maximum Observed Degree. This work has been done in the context of THANES Joint team (see §8.3.1.1) and Inria Alcatel-Lucent Bell Labs joint laboratory's ADR "Network Science" (see §7.1.2).

In [66] G. Neglia, in collaboration with X. Ye (Politecnico di Torino, Italy), M. Gabielkov and A. Legout (from the DIANA team) consider how to maximize users influence in Online Social Networks (OSNs) by exploiting social relationships only. Their first contribution is to extend to OSNs the model of Kempe, Kleinberg and Tardös on the propagation of information in a social network and to show that a greedy algorithm is a good approximation of the optimal algorithm that is NP- hard. However, the greedy algorithm requires global knowledge, which is hardly practical. Their second contribution is to show on simulations on the full Twitter social graph that simple and practical strategies perform close to the greedy algorithm.

6.2.5. Clustering

Clustering of a graph is the task of grouping its nodes in such a way that the nodes within the same cluster are well connected, but they are less connected to nodes in different clusters. In [45] K. Avrachenkov, M. El Chamie and G. Neglia propose a clustering metric based on the random walks' properties to evaluate the quality of a graph clustering. They also propose a randomized algorithm that identifies a locally optimal clustering of the graph according to the metric defined. The algorithm is intrinsically distributed and asynchronous. If the graph represents an actual network where nodes have computing capabilities, each node can determine its own cluster relying only on local communications. They show that the size of clusters can be adapted to the available processing capabilities to reduce the algorithm's complexity.

6.2.6. Average consensus protocols

In [54], [82], M. El Chamie in collaboration with J. Liu and T. Başar (Univ. of Illinois at Urbana Champaign, USA) studies the performance of a subclass of distributed averaging algorithms where the information exchanged between neighboring nodes (agents) is subject to deterministic uniform quantization. They give the convergence properties of linear averaging due to such quantization (which is a practical concern for many applications) that cause nonlinearity in the system. This is the first attempt to solve the exact model.

In [53], M. El Chamie in collaboration with T. Başar (Univ. of Illinois at Urbana Champaign, USA) considers optimal design strategies in consensus protocols for networks vulnerable to adversarial attacks. They provide a game theoretical model for the problem of a network with an adversary corrupting the control signal with noise. They derive the optimal strategies for both players (the adversary and the network designer) of the resulting game using a saddle point equilibrium solution in mixed strategies.

6.3. Wireless Networks

Participants: Eitan Altman, Abdulhalim Dandoush, Majed Haddad, Jithin Kazhuthuveettil Sreedharan.

6.3.1. 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. In collaboration with M. S. Elgamel (Univ. of Louisiana, USA), A. Dandoush has revised the Trigonometric based Ad-hoc Localization System (TALS) proposed in the literature. In [83], they propose a new technique to optimize the system: by eliminating the need of solving a linear system of equations via least square methods or its variants or the need for any square root operations, the computational overhead is reduced. Also, a novel modified Manhattan distance is proposed and used in the elimination process ensuring thereby a very good accuracy with less complexity than the basic TALS. Through a mathematical analysis and intensive simulations, the optimized TALS is shown to present superior performance and accuracy results compared to other localization techniques.

6.3.2. Channel management

The enhanced Inter Cell Interference Coordination (eICIC) feature has been introduced to solve the interference problem in small cells. It involves two parameters which need to be optimized, namely the Cell Range Extension (CRE) of the small cells and the ABS ratio (ABSr) which defines a mute ratio for the macro cell to reduce the interference it produces. In [72], A. Tall, Z. Altman (Orange Labs, Issy les Moulineaux) and E. Altman propose self-optimizing algorithms for the eICIC. The CRE is adjusted by means of a load balancing algorithm. The ABSr parameter is optimized by maximizing a proportional fair utility of user throughputs. The convergence of the algorithms is proven using Stochastic Approximation theorems. Numerical simulations illustrate the important performance gain brought about by the different algorithms.

Cognitive Radios are proposed as a solution to scarcity of wireless spectrum and one of the main challenges here is to gain knowledge about the spectrum usage by the licensed users, termed as spectrum sensing. In [29], Vinod Sharma (Indian Institute of Science, Bangalore, India) and J. K. Sreedharan study novel algorithms for spectrum sensing which minimize the expected time for spectrum sensing with stringent constraints on the probability of wrong detection. Algorithms are distributed in nature and the work proves that the algorithms are asymptotically optimal distributed sequential hypothesis tests. Along with theoretical guarantees, many practical scenarios in Cognitive Radios are also investigated.

6.3.3. Self-Organizing Network (SON)

The fast development of SON technology in mobile networks renders critical the problem of coordinating SON functionalities operating simultaneously. SON functionalities can be viewed as control loops that may need to be coordinated to guarantee conflict free operation, to enforce stability of the network and to achieve performance gain. In [30], A. Tall and Z. Altman (Orange Labs, Issy les Moulineaux), R. Combes (SUPELEC), and E. Altman propose a distributed solution for coordinating SON functionalities. It uses Rosen's concave games framework in conjunction with convex optimization. The SON functionalities are modeled as linear Ordinary Differential Equation (ODE)s. The stability of the system is first evaluated using a basic control theory approach together with strict diagonal concavity notion that originates from game theory. The coordination solution consists in finding a linear map (called coordination matrix) that stabilizes the system of SON functionalities. It is proven that the solution remains valid in a noisy environment using Stochastic Approximation.

6.4. Network Engineering Games

Participants: Eitan Altman, Ilaria Brunetti, Majed Haddad, Alexandre Reiffers.

6.4.1. The association problem

In [57], M. Haddad, S. Habib (Orange Labs, Issy les Moulineaux), and P. Wiecek (Wroclaw Univ. of Technology, Poland) and E. Altman develop a hierarchical Bayesian game framework for automated dynamic offset selection. Users compete to maximize their throughput by picking the best locally serving radio access network (RAN) with respect to their own measurement, their demand and a partial statistical channel state information of other users. In particular, they investigate the properties of a Stackelberg game, in which the base station is a player on its own. They derive analytically the utilities related to the channel quality perceived by users to obtain the equilibria. They study the Price of Anarchy of such system, which is defined as the ratio of the social welfare attained when a network planner chooses policies to maximize social welfare versus the social welfare attained at a Nash/Stackeleberg equilibrium when users choose their policies strategically.

6.4.2. Cognitive radio

In [26], M. Haddad, P. Wiecek (Wroclaw Univ. of Technology, Poland), O. Habachi and Y. Hayel (both with Univ. of Avignon) propose a game theoretical approach that allows cognitive radio pairs, namely the primary user (PU) and the secondary user (SU), to update their transmission powers and frequencies simultaneously. Specifically, a Stackelberg game model in which individual users attempt to hierarchically access to the wireless spectrum while maximizing their energy efficiency was addressed. A thorough analysis of the existence, uniqueness and characterization of the Stackelberg equilibrium was conducted. In particular, it was shown that a spectrum coordination naturally occurs when both actors in the system decide sequentially about their powers and their transmitting carriers. As a result, spectrum sensing in such a situation turns out to be a simple detection of the presence/absence of a transmission on each sub-band. An algorithmic analysis on how the PU and the SU can reach such a spectrum coordination using an appropriate learning process is provided.

In [59], the same authors present a hierarchical game to model distributed joint power and channel allocation for multi-carrier energy efficient cognitive radio systems. A thorough analysis of the existence, uniqueness and characterization of the Stackelberg equilibrium is conducted. It was proved that, at the Stackelberg equilibrium, each of the two users transmits on only one carrier depending on the fading channel gains. This results contrast with capacity-based approaches in which a certain number of carriers is exploited depending on the channel gains. Interestingly, it was shown that, for the vast majority of cases, introducing a certain degree of hierarchy in a multi-carrier system induces a natural coordination pattern where users have incentive to choose their transmitting carriers in such a way that they always transmit on orthogonal channels. Analytical results were provided for assessing and improving the performances in terms of energy efficiency between the non-cooperative game with synchronous decision makers and the proposed Stackelberg game.

6.4.3. Routing Games

In [39], E. Altman, J. Kuri (Indian Institute of Science, Bangalore, India) and R. El-Azouzi (Univ. of Avignon) study a routing game that models competition over a simple network with losses. Packets may be lost in the network due to either congestion losses or to channel random losses. They compute the equilibrium and establish its properties. They identify a Braess type paradox in which by adding a link the loss probabilities of all players increase.

G. Accongiagioco (Institute for Advanced Studies, Lucca, Italy), E. Altman, E. Gregori (Italian National Research Council, Italy) and L. Lenzini (Univ. of Pisa, Italy) analyze in [36] the decisions taken by an Autonomous System (AS) when joining the Internet. They first define a realistic model for the interconnection costs incurred and then they use this cost model to perform a game theoretic analysis of the decisions related to the creation of new links in the Internet. The proposed model does not fall into the standard category of routing games, hence they devise new tools to solve it by exploiting peculiar properties of the game. They prove analytically the existence of multiple equilibria for specific cases, and provide an algorithm to compute the stable ones. The analysis of the model's outcome highlights the existence of a Price of Anarchy and a Price of Stability.

6.4.4. Network neutrality and collusion

Representatives of several Internet access providers have expressed their wish to see a substantial change in the pricing policies of the Internet. In particular, they would like to see content providers pay for use of the network, given the large amount of resources they use. This would be in clear violation of the "network neutrality" principle that had characterized the development of the wireline Internet. In [14], E. Altman, M. K. Hanawal (former PhD student in MAESTRO) and R. Sundaresan (Indian Institute of Science, Bangalore, India) proposed and studied possible ways of implementing such payments and of regulating their amount. The results were reported already in a previous report, but were substantially revised during the period of this project.

6.4.5. Competition over popularity in social networks

We have pursued our analysis of competition over popularity and visibility in social networks. In [68], A. Reiffers and E. Altman, together with Y. Hayel (Univ. of Avignon) study a game model that arises when the rate of transmission of packets of each source can be accelerated in order to optimize a weighted sum of its acceleration cost and the expected number of its contents on the timelines of those who follow that content. While this paper considers equilibrium within static policies (in which the acceleration rate does not change in time), the same authors study in [51] the structure of dynamic equilibrium policies which are allowed to change as a function of the time (or of the state). A problem with a similar tradeoff is studied by E. Altman in a mobile context in [13] where the question of accelerating the transmission rate of content arises in a context of competition over content where it is assumed that if a content reaches a given destination then that destination will not be interested any more in receiving competing content.

In [67], A. Reiffers, E. Altman and Y. Hayel (Univ. of Avignon) extend the work in [68], and model the situation in which several social networks are available and a source may control not only the rate of transmission (acceleration) but may also decide how to split its content to the various social networks.

A competition over the timing of the transmission of a content was studied by E. Altman and N. Shimkin (Israel Institute of Technology, Israel) in [41]. Uniqueness of a symmetric equilibrium was established under the assumption of Poisson arrival of requests.

6.5. Green Networking and Smart Grids

Participants: Sara Alouf, Eitan Altman, Alberto Benegiamo, Ioannis Dimitriou, Majed Haddad, Alain Jean-Marie, Giovanni Neglia.

6.5.1. Energy efficiency in wireless networks

In [25], M. Haddad, P. Wiecek (Wroclaw Univ. of Technology, Poland), O. Habachi and Y. Hayel (both with Univ. of Avignon) investigated the achievable performances of multi-carrier energy efficient power control game. Both the simultaneous-move and the hierarchical games were addressed. For the first time, the analytical closed-form expressions of the spectrum coordination and the spectral efficiency of such models was derived. Results indicate that the spectrum coordination capability induced by the power control game model enables the wireless network to enjoy the energy efficiency improvement while still achieving a high spectral efficiency.

In [58], the same authors studied energy efficiency of heterogeneous networks for both sparse and dense (two-tier and multi-tier) small cell deployments. The problem is formulated as a hierarchical (Stackelberg) game in which the macro cell is the leader whereas the small cell is the follower. Both players want to strategically decide on their power allocation policies in order to maximize the energy efficiency of their registered users. A backward induction method has been used to obtain a closed-form expression of the Stackelberg equilibrium. It was shown that the energy efficiency is maximized when only one sub-band is exploited for the players of the game depending on their fading channel gains.

In [34], R. A. Vaca Ramirez and J. S. Thompson (Univ. of Edinburgh, UK), E. Altman and V. M. Ramos Ramos (Univ. Autonoma Metropolitana, Mexico) aim to reduce the power expenditure in the reverse link during low network load periods, by allocating extra resource blocks (RBs) to the mobile users. This is in contrast with other approaches in which resources are reduced in hours of low energy consumption. The user's rate demands are split among its allocated RBs in order to transmit in each of them by using a low level modulation order. In this low SINR regime the transmission is much more energy efficient since the log appearing in Shannon formula is in close to linear. We model the bandwidth expansion (BE) process by a game theory framework derived from the concept of stable marriage with incomplete lists (SMI).

P. Wiecek (Wroclaw Univ. of Technology, Poland) and E. Altman consider in [42] dynamic Multiple Access games between a random number of players competing over collision channels. Each of several mobiles involved in an interaction determines whether to transmit at a high or at a low power. High power decreases the lifetime of the battery but results in smaller collision probability. They formulated this game as an anonymous sequential game with undiscounted reward and computed the equilibrium [42]. The internal state of a player corresponds to the amount of energy left in the battery and the actions correspond to the transmission power.

I. Dimitriou investigated in [52] the power management of mobile devices, using a variant of an M/G/1 queue with probabilistic inhomogeneous multiple vacations and generalized service process. Under the vacation scheme, at the end of a vacation the server goes on another vacation, with a different probability distribution, if during the previous vacation there have been no arrivals. The modified vacation policy depends on the initial vacation interval and the server selects randomly over M such vacation policies. The theoretical system can be applied for modeling the power saving mode of mobile devices in modern wireless systems. Moreover, the form of the service process properly describes the incremental redundancy retransmission scheme that provides different types of retransmissions in such systems. Steady state analysis is investigated, energy and performance metrics are obtained and used to provide numerical results that are also validated against simulations.

6.5.2. Energy efficiency in delay tolerant networks

Energy efficiency in mobile networks is further studied in [28] where L. Sassatelli (Univ. of Nice Sophia Antipolis), A. Ali, M. Panda and T. Chahed (all with Telecom SudParis) and E. Altman tackle the issue of reliable transport in Delay-Tolerant mobile ad hoc Networks, that are operated by some opportunistic routing algorithm. We propose a reliable transport mechanism that relies on Acknowledgements (ACK) and coding at the source. The various versions of the problem depending on buffer management policies are formulated, and a fluid model based on a mean-field approximation is derived for the designed reliable transport mechanism. This model allows to express both the mean file completion time and the energy consumption up to the delivery of the last ACK at the source.

6.5.3. Modeling of a smart green base station

S. Alouf, I. Dimitriou A. Jean-Marie have considered the modeling of wireless communication base stations with autonomous energy supply (solar, wind). They proposed and analyzed a queueing model to assess performance of a base station fully powered by renewable energy sources. The system operates in a finite state Markovian random environment that properly describes the intermittent nature of renewable energy sources and the data traffic. The base station is considered to be "smart" in the sense that it is able to dynamically adjust its coverage area, controlling thereby the traffic rate and its energy consumption. They show how the matrix-analytic formalism enables to construct and study the performance of a smart green base station operating in random environment. More precisely, the behavior of such a system is described by a five-dimensional Markov process, which is a homogeneous finite Quasi Birth-Death (QBD) process. Several existing algorithms can be used in order to obtain the stationary probability vector, which is the basis for the calculation of interesting performance metrics. This work is on-going and has not been submitted for publication yet.

6.5.4. Direct Load Control

Balancing energy demand and production is becoming a more and more challenging task for energy utilities also because of the larger penetration of renewable energies which are more difficult to predict and control.

While the traditional solution is to dynamically adapt energy production to follow the time-varying demand, a new trend is to drive the demand itself. Most of the ongoing actions in this direction involve greedy energy consumers, like industrial plant, supermarkets or large buildings. Pervasive communication technologies may allow in the near future to push further the granularity of such approach, by having the energy utility interacting with residential appliances. In [65] and in its extension [64], G. Neglia, in collaboration with G. Di Bella, L. Giarré and I. Tinnirello (Univ. of Palermo, Italy) study large scale direct control of inelastic home appliances whose energy demand cannot be shaped, but simply deferred. Their solution does not suppose any particular intelligence at the appliances. The actuators are rather smart plugs (simple devices with local communication capabilities that can be inserted between appliances plugs and power sockets) and are able to interrupt/reactivate power flow through the plug. A simple control message can be broadcast to a large set of smart plugs for probabilistically enabling or deferring the activation requests of a specific load type in order to satisfy a probabilistic bound on the aggregated power consumption. The control law and the most important performance metrics can be easily derived analytically.

6.5.5. Charge of Electric Vehicles

The massive introduction of Electric Vehicles (EVs) is expected to significantly increase the power load experienced by the electrical grid, but also to foster the exploitation of renewable energy sources: if the charge process of a fleet of EVs is scheduled by an intelligent entity such as a load aggregator, the EVs' batteries can contribute in flattening energy production peaks due to the intermittent production patterns of renewables by being recharged when energy production surpluses occur. To this aim, time varying energy prices are used, which can be diminished in case of excessive energy production to incentivize energy consumption (or increased in case of shortage to discourage energy utilization). In [70] G. Neglia, in cooperation with C. Rottondi and G. Verticale (Politecnico di Milano, Italy), evaluate the complexity of the optimal scheduling problem for a fleet of EVs aimed at minimizing the overall cost of the battery recharge in presence of time-variable energy tariffs. The scenario under consideration is a fleet owner having full knowledge of customers' traveling needs at the beginning of the scheduling horizon. They prove that the problem has polynomial complexity, provide complexity lower and upper bounds, and compare its performance to a benchmark approach which does not rely on prior knowledge of customers' requests, in order to evaluate whether the additional complexity required by the optimal scheduling strategy w.r.t. the benchmark is worthy the achieved economic advantages. Numerical results show considerable cost savings obtained by the optimal scheduling strategy.

6.6. Content-Oriented Systems

Participants: Sara Alouf, Eitan Altman, Konstantin Avrachenkov, Nicaise Choungmo Fofack, Abdulhalim Dandoush, Majed Haddad, Alain Jean-Marie, Philippe Nain, Giovanni Neglia, Marina Sokol.

6.6.1. Modeling modern DNS caches

N. Choungmo Fofack and S. Alouf have pursued their study of the modern behavior of DNS (Domain Name System) caches. The entire set of traces collected in 2013 by Inria's IT services in Sophia Antipolis at one of the Inria's DNS caches have been processed and analyzed with the help of N. Nedkov (4-month intern in MAESTRO). This allowed to strengthen the validation of the theoretical models developed in 2013 (see [86]). On the other hand, parts of [86] have been revisited and derived under more general assumptions. As a direct consequence, the exact analysis derived on linear cache networks is extended to a large class of hierarchical cache networks called *linear-star* networks which include linear and two-level tree/star networks. In addition, closed-form expressions for the cache consistency measures (refresh rate and correctness probability) are provided under the assumption that contents requests and updates occur according to two independent renewal processes.

6.6.2. Analysis of general and heterogeneous cache networks

There has been considerable research on the performance analysis of *on-demand* caching replacement policies like Least-Recently-Used (LRU), First-In-First-Out (FIFO) or Random (RND). Much progress has been made

on the analysis of a single cache running these algorithms. However it has been almost impossible to extend the results to networks of caches. In [22], N. Choungmo Fofack, P. Nain and G. Neglia, in collaboration with D. Towsley (Univ. of Massachusetts, Amherst, USA), introduce a Time-To-Live (TTL) based caching model, that assigns a timer to each content stored in the cache and redraws it every time the content is requested (at each hit/miss). They derive the performance metrics (hit/miss ratio and rate, occupancy) of a TTL-based cache in isolation fed by stationary and ergodic request processes with general TTL distributions. Moreover they propose an iterative procedure to analyze TTL-based cache networks under the assumptions that requests are described by *renewal processes* (that generalize Poisson processes or the standard IRM assumption). They validate the theoretical findings through event-driven and Monte-Carlo simulations based on the Fourier Amplitude Sensitivity Test to explore the space of the input parameters. The analytic model predicts remarkably well all metrics of interest with relative errors smaller than 1%.

Jointly with M. Badov, M. Dehghan, D. L. Goeckel and D. Towsley (all with the Univ. of Massachusetts, Amherst, USA), N. Choungmo Fofack proposes in [81] approximate models to assess the performance of a cache network with arbitrary topology where nodes run the Least Recently Used (LRU), First-In First-Out (FIFO), or Random (RND) replacement policies on arbitrary size caches. The authors take advantage of the notions of "cache characteristic time" and "Time-To-Live (TTL)-based cache" to develop a unified framework for approximating metrics of interest of interconnected caches. This approach is validated through event-driven simulations, and when possible, compared to the existing *a-NET* model.

6.6.3. Data placement and retrieval in distributed/peer-to-peer systems

Distributed systems using a network of peers have become an alternative solution for storing data. These systems are based on three pillars: data fragmentation and dissemination among the peers, redundancy mechanisms to cope with peers churn and repair mechanisms to recover lost or temporarily unavailable data. In previous years, A. Dandoush, S. Alouf and P. Nain have studied the performance of peer-to-peer storage systems in terms of data lifetime and availability using the traditional redundancy schemes. This work has now been published in [23].

A. Jean-Marie and O. Morad (Univ. Montpellier 2) have proposed a control-theoretic model for the optimization of prefetching in the context of hypervideo or, more generally, connected documents. The user is assumed to move randomly from document to document, and the controller attempts at downloading in advance the documents accessed. A penalty is incurred when the document is not completely present. The model is flexible in the sense that it allows several variants for the network model and the cost metric [63]. They have proposed exact algorithms and heuristics for the solution of this problem, and compared them on a benchmark of different user behaviors [62].

The question of whether it is possible to prefetch documents so that the user never experiences blocking, has been modeled with a "cops-and-robbers" game jointly with F. Fomin (Univ. Bergen), F. Giroire and N. Nisse (both from Inria project-team COATI) and D. Mazauric (former PhD student in MAESTRO and MASCOTTE) [24] (see also MAESTRO's 2011 activity report).

6.6.4. Streaming optimization

In streaming applications such as youtube, packets have to be played at the destination at the same rate they were created. If a packet is not available at the destination when it has to be played then a starvation occurs. This results in an unpleasant frozen screen and in an interruption in the video. To decrease the probability of a starvation the destination first waits till it has received some target number of packets and only then starts to play them. In [32], E. Altman and M. Haddad together with Y. Xu (Fudan Univ. China), R. El-Azouzi and T. Jimenez (Univ. of Avignon), and S.-E. Elayoubi (Orange Labs, Issy les Moulineaux) compute the starvation probability as a function of the initial buffering and study tradeoffs between the two performance measures: starvation probabilities and the pre-buffering delay.

6.6.5. Stochastic geometry and network coding for distributed storage

In [37] E. Altman and K. Avrachenkov in collaboration with J. Goseling (Twente Univ., The Netherlands) consider storage devices located in the plane according to a general point process and specialize the results for

the homogeneous Poisson process. A large data file is stored at the storage devices, which have limited storage capabilities. Hence, they can only store parts of the data. Clients can contact the storage devices to retrieve the data. The expected costs of obtaining the complete data under uncoded or coded data allocation strategies are compared. It is shown that for the general class of cost measures where the cost of retrieving data is increasing with the distance between client and storage devices, coded allocation outperforms uncoded allocation. The improvement offered by coding is quantified for two more specific classes of performance measures. Finally, the results are validated by computing the costs of the allocation strategies for the case that storage devices coincide with currently deployed mobile base stations.

6.7. Advances in Methodological Tools

Participants: Eitan Altman, Konstantin Avrachenkov, Ilaria Brunetti, Ioannis Dimitriou, Mahmoud El Chamie, Majed Haddad, Alain Jean-Marie, Philippe Nain, Giovanni Neglia.

6.7.1. Queueing theory

In [21] K. Avrachenkov and P. Nain in collaboration with U. Yechiali (Tel Aviv Univ., Israel) study a retrial queueing system with two independent Poisson streams of jobs flowing into a single-server service system, having a limited common buffer that can hold at most one job. If a type-i job ($i = 1, 2$) finds the server busy, it is blocked and routed to a separate type-i retrial (orbit) queue that attempts to re-dispatch its jobs at its specific Poisson rate. This creates a system with three dependent queues. Such a queueing system serves as a model for two competing job streams in a carrier sensing multiple access system. They study the queueing system using multi-dimensional probability generating functions, and derive its necessary and sufficient stability conditions while solving a Riemann-Hilbert boundary value problem. Various performance measures are calculated and numerical results are presented. In particular, numerical results demonstrate that the proposed multiple access system with two types of jobs and constant retrial rates provides incentives for the users to respect their contracts.

In [19] K. Avrachenkov in collaboration with E. Morozov (Petrozavodsk State Univ., Russia) consider a finite buffer capacity GI/GI/c/K-type retrial queueing system with constant retrial rate. The system consists of a primary queue and an orbit queue. The primary queue has c identical servers and can accommodate up to K jobs (including c jobs under service). If a newly arriving job finds the primary queue to be full, it joins the orbit queue. The original primary jobs arrive to the system according to a renewal process. The jobs have i.i.d. service times. The head of line job in the orbit queue retries to enter the primary queue after an exponentially distributed time independent of the length of the orbit queue. Telephone exchange systems, medium access protocols, optical networks with near-zero buffering and TCP short-file transfers are some telecommunication applications of the proposed queueing system. The model is also applicable in logistics. They establish sufficient stability conditions for this system. In addition to the known cases, the proposed model covers a number of new particular cases with the closed-form stability conditions. The stability conditions that they obtained have clear probabilistic interpretation.

In [20] K. Avrachenkov in collaboration with E. Morozov and R. Nekrasova (Petrozavodsk State Univ., Russia) and B. Steyaert (Ghent Univ., Belgium) study a retrial queueing system with N classes of customers, where a class-i blocked customer joins orbit i. Orbit i works like a single-server queueing system with (exponential) constant retrial time (with rate μ_{0i}) regardless of the orbit size. Such a system is motivated by multiple telecommunication applications, for instance wireless multi-access systems, and transmission control protocols. First, they present a review of some corresponding recent results related to a single-orbit retrial system. Then, using a regenerative approach, they deduce a set of necessary stability conditions for such a system. They will show that these conditions have a very clear probabilistic interpretation. They also performed a number of simulations to show that the obtained conditions delimit the stability domain with a remarkable accuracy, being in fact the (necessary and sufficient) stability criteria, at the very least for the 2-orbit M/M/1/1-type and M/Pareto/1/1-type retrial systems that they focus on.

In [75], I. Dimitriou investigates a single server system accepting two types of retrial customers and paired services. The service station can handle at most one customer, and if upon arrival a customer finds the server busy it is routed to an infinite capacity orbit queue according to its type. Upon a service completion epoch, if at least one orbit queue is non-empty, the server seeks to find customers from the orbits. If both orbit queues are non-empty, the seeking process will bring to the service area a pair of customers, one from each orbit. If only one is non-empty, then a customer from this orbit queue will be brought to the service area. However, if a primary customer arrives during the seeking process it will occupy the server immediately. It is shown that the joint stationary orbit queue length distribution at service completion epochs, can be determined via transformation to a Riemann boundary value problem. Stability condition is investigated, while an extension of the model is also discussed and analyzed. Numerical results are obtained and yield insight into the behavior of the system. The theoretical system can be used to model a relay node for two connections in wireless communication, where network coding is used.

When individuals have to take a decision on whether or not to join a queue, one may expect to have threshold equilibria in which customers join the queue if its size is smaller than a threshold and do not join if it exceeds the threshold. In [74], P. Wiecek (Wroclaw Univ. of Technology, Poland), E. Altman and A. Ghosh (Univ. of Pennsylvania, USA) have studied queueing in which the congestion cost per user decreases in the queue size. An example for such a situation is multicast communication where all individuals that participate in the multicast session share the transmission cost. They showed that many equilibria exist and computed the asymptotic system behavior as the arrival rate of individuals grows.

6.7.2. Markov processes

In [16] K. Avrachenkov in collaboration with A. Eshragh (Univ. of Adelaide, Australia) and J. Filar (Flinders Univ., Australia) present some algebraic properties of a particular class of probability transition matrices, namely, Hamiltonian transition matrices. Each matrix P in this class corresponds to a Hamiltonian cycle in a given graph G on n nodes and to an irreducible, periodic, Markov chain. They show that a number of important matrices traditionally associated with Markov chains, namely, the stationary, fundamental, deviation and the hitting time matrix all have elegant expansions in the first $n-1$ powers of P, whose coefficients can be explicitly derived. They also consider the resolvent-like matrices associated with any given Hamiltonian cycle and its reverse cycle and prove an identity about the product of these matrices. As an illustration of these analytical results, they exploit them to develop a new heuristic algorithm to determine a non-Hamiltonicity of a given graph.

6.7.3. Control theory

In [17] K. Avrachenkov and O. Habachi (former post-doc in MAESTRO) in collaboration with A. Piunovskiy and Y. Zhang (both from the 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 non-trivially 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.

6.7.4. Game theory

6.7.4.1. Estimating the Shapley-Shubik index

In [15] K. Avrachenkov in collaboration with L. Cottatellucci (EURECOM) and L. Maggi (CREATE-NET, Italy) consider simple Markovian games, in which several states succeed each other over time, following an exogenous discrete-time Markov chain. In each state, a different simple static game is played by the same set of players. They investigate the approximation of the Shapley-Shubik power index in simple Markovian games (SSM). They prove that an exponential number of queries on coalition values is necessary for any deterministic algorithm even to approximate SSM with polynomial accuracy. Motivated by this, they propose and study three

randomized approaches to compute a confidence interval for SSM. They rest upon two different assumptions, static and dynamic, about the process through which the estimator agent learns the coalition values. Such approaches can also be utilized to compute confidence intervals for the Shapley value in any Markovian game. The proposed methods require a number of queries, which is polynomial in the number of players in order to achieve a polynomial accuracy.

6.7.4.2. Evolutionary games

Evolutionary games attempt to explain the evolution of species and the dynamics of competition. The player's utility is called "fitness" and a larger fitness indicates a larger rate of reproducibility. In standard evolutionary games, one studies interactions between individuals each of which is consider as a player. In [49], I. Brunetti, E. Altman, and R. El-Azouzi (Univ. of Avignon) argue that in many situations both in biology as well as in networking, one cannot attribute a fitness to an individual but rather to a group of individuals that behaves as an altruistic entity. For example, in a hive of bees it is only the queen that reproduces and thus one cannot model a single bee as a selfish player. They present new definitions for evolutionary games for such situations and study their equilibrium.

This, as well as other considerations in multi-population evolutionary games, is applied in [56] by H. Gaiech and R. El-Azouzi (Univ. of Avignon), M. Haddad, E. Altman and I. Mabrouki (Univ. of Manouba, Tunisia) to Multiple Access Control for which the equilibrium is explicitly computed.

In [84] E. Altman presents a summary of the foundations of classical evolutionary games addressed to a wide public. Both the equilibrium notion of ESS (Evolutionary Stable Strategy) as well as the replicator dynamics (which describes the non-equilibrium behavior) are presented.

6.7.4.3. Sequential Anonymous Games

Stationary anonymous sequential games are a special class of games that combines features from both population games (infinitely many players) with stochastic games. It allows studying competition in complex systems where each individual belongs to a community (which we call individual state) which may change in time as a result of actions taken by the individual. Unlike standard evolutionary games, a player does not just optimize its immediate reward (fitness) but some long term reward over the time. P. Wiecek (Wroclaw Univ. of Technology, Poland) and E. Altman proved in [42] the existence of an equilibrium for the general model and studied the two applications. The first one is described in §6.5.1 .

The second application is a maintenance repair problem: each of a large number of cars can decide whether to behave gently or to drive fast. By driving fast it takes larger risks for having an accident. The probability of an accident depends on the fraction of drivers that drive fast. An internal state of the car is either good (g) or bad (b). A car gets to a state b as a result of an accident and then it has some penalty and costs for repair. The advantage of driving fast is reducing delay costs. This problem is formulated as a sequential anonymous game and its equilibrium is computed. They computation makes use of the linear structure of both the transition probabilities and the immediate fitness in the global state.

6.7.5. Optimization

In [55] M. El Chamie and G. Neglia provide a methodology for solving smooth norm optimization problems under some linear constraints using the Newton's method. This problem arises in many machine learning and graph optimization applications. They show how Newton's method significantly outperforms gradient methods both in terms of convergence speed and in term of robustness to the step size selection.

MUSE Team

6. New Results

6.1. Pinpointing Home and Access Network Delays Using WiFi Neighbors

Participants: Lucas Di Cioccio (LIP6/Technicolor), Martin May (Technicolor), Jim Kurose (University of Massachusetts, Amherst), Renata Teixeira

Home Internet users and Internet access providers need tools to assist them in diagnosing and troubleshooting network performance problems. Today, expert users may rely on simple techniques using round-trip-measurements to local and remote points to locate delays on an end-to-end path. Unfortunately, round-trip measurements do not provide accurate diagnoses in the presence of asymmetric link capacities and performance, which is often the case in residential access. Our work [8] introduces *neighbor-assisted delay diagnosis* (NADD) - an approach for pinpointing the location of delays (among the home, access, and wide-area network), leveraging end-host multi-homing capabilities. NADD runs on an end host connected simultaneously to the home gateway and to a neighbor WiFi access point. Our evaluation shows that NADD efficiently detect and distinguish uplink and downlink delays with small error. In addition, we learn from a proof-of-concept deployment in five homes in France that our techniques can work "in the wild." Technicolor filed a patent on this work [8].

6.2. Locating Throughput Bottlenecks in Home Networks

Participants: Srikanth Sundaresan (ICSI), Nick Feamster (Princeton), Renata Teixeira

We developed *WTF (Where's The Fault?)* [4], a system that localizes performance problems in home and access networks. We implement WTF as custom firmware that runs in an off-the-shelf home router. WTF uses timing and buffering information from passively monitored traffic at home routers to detect both access link and wireless network bottlenecks. The Federal Communication Commission (FCC) in the United States deployed WTF in 3000 homes for a few days in November 2014. We are currently analyzing the resulting dataset to help shed light on common pathologies that occur in home networks.

6.3. Measuring the Performance of User Traffic in Home Wireless Networks

Participants: Srikanth Sundaresan (ICSI), Nick Feamster (Princeton), Renata Teixeira

This work [5] studies how home wireless performance characteristics affect the performance of user traffic in real homes. Previous studies have focused either on wireless metrics exclusively, without connection to the performance of user traffic; or on the performance of the home network at higher layers. In contrast, we deploy a passive measurement tool on commodity access points to correlate wireless performance metrics with TCP performance of user traffic. We implement our measurement tool, deploy it on commodity routers in 66 homes for one month, and study the relationship between wireless metrics and TCP performance of user traffic. We find that, most of the time, TCP flows from devices in the home achieve only a small fraction of available access link throughput; as the throughput of user traffic approaches the access link throughput, the characteristics of the home wireless network more directly affect performance. We also find that the 5 GHz band offers users better performance better than the 2.4 GHz band, and although the performance of devices varies within the same home, many homes do not have multiple devices sending high traffic volumes, implying that certain types of wireless contention may be uncommon in practice.

6.4. Characterizing Bufferbloat and its Impact at End-hosts

Participants: Stephane Wustner, Jaideep Chandrashekar (Technicolor), Renata Teixeira

While, on routers and gateways, buffers on forwarding devices are required to handle bursty Internet traffic, overly large or badly sized buffers can interact with TCP in undesirable ways. This phenomenon is well understood and is often called "bufferbloat". Although a number of previous studies have shown that buffering (particularly, in home) can delay packets by as much as a few seconds in the worst case, there is less empirical evidence of tangible impacts on end-users. In [3], we develop a modified algorithm that can detect bufferbloat at individual end-hosts based on passive observations of traffic. We then apply this algorithm on packet traces collected at 55 end-hosts, and across different network environments. Our results show that 45 out of the 55 users we study experience bufferbloat at least once, 40% of these users experience bufferbloat more than once per hour. In 90% of cases, buffering more than doubles RTTs, but RTTs during bufferbloat are rarely over one second. We also show that web and interactive applications, which are particularly sensitive to delay, are the applications most often affected by bufferbloat.

6.5. Measuring and Characterising User Online Activity

Participants: Omayma Belkadi, Mauricio Santoro, Anna-Kaisa Pietilainen, Renata Teixeira

The goal of our work is to identify what people are doing online (or the online user activity) from passively collected network traffic traces. Our analysis of network traffic and application information from 12 end-hosts shows that this task is challenging because there are often many applications running on each user's device, whereas the user is only interacting with one application at a time. Our work with two master students presents the first evaluation of the set of features computable from network traffic alone that can help distinguish user activity traffic from all other traffic flows [6], [10]. We obtain ground truth on user activities and network traffic traces in a controlled setting, and complement this dataset with traces collected by the HostView monitoring tool on the devices of 12 users over several months. We develop simple heuristics to extract user activities for the HostView dataset based on the foreground application and on keyboard/mouse activity. Then, we analyze which network traffic features allow us to distinguish between online user activity and background network traffic. Features related to traffic volumes and timings show the most significant differences.

6.6. WeBrowse: a Passive Content Curation System Based on HTTP Logs

Participants: Giuseppe Scavo, Zied Ben Houidi (Alcatel-Lucent), Renata Teixeira, Stefano Traverso (Politecnico di Torino), Marco Mellia (Politecnico di Torino)

Content curation refers to the act of assisting users to identify relevant and interesting information in the overwhelming amount of online content available today. Existing curation services rely either on experts or on crowdsourcing to promote content. This work designs, implements, and evaluates WeBrowse, the first passive crowdsourced content curation system. WeBrowse requires no active user engagement to promote content. Instead, it extracts the URLs users visit from traffic traversing an ISP network to identify popular and interesting content. A key challenge to design such a passive curation system is to process network traffic in real-time to identify the small set of URLs that are interesting to users. WeBrowse contains a set of heuristics to identify the set of URLs users visit and to select the subset that are interesting, while preserving their privacy at the same time. We prototype WeBrowse and evaluate it using traces collected at a large European ISP, and in a deployment in a large campus network. We have tested and improved WeBrowse with a small number of users from September 2014 to January 2015. The plan is to announce WeBrowse to all users of the campus network early 2015 to get feedback on their experience with the system.

Available at: http://tstat.polito.it/netcurator/

RAP Project-Team

4. New Results

4.1. Random Graphs

Participants: Nicolas Broutin, Henning Sulzbach.

4.1.1. Universality of scaling limits of random graphs

Random graphs are one of the most studied models of networks, and they turn out to be related to crucial questions in physics about the behaviour of matter at the phase transition, or in combinatorial optimization about the hardness of computation. In recent years, we have constructed the scaling limit of the classical Erdos-Renyi random graph model, and conjectured that this limit also happened to be universal.

The funding of the Associated Team RNA has permitted to invite Shankar Bhamidi. During his visit, we have worked and found a new way to construct the scaling limit of random graph processes in the critical window. This method is especially important since it is robust enough to prove universality of the limit, that is that many models have the same limit. The method relies on the dynamics of the coalescence of clusters as the edges are added, and allows us to hope for proofs that would be able to treat the more complex geometric models.

4.1.2. Cutting down random tree and the genealogy of fragmentations

The study of the internal structure of random combinatorial object such as graphs and trees led to question about whether such objects exhibit invariance by certain complex surgical operations (disconnect some pieces, and re-attach them somewhere else). In the context of graphs, this is related to the so-called self-organized criticality: certain distributions that yield fractal objects should naturally appear in nature because they are the fixed points of some recombination procedures. In the context of trees, it turns out that certain fragmentations arising when chopping off a random tree have a genealogy that has the same distribution as the original tree. We have investigated this with Minmin Wang, and obtained results about p-trees and the genealogy of the fragmentation on Aldous' celebrated continuum random tree. These may also be interpreted in terms of complex path transformations for Brownian excursions and other random processes with exchangeable increments, and hence relate to very classical questions in probability theory.

4.1.3. New encodings for combinatorial coalescent processes

In 2013, we had constructed the scaling limit of the minimum spanning tree of a complete graph using crucial information about the scaling limit of random graphs, and especially about the way the cluster merge as the edges are added in the graph. With J.-F. Marckert (LaBRI, Bordeaux) we have found a novel construction of the important multiplicative coalescent that describes how the connected components of a random graph coalesce as the edges are added. This unveils yet more interesting links between the minimum spanning tree and the random graph, since Prim's celebrated algorithm is used to construct a consistent ordering of the vertices that ensures that the connected components are intervals.

4.1.4. Navigation in random Delaunay triangulations

Navigation or routing algorithms are fundamental routines: in order to solve many problems, one of the first steps consists in locating a node in a data structure. Unfortunately, the current algorithms are based on heuristics and very few rigorous results about the performance of such algorithms are known when the model for data is more realistic than the worst-case.

With O. Devillers and R. Hemsley, we have initiated a program that aims at finding rigourous estimates for the performance of routing algorithms in geometric structures such as Delaunay tesselations. So far we have managed to develop some tools that permitted us to analyse a simple algorithm. Although this algorithm has been designed for most of the analysis to work, this work paves the way towards the rigorous analysis of other more natural and widely used algorithms.

4.1.5. Connectivity and sparsification of sparse wireless networks

Many models of wireless networks happen to be connected only when the average degree is tending to infinity with the size of the network, more precisely when it is about the logarithm of the number of nodes. This raises questions about the potential issues in scaling such models. With L. Devroye (McGill) and G. Lugosi (ICREA and Pompeu Fabra), we have worked at analysing models in which we try to construct connected or almost connected networks in a distributed way (that is that no global optimization is allowed in designing the network, and every device should proceed in the same way to choose its neighbors). We have managed to analyse an algorithm for constructing such a network, and to obtain tight results about the number of links that a typical device should have in order for the global network to be connected. We further proved that this is asymptotically optimal when one only requires that most nodes should be in the same connected component.

4.2. Resource Allocation Algorithms in Large Distributed Systems

Participants: Christine Fricker, Philippe Robert, Guilherme Thompson.

This is a collaboration with Fabrice Guillemin from Orange Labs which started in February 2014.

4.2.1. Controlling impatience in cellular networks using QoE-aware radio resource allocation

Impatience of users when using a data service has a major impact on the quality of service offered by telecommunication networks, especially in cellular networks with scarce radio resources. Impatience is negative for users, it is due to many factors related to the performance of servers, customer devices, etc., but also to bandwidth sharing in the network.

While impatience can be seen as a negative phenomenon, it can also be used as a lever to discourage customers when the system becomes too much overloaded. This can be achieved in cellular networks by modulating the capacity available to customers being at a certain distance of the antenna. This general idea can be applied in several manners and can be viewed as a network optimization mechanism. In this paper, we reuse the general framework of α-fair scheduler in order to perform this control. This has the advantage of being easy to implement in realistic settings as α-fair schedulers (and especially the Proportional Fair (PF) one) are widely adopted in mobile networks. This also reduces the dimension of our problem as it narrows the optimization problem to the tuning of a single parameter α.

In order to achieve this goal, we first derive a model for reneging probabilities under a general α-fair scheduler. In particular, we consider a heavy load regime and develop a fluid flow analysis of impatience in cellular networks. We notably establish a fixed point formulation for the computation of the reneging probability and introduce a new metric, namely QoE perturbation, expressing how much a particular flow impacts the reneging probability in the system. We then use this QoE perturbation metric to design of a new radio resource management scheme that controls the parameter of the scheduler in order to reduce the global reneging in the system. For instance, recognizing that customers far from the base station degrade the global performance of the system, impatience and α-fair scheduling can be used to discourage those customers and in some sense to perform an implicit admission control in order to optimize the use of radio resources.

4.2.2. Resource Allocation in Large Data Centers

The goal of this study is to investigate the design of allocation algorithms of requests requiring different classes of quality of video streams as well as their performances. The class of algorithms considered may downgrade the quality of some of the transmission to maximize the utilization of the servers.

4.3. Stochastic networks: large bike sharing systems

Participants: Christine Fricker, Hanène Mohamed, Cédric Bourdais, Yousra Chabchoub.

Vehicle sharing systems are becoming an urban mode of transportation, and launched in many cities, as Velib' and Autolib' in Paris. One of the major issues is the availability of the resources: vehicles or free slots to return them. These systems became a hot topic in Operation Research and now the impact of stochasticity on the system behavior is commonly admitted. The problem is to understand their behavior and how to manage them in order to provide both resources to users.

Our stochastic model is the first studying the impact of the finite number of spots at the stations on the system behavior.

With Danielle Tibi, we use limit local theorems to obtain the asymptotic stationary joint distributions of several node (station or route) states when the system is large (both numbers of stations and bikes), also in the case of finite capacities of the stations. This gives an asymptotic independence preperty for node states. This widely extends the existing results on heterogeneous bike-sharing systems.

Second we investigate the impact of finite capacity of stations and reservation in car-sharing systems. The large-scale asymptotic joint stationary distribution of the numbers of vehicles and reserved parking places is given as the joint distribution in a tandem of queues with a constrained total capacity where rates are solutions of a system of two fixed point equations. Analytical expressions are given for performance in light and heavy traffic cases. As expected, reservation impact drastically increases with traffic. Even if the equilibrium is identified and analyzed, the question of convergence is still open.

JC Decaux provides us data describing Velib' user trips. These data are useful to measure the system behavior. With Yousra Chabchoub, we test clustering to obtain a typology of the stations. Then we focus on the resources availability (free docks and available bikes) and separate the Velib' stations into three clusters (balanced, overloaded and underloaded stations), using Kmeans clustering algorithm, along with the Dynamic Time Wraping (DTW) metric. We choose to update the centers of the clusters using the efficient Dtw Barycenter Averaging (DBA) method.

4.4. Scaling Methods

Participants: Philippe Robert, Wen Sun, Mohammadreza Aghajani.

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. A fluid scaling analysis of such a network is presented. We have shown that the interaction of several time scales plays an important role in the evolution of such a system, in particular its coordinates may live on very different time and space scales. As a consequence, the associated stochastic processes turn out to have unusual scaling behaviors which give an interesting fairness property to this class of algorithms. A heavy traffic limit theorem for the invariant distribution has also been proved. A generalization to the resource sharing algorithm for which the $\log$ function is replaced by an increasing function. 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

The Distributed Hash Table (DHTs) consists of a large set of nodes connected through the Internet. Each file contained in the DHT is stored in a small subset of these nodes. Each node breaks down periodically and it is necessary to have back-up mechanisms in order to avoid data loss. 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.

4.5. Stochastic Models of Biological Networks

Participants: Renaud Dessalles, Sarah Eugene, Emanuele Leoncini, 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 procaryots 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. All molecules are supposed to move freely into the cytoplasm, it is assumed that the the encounter time between a given entity and its target is exponentially distributed.

4.5.1.1. Transcription-translation model for all proteins

The first model that has been studied integrates the production of all the proteins. Each gene has to be transcribed in mRNA and each mRNA has to be translated in protein. The transcription step needs a RNA-Polymerase molecule that is sequestered during the time of elongation. Likewise, each mRNA needs a ribosome in order to produce a protein. RNA-Polymerases/Ribosomes are present in limited amount and the genes/mRNAs sequester these molecules during the whole the time of elongation. Finally each mRNA has an exponentially distributed lifetime with an average value of 4 min and the proteins disappear at a rate of one hour, hence simulating the global dilution in the growing bacteria.

This global sharing of Ribosomes/RNA-Polymerases among all proteins induces a general regulation: each gene competing to each other to have access to these common resources. Because of the parameters of affinity (between gene and RNA-Poymerase and between mRNA and ribosome) are specific to each gene, it allows a large range of average protein production but induce some noise, especially for highly expressed proteins.

We developed a Python simulation, and using the biological experiments of Tanichuchi et al. (2010), and we have investigated a biologically coherent range of parameters. By making the simulations, we have been able to reproduce certain aspects of the biological measures, especially for the high amount of noise for well expressed proteins.

4.5.1.2. Simple 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 try to determine if, for instance, for the same average protein level, the feedback permits a noise reduction of protein distribution compared to the "open loop" model; or if it rather allows a better efficiency in case of a change of command for a new level of production (due, for example, to a radical change in the environment) by reducing the respond time to reach this new average.

4.5.2. Stochastic Modelling of Protein Polymerization

This is a collaboration with Marie Doumic, Inria MAMBA team.

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 a translation.

The first study we did proposed a simplified stochastic model to analyze this phenomenon. For this model, when the volume gets large, the quantity of polymers has the typical sigmoidal shape. A second order result has also been obtained for this model. We were able to compute the asymptotic distribution of the lag time and express its variance. The parameters of the model have been obtained by using data given by Wei-Feng Xue, University of Kent.

The current project concerns a more sophisticated mathematical model. Indeed, we have added a conformation step: before polymerizing, 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 follows the conformation step: this modelling leads to the study of averaging principles.

SOCRATE Project-Team

6. New Results

6.1. Highlights of the Year

6.1.1. FIT/CortexLab Inauguration

FIT(Future Internet of Things) is a french Equipex (Équipement d'excellence) which aims to develop an experimental facility, a federated and competitive infrastructure with international visibility and a broad panel of customers. FIT is be composed of four main parts: a Network Operations Center (NOC), a set of Embedded Communicating Object (ECO) test-beds, a set of wireless OneLab test-beds, and a cognitive radio test-bed (CorteXlab) deployed by the Socrate team in the Citi lab. In 2014 the construction of the room was finished see Figure 5 . SDR nodes have installed in the room, 42 industrial PCs (Aplus Nuvo-3000E/P), 22 NI radio boards (USRP) and 18 Nutaq boards (PicoSDR, 2x2 and 4X4) can be programmed from internet now.

A very successfully inauguration took place on the 28th October 2014 [0], with the noticable venue of Vincent Poor, Dean of School of Engineering and Applied Science of Princeton University.

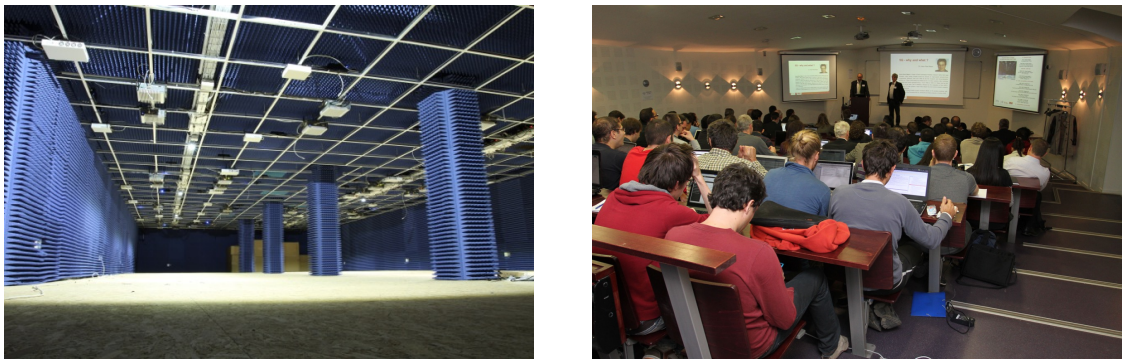

Figure 5. Photo of the FIT/CortexLab experimentation room installed and a snaptshot of the inauguration meeting

6.2. Flexible Radio Front-End

The innovative Wake-Up radio architecture proposed by the Socrate team, based on a classical WiFi standard with a specific OFDM pattern, has been deeply studied in theory and simulations [1], [25], [24]. Great enhancements on the sensitivity study, the choice of identifiers and the comparison of the energy consumption relative to classical systems have led to the development of a first prototype (ongoing work).

6.2.1. Wake-Up Radios

The innovative Wake-Up radio architecture proposed by the Socrate team, based on a classical WiFi standard with a specific OFDM pattern, has been deeply studied in theory and simulations [HUTU-JWCN][KHOUMERI-ECUMICT][HUTU-RWS]. Great enhancements on the sensitivity study, the choice of identifiers and the comparison of the energy consumption relative to classical systems have led to the development of a first prototype (ongoing work).

[0] http://www.inria.fr/centre/grenoble/actualites/inauguration-reussie-de-la-plateforme-cortexlab-equipex-fit

6.2.2. Full-Duplex systems

In the development of wideband OFDM Full-Duplex systems, [33] proposes an analysis of the impact of the thermal noise on the quality of the self-interference cancellation in such systems. A method is proposed to reduce the impact on the bit-error-rate by increasing the level of certain parts of the preamble in each frame. [35] add to the analog RF cancellation proposed previously a stage of digital cancellation enabling to increase more the performance of Full-Duplex terminals.

Furthermore, [34] extend the study to a dualband Full-Duplex systems, enabling the very promising combination of Full-Duplex and carrier aggregation. The proposed structure being sensitive to IQ impairments, a digital mitigation algorithm is also designed.

6.2.3. SDR for SRD

In collaboration with Orange labs, [32] analyses the requirements of an SDR gateway for urban networks collecting SRD (short range devices) information. This study is particularly focused on the ADC resolution, showing that the required resolution in realistic scenarios is too high, therefore emphasizing the need to develop specific hardware techniques.

6.2.4. Experimental Facilities

For the development of the CorteXlab testbed, lots of radio hardware and propagation constraints had to be taken into account [15], [14]. Moreover,[36] had proposed a first implementation of Full-Duplex on USRPs which is expected to be deployed on this tesbed.

Another testbed dedicated to the measurement of the energy consumption of radio devices was also designed and implemented.

6.3. Agile Radio Resource Sharing

This axis addresses the challenges relative to the network perspective of software radio. While the two other axes work on the design of the software radio nodes, we focus herein on their coexistence in a multi-user communications perspective. We are first interested in theoretical limits of some reference scenarios where trade-offs between spectral efficiency, energy efficiency, stability and/or fairness are analyzed. Our research activities are further driven by applicative frameworks. We focused on radio access networks with new results on energy efficiency-spectral efficiency trade-off in LTE networks and multi-band CSMA strategies in Wifi networks. We also studied pure random access and success probabilities for the challenging ultra-narrow band (UNB) technology of SigFox. Lot of efforts has been put on body area networks [8] with deep studies on positioning strategies and distributed decisions and information gathering. As mentioned above, our research follows three objectives:

- Establishing theoretical limits of cooperative wireless networks in the network information theory framework.
- Designing MAC procedures, coding and signal processing techniques for optimal transmissions (e.g. interference alignment).
- Developing distributed mechanisms for distributed decision at layer 1 and 2, using game theory, consensus and graph modeling.

6.3.1. Theoretical limits from information theory

The group strengthened his activities from a formal perspective in the framework of network information theory as initiated with the recruitment of Samir Perlaza and the sabbatical of Jean-Marie Gorce at Princeton University in the group of Prof. H. Vincent Poor. The first scenario is devoted to cellular networks with a random distribution on base stations. The main contribution concerns the broadcast channel (BC) generalized to a continuum of users. The second scenario concerns the interference channel (IC) and the main contribution is relative to the characterization of the Nash stable region for the interference channel with noisy feedback.

6.3.1.1. Broadcast channel with a continuum of users in a typical cell

The theoretical Energy efficiency-Spectral efficiency Pareto optimal front in a typical cell has been evaluated by associating stochastic geometry (Poisson point processes, PPP) and information theory.[21], the broadcast channel is extended to a continuum of users. We derived the theoretical uniform achievable rate with superposition coding principles. We show the potential gain of superposition coding techniques compared to the conventional time sharing. These results are however limited to Gaussian channels and the extension to the vector Gaussian channel is still under investigation. The PPP modeling for multi-cells has been also introduced as well as the price of interference management.

6.3.1.2. Interference Channel with feedback

The decentralized interference channel (DIC) with noisy feedback has been analyzed. In [31], all the rate-pairs that are achievable at a Nash equilibrium (NE) in the two-user linear deterministic symmetric decentralized interference channel (LD-S-DIC) with noisy feedback are identified. A second result provides closed form expressions for the PoA, which allows the full characterization of the reduction of the sum rate due to the anarchic behavior of all transmitter-receiver pairs. The price of anarchy (PoA) and the price of stability (PoS) of the game in which transmit-receiver pairs seek an optimal individual transmission rate are fully characterized in [9]. In particular, it is shown that in all interference regimes, there always exists at least one Pareto optimal Nash equilibrium (NE).

6.3.2. Coding, signal processing and MAC procedures for optimal transmissions

6.3.2.1. Implementation

While theoretical studies provide interesting insights about potential gain and limits of cognitive networks, the achievable efficiency may depend on practical issues related to quantization, synchronization and real-time processing limits. We developed the CortexLab facility offering a reproducible environment for fostering the validation of cooperative communication schemes. The first demo has been presented at the Infocom conference [28] and also at the Melbourne Greentouch meeting. We also contributed to the implementation and analysis of a cognitive transceiver for opportunistic networks [ref Maso JWCN]. The work first focused on a previously introduced dynamic spectrum access (DSA) - cognitive radio (CR) solution for primary-secondary coexistence in opportunistic orthogonal frequency division multiplexing (OFDM) networks, called cognitive interference alignment (CIA). The implementation is based on software-defined radio (SDR) and uses GNU Radio and the universal software radio peripheral (USRP) as the implementation toolkit. The proposed flexible transceiver architecture allows efficient on-the-fly reconfigurations of the physical layer into OFDM, CIA or a combination of both.

6.3.2.2. Interference alignment

In the framework of Greentouch, we studied interference alignment as a mean for improving the EE-SE tradeoff in cellular networks [43]. We combined theoretical studies with stochastic geometry and simulations to show the potential interest. We are also developing a demo with CorteXlab enhancing the IA capability from a real perspective.

6.3.2.3. Multiband MAC

In collaboration with CEA-Leti, we studied MAC strategies for multiband systems. The main idea is based on exploiting the multiband system as a slotted Aloha channel for the RTS/CTS initiation but keeping the total band as a whole for data transmission. We proved that this strategy outperforms classical approaches [39], [40], [30].

6.3.2.4. MAC for localization

In the context of the ANR Cormoran project, we account for radiolocation experiments aiming at both indoor navigation and mobility detection applications for Wireless Body Area Networks (WBAN) [7]. We also studied the relation between the MAC protocol and ranging techniques for localization. The impact of mobility on the distance estimation between 2 nodes of a Wireless Body Area Network (WBAN) by comparing the Two-Way Ranging (2WR) and Three-Way Ranging (3WR) protocols has been proposed in [23]. We also investigated the impact of mobility on the Motion Capture applications [22].

6.3.2.5. random access in Ultra-narrow band networks

Ultra narrow band (UNB) transmission is a very promising technology for low-throughput wireless sensor networks. This technology has already been deployed and has proved to be ultra-efficient for point-to-point communications in terms of power-efficiency, and coverage area. We studied the scalability of UNB for a multi-point to point network. In particular, we proposed a new multiple access scheme: random frequency division multiple access (R-FDMA) and studied the impact of the induced interference on the system performance in terms of bit error rate and outage probability [20]. We also analyzed the system performance in terms of bit error rate and outage probability [37].

6.3.3. Distributed decision mechanisms

Distributed decisions appear in many situations in the wireless world. Resource allocation, power management or relaying techniques are all expecting distributed decisions. To avoid strong coordination, distributed mechanisms inspired e.g. by game theory or consensus algorithms are appealing. Some of the results obtained below also rely on information theory but with a more important focus on algorithms and decision processes when several pairs of wireless transceivers are willing to simultaneously transmit in the same environment.

6.3.3.1. Cognitive radio networks

The problem of joint channel selection and power control is analyzed in the context of multiple-channel clustered ad-hoc networks in[ref Rose [3], i.e., decentralized networks in which radio devices are arranged into groups (clusters) and each cluster is managed by a central controller (CC). The problem is modeled by a game in normal form in which the corresponding utility functions are designed for making some of the Nash equilibria (NE) to coincide with the solutions to a global network optimization problem. A second scenario has been considered where multiple source-destination pairs communicate with each other via an energy harvesting relay [5]. The focus was put on the relay's strategies to distribute the harvested energy among the multiple users and their impact on the system performance. Specifically, a non-cooperative strategy that uses the energy harvested from the i-th source as the relay transmission power to the i-th destination is considered first. An auction based power allocation scheme is also proposed to achieve a better tradeoff between system performance and complexity.

6.3.3.2. Distributed decisions and consensus in MANETs

In the large research area of wireless body area networks, cooperative applications involving several users is attracting strong interests. This cooperation may target a simple information exchange or even some cooperative decision such as swarm coordination. We considered in [26]such a swarm of users moving in a common direction and we are interested in the mechanisms allowing to propagate and share some common information. We extend and improve a previous algorithm derived as a max-consensus approach. We describe a complete experimental setup deployed during a real bike race with 200 runners.

6.4. Software Radio Programming Model

6.4.1. Data Flow Programming

Software defined radio (SDR) technology has evolved rapidly and is now reaching market maturity. However, no standard has emerged for programming the new type of machine that will manage the access to the radio channel. Mickaël Dardaillon, Kevin Marquet, Tanguy Risset have been working in collaboration with the CEA LETI on compiling waveform for heterogeneous Multi-processor SoCs . This research leaded to a prototype compiler for the Magali MP-SoC developped in Mickael Dardaillon' PhD thesis (passed in November 2014) which was the first attempt to compile the SPDF format to a real architecture [18], [16], [17]. This study highlighted in particular the fact that SPDF was a good computation model for waveform description langage, easier to compile than dynamic dataflow format.

6.4.2. Non-volatile memory management for ultra low power systems

To enable non-trivial computation on very resource-constrained platforms powered by energy harvested from RF communications, an embedded OS has to save and restore program state to and from non-volatile memory. By doing so, the application program does not lose all progress when power is lost, which happens very often in environmentally-powered systems. This can be achieved [13] thanks to an incremental checkpointing scheme which aims at minimizing the amount of data written to non-volatile memory, while keeping the execution overhead as low as possible.

6.4.3. FPGA-based Implementation of physical Layers for SDR

A VHDL implementation of the three available options of the IEEE 802.15.4 physical layer was developed [29] in the context of FIT/CorteXlab. This parametrized design was validated on a Nutaq platform which combines Xilinx Virtex-6 FPGA and tunable Radio420x RF transceiver. This work participates to the building of an open source hardware SDR library similar to GNU radio but targeted to FPGA-based platforms.

6.4.4. Towards filters and functions computing just right

A FIR filter is specified by its coefficients (real numbers) and its input and output formats. The implementation of a FIR should be as accurate as its output format allows, but no more. This very simple specification enables the automatic construction of FIR filter implementations that are provably accurate at a minimal hardware cost [19]. The corresponding FIR generator is available in FloPoCo.

The fixed-point Atan2 function is very useful to recover the phase of a complex signal. A careful study of three implementation techniques (including a novel one based on two-variable quadratic approximation) shows that, on current FPGAs, the good old CORDIC technique is more efficient than multiplier-based techniques [46].

URBANET Team

6. New Results

6.1. Highlights of the Year

Two scientific results can be distinguished in UrbaNet activity this year. First of all, the work did in collaboration with Orange Labs during the PhD thesis of O. Erdene-Ochir (defended in 2013) led to a patent [38] related to routing in wireless sensor networks under resiliency constraints.

A second important result is represented by the book chapter "Wireless Access Networks for Smart Cities" [31], a common contribution of all the permanent members of the team. We hope that this chapter will become the reference on wireless networking within the new and dynamic smart cities community.

6.2. Characterizing and measuring urban networks

Participants: R. Domga Komguem, M. Fiore, D. Naboulsi, P. Raveneau, R. Stanica, F. Valois

6.2.1. 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, such as network quality or weather forecast applications.

In this sense, we contributed with a novel privacy-preserving mobile data collection platform [21], leveraging the dynamic deployment of crowdsourcing tasks across a pouplation of mobile phones.

Using such data, or other datasets coming from network operators, we can propose dynamic access network mechanisms that adapt to customers' demands. To that end, one must be able to process large amount of mobile traffic data and outline the network utilization in an automated manner. In [28], we propose a framework to analyze broad sets of Call Detail Records (CDRs) so as to define categories of mobile call profiles and classify network usages accordingly. We evaluated our framework on a CDR dataset including more than 300 million calls recorded in an urban area over 5 months. We showed how our approach allows to classify similar network usage profiles and to tell apart normal and outlying call behaviors.

6.2.2. Generation and Analysis of Vehicular Mobility Datasets

The surge in vehicular network research has led, over the last few years, to the proposal of countless network solutions specifically designed for vehicular environments. A vast majority of such solutions has been evaluated by means of simulation, since experimental and analytical approaches are often impractical and intractable, respectively. The reliability of the simulative evaluation is thus paramount to the performance analysis of vehicular networks, and the first distinctive feature that has to be properly accounted for is the mobility of vehicles, i.e., network nodes. Notwithstanding the improvements that vehicular mobility modeling has undergone over the last decade, no vehicular mobility dataset was publicly available that captures both the macroscopic and microscopic dynamics of road traffic over a large urban region.

In [12], we present a realistic synthetic dataset, covering 24 hours of car traffic in a 400-km2 region around the city of Ko¨ln, in Germany. We describe the generation process and outline how the dataset improves the traces currently employed for the simulative evaluation of vehicular networks. We also show the potential impact that such a comprehensive mobility dataset has on the network protocol performance analysis, demonstrating how incomplete representations of vehicular mobility may result in over-optimistic network connectivity and protocol performance.

Moreover, using a similar methodology we contribute to the ongoing effort to define such mobility scenarios by introducing a second set of traces for vehicular network simulation, this time focusing on a highway environment. Our traces are derived from high-resolution real-world traffic counts, and describe the road traffic on two highways around Madrid, Spain, at several hours of different working days. We provide a thorough discussion of the real-world data underlying our study, and of the synthetic trace generation process [20] [35] [29]. Finally, we assess the potential impact of our dataset on networking studies, by characterizing the connectivity of vehicular networks built on the different traces. Our results underscore the dramatic impact that relatively small communication range variations have on the network. Also, they unveil previously unknown temporal dynamics of the topology of highway vehicular networks, and identify their causes.

6.2.3. Characterizing Novel Wireless Networks for Urban Intelligent Transportation Solutions

Vehicular networks are not the only contribution communication technologies can bring in the field of Intelligent Transportation Systems. Two other examples have been studied this year in the team.

The first example is related to traffic light control in an urban environment [17]. A traffic light controller takes as input an estimation of the number of vehicles entering the intersection and produces as output a light plan, with the objective to reduce the traffic jam. The quality of the input traffic estimation is a key consideration on the performance of the traffic light controller. The advent of Wireless Sensor Networks, with their relatively low deployment and operation price, led to the development of several sensor-based architectures for intersection monitoring. We show in this work that the solutions proposed in the literature are unrealistic in terms of communication possibilities and that they do not allow a measure of the vehicular queue length at a lane level. Based on extensive experimental results, we propose an energy efficient, low cost and lightweight multi-hop wireless sensor network architecture to measure with a good accuracy the vehicle queue length, in order to have a more precise vision of traffic at the intersection.

On a second example, these last years have witnessed the rise of the smart cities and several mechanisms to render the cities more sustainable and more energy-efficient. Among all different aspects, a noteworthy one is urban bike development. Besides the growing enthusiast provoked by bicycles and the benefit for health they bring, there still exists some reluctance in using bikes because of safety, road state, weather, etc. To counter-balance these feelings, there is a need to better understand bicycle users habits, path, road utilization rate in order to improve the bicycle path quality. In this perspective, in [25], we propose to deploy a set of mobile sensors on bicycles to gather this different data and to exploit them to make the bike easier and make people want to ride bicycles more often. Such a network will also be useful for several entities like city authorities for road maintenance and deployment, doctors and environment authorities, etc. Based on such a framework, we propose a first basis model that help to dimension the network infrastructure and the kind of data to be real time gathered from bikes. More specifically, we present a theoretical model that computes the quantity of data a bike will be able to send along a travel and the quantity of data a base station should be able to absorb. We have based our study on real data to provide first numerical results and be able to draw some preliminary conclusions and open new research directions.

6.3. Technology specific solutions

Participants: I. Augé-Blum, W. Bechkit, J. Cui, A. Mouradian, T. Lin, H. Rivano, R. Stanica, F. Valois

6.3.1. Medium Access Control in Wireless Sensor Networks

Protocols developed during the last years for Wireless Sensor Networks (WSNs) are mainly focused on energy efficiency and autonomous mechanisms (e.g. self-organization, self-configuration, etc.). Nevertheless, with new WSN applications, new QoS requirements appear, such as time constraints. Real-time applications require the packets to be delivered before a known time bound which depends on the application requirements. We particularly focus on applications which consist in alarms sent to the sink node. We propose Real-Time X-layer Protocol (RTXP) [8], a real-time communication protocol. RTXP is a MAC and routing real-time communication protocol that is not centralized, but instead relies only on local information. To the best of our knowledge, it is the first real-time protocol for WSNs using an opportunistic routing scheme in order

to increase the packet delivery ratio. In the paper above, we describe the protocol mechanisms. We give theoretical bounds on the end-to-end delay and the capacity of the protocol. Intensive simulation results confirm the theoretical predictions and allow to compare RTXP with a real-time scheduled solution. RTXP is also simulated under harsh radio channel, in which case the radio link introduces probabilistic behavior. Nevertheless, we show that RTXP performs better than a non-deterministic solution. It thus advocates for the usefulness of designing real-time (deterministic) protocols even for highly unreliable networks such as WSNs.

Continuing on the idea of WSN applications with strict temporal constraints, these critical applications require correct behavior, reliability, and, of course, the respect of time constraints. Otherwise, if they fail, consequences on human life and the environment could be catastrophic. For this reason, we argue that the WSN protocols used in these applications must be formally verified. Unfortunately the radio link is unreliable, it is thus difficult to give hard guarantees on the temporal behavior of the protocols (on wired systems the link error probability is very low, so they are considered reliable). Indeed, in WSN a message may experience a very high number of retransmissions. The temporal guarantee has thus to be given with a probability that it is achieved. This probability must meet the requirements of the application. Network protocols have been successfully verified on a given network topology without taking into account unreliable links. Nevertheless, the probabilistic nature of radio links may change the topology (links which appear and disappear). Thus, instead of a single topology we have a set of possible topologies, each topology having a probability to exist. In this paper, we propose a method that produces the set of topologies, checks the property on every topology, and gives the probability that the property is verified. This technique is independent from the verification technique, i.e. each topology can be verified using any formal method which can give a “yes” or “no” answer to the question: “Does the model of the protocol respect the property?”. In [27], we apply this method on the previously proposed f-MAC protocol, a real-time medium access protocol for WSNs. We use UPPAAL model checker as verification tool, and we perform simulations to observe the difference between average and worst case behaviors.

One WSN application gaining a lot of importance in the team in the last few years targets Intelligent Transportation Systems (ITS), as also explained in the previous section. In this ITS field, parking sensor networks are rapidly deploying around the world and are also regarded as one of the first implemented urban services in smart cities. To provide the best network performance in this context, the MAC protocol shall be adaptive enough in order to satisfy the traffic intensity and variation of parking sensors. In this sense, in [24] and [36], we compare the performance of two off-the-shelf medium access control protocols on two different kinds of traffic models, and then evaluate their application-end information delay and energy consumption while varying traffic parameters and network density. From the simulation results, we highlight some limits induced by network density and occurrence frequency of event-driven applications. When it comes to real-time urban services, a protocol selection shall be taken into account - even dynamically - with a special attention to the energy-delay trade-off. In a follow-up study [23], we use real world data, more precisely the heavy-tailed parking and vacant time models from the SmartSantander platform, and then we apply the traffic model in the simulation with four different kinds of MAC protocols, that is, contention-based, schedule-based and two hybrid versions of these. The result shows that the packet inter-arrival time is no longer heavy-tailed while collecting a group of parking sensors, and then choosing an appropriate MAC protocol highly depends on the network configuration. Also, the information delay is bounded by traffic and MAC parameters which are important criteria while the timely message is required.

6.3.2. Routing in Wireless Sensor Networks

Routing represents another major challenging issue in WSN, because of the application diversity and energy efficiency constraints. Gradient broadcast routing is a robust scheme for data gathering in WSNs. At each hop, the sender broadcasts the packet to its neighbors and one or more nodes among its neighbors closer to the sink forward it. As long as a node has at least one neighbor with a smaller hop-count, it can route packets. Nevertheless, nodes can disappear because of energy depletion, hardware failure, etc. In this case, it cannot be ensured that a packet reaches the sink. Usually this issue is addressed by updating the gradient with a periodical flooding. Nevertheless, it consumes an important amount of energy, moreover, parts of the network may not need to be updated. In [26], we propose GRABUP (GRAdient Broadcast UPdate), a traffic-based gradient

maintenance algorithm which updates the gradient thanks to the data packets. We simulate the proposition and compare it with the classic gradient broadcast routing.

Another specific application that we target is smart metering, which heavily rely on the communication network for efficient data gathering, thus eliminating manual meter reading. Smart electronic devices are deployed in open, unattended and possibly hostile environment such as consumer's home and office areas, making them particularly vulnerable to physical attacks. Resilience is needed in this case to mitigate such inherent vulnerabilities and risks related to security and reliability. In [18], a general overview of the resilience including definition, metric and resilient techniques relevant for smart metering is presented. A quantitative metric, visual and meaningful, based on the graphical representation is adopted to compare routing protocols in the sense of resilience against active insider attacks. Five well-known routing protocols from the main categories have been studied through simulations and their resilience is evaluated according to the given metric. Resilient techniques introduced to these protocols have enhanced significantly the resilience against attacks providing route diversification.

6.3.3. Other Research Issues Related to Wireless Sensor Networks

Important features of WSNs, such as low battery consumption, changing topology awareness, open environment, non reliable radio links, raise other research issues than classical MAC and routing problems. For example, in [32], we investigate the benefits of Network Coding in WSN, especially with respect to resiliency. We have seen in our previous work that resiliency could be described as a multi dimensional metric, taking parameters such as Average Delivery Ratio, Delay Efficiency, Energy Efficiency, Average Throughput and Delivery Fairness into account. Resiliency can then be graphically represented as a kiviat diagram created by the previous weighted parameters. In order to introduce these metrics, previous works have been leaded on the Random Gradient Based Routing, which proved good resiliency in malicious environment. We look for seeing the improvements in term of resiliency, when adding network coding in the Random Gradient Based Routing with malicious nodes.

Another challenge is represented by the deployment of sensor nodes, which can take into account the impact of multiple parameters. For example, temperature variations have a significant effect on low power WSNs 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, we develop [4] 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. We use 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, 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 identify the optimal operation of each algorithm and show that a significant amount of energy can be saved by taking temperature into account.

Finally, the notion of Shared Risk Link Groups (SRLG) captures survivability issues when a set of links of a network may fail simultaneously, such as a WSN where link conditions are extremely dynamic. 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, for example a battery depleted sensor) and the span 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 [33], we extend these notions to characterize new cases in which these optimization problems can be solved in polynomial time or are fixed parameter tractable. We also investigate on 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.

6.3.4. Data Aggregation and Gathering

In the data gathering problem, a particular network node, the base station or the sink, aims at receiving messages from some other network nodes. In [5], we model this network as a graph, and we consider that, 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 below a certain threshold 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 in this case to the personalized broadcasting problem where the base station has to send messages to some nodes in the graph, with same transmission constraints. We focus on the gathering and personalized broadcasting problem in grids (regular networks, with nodes deployed in a grid-like shape, e.g. parking or intersection monitoring WSNs). 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 optimal schedules for data gathering.

Data aggregation is a particular solution for the data gathering problem, which reduces the amount of data sent to the base station. In [16], we show that data aggregation can effectively reduce the energy consumption and improve the network capacity. Moreover, we present the state-of-the-art aggregation functions, including compressing-based and forecasting-based method; compressing-based aggregation focuses on compressing the data packets accompanied with transmitting based on spatial correlation, while forecasting aggregation tends to use mathematical models to fit the time series and predict the new value due to highly temporal correlation. We detail these two methods and characterize them respectively. We propose comparison between A-ARMA and Compressing Sensing, which are noteworthy examples of forecasting aggregation and compressing aggregation respectively.

6.3.5. Safety Vehicular Ad Hoc Networks

Vehicular ad hoc networks can play an important role in enhancing transportation efficiency and improving road safety. Therefore, direct vehicle-to-vehicle communications are considered as one of the main building blocks of a future Intelligent Transportation System. The success and availability of IEEE 802.11 radios made this technology the most probable choice for the medium access control layer in vehicular networks. However, IEEE 802.11 was originally designed in a wireless local area network context and it is not optimized for a dynamic, ad hoc vehicular scenario. In [11], we investigate the compatibility of the IEEE 802.11 medium access control protocol with the requirements of safety vehicular applications. As the protocols in this family are well-known for their scalability problems, we are especially interested in high density scenarios, quite frequent on today's roads. Using an analytical framework, we study the performance of the back-off mechanism and the role of the contention window on the control channel of a vehicular network. Based on these findings, we propose a reverse back-off mechanism, specifically designed with road safety applications in mind. Extensive simulations are carried out to prove the efficiency of the proposed enhancement scheme and to better understand the characteristics of vehicular communications.

One of the major roles of vehicular communication is the dissemination of information on the road in order to increase the awareness of the drivers. The facilities layer is a recently standardized component in the vehicular communication architecture, with an important role to play in the process of information dissemination. In [22], we propose facilities layer-based mechanisms for information propagation and we show they outperform classical network layer solutions. We also demonstrate that previous studies that do

not consider the cohabitation of different types of safety messages on the vehicular control channel highly under-estimate the dissemination delay, which can lead to unrealistic assumptions in the design of safety applications.

6.4. Capillary solutions

Participants: M. Fiore, G. Gaillard, D. Naboulsi, H. Rivano, R. Stanica, F. Valois

6.4.1. Connected Vehicles

Bandwidth availability in the cellular backhaul is challenged by ever-increasing demand by mobile users. Vehicular users, in particular, are likely to retrieve large quantities of data, choking the cellular infrastructure along major thoroughfares and in urban areas. It is envisioned that alternative roadside network connectivity can play an important role in offloading the cellular infrastructure. We investigate [7] the effectiveness of vehicular networks in this task, considering that roadside units can exploit mobility prediction to decide which data they should fetch from the Internet and to schedule transmissions to vehicles. Rather than adopting a specific prediction scheme, we propose a fog-of-war model that allows us to express and account for different degrees of prediction accuracy in a simple, yet effective, manner. We show that our fog-of-war model can closely reproduce the prediction accuracy of Markovian techniques. We then provide a probabilistic graph-based representation of the system that includes the prediction information and lets us optimize content prefetching and transmission scheduling. Analytical and simulation results show that our approach to content downloading through vehicular networks can achieve a 70% offload of the cellular network.

Vehicles also produce large quantities of Floating Car Data (FCD), which consist of information generated by moving vehicles and uploaded to Internet-based control centers for processing and analysis. As upcoming mobile services based on or built for networked vehicles largely rely on uplink transfers of small-sized but high-frequency messages, FCD traffic is expected to become increasingly common in the next few years. Presently, FCD are managed through a traditional cellular network paradigm: however, the scalability of such a model is unclear in the face of massive FCD upload, involving large fractions of the vehicles over short time intervals. In [13], we explore the use of vehicle-to-vehicle (V2V) communication to partially relieve the cellular infrastructure from FCD traffic. Specifically, we study the performance boundaries of such a FCD offloading approach in presence of best- and worst-case data aggregation possibilities at vehicles. We show the gain that can be obtained by offloading FCD via vehicular communication, and propose a simple distributed heuristic that has nearly optimal performance under any FCD aggregation model.

We also advocate the use of a data shuttle service model to offload bulk transfers of delay-tolerant data from the Internet onto standard vehicles equipped with data storage capabilities [14]. We first propose an embedding algorithm that computes an offloading overlay on top of the road infrastructure. The goal is to simplify the representation of the road infrastructure as raw maps are too complex to handle. In this overlay, each logical link maps multiple stretches of road from the underlying road infrastructure. We formulate then the data transfer assignment problem as a novel linear programming model that determines the most appropriate logical paths in the offloading overlay for a data transfer request. We evaluate our proposal using actual road traffic counts in France. Numerical results show that we can satisfy weekly aggregate requests in the petabyte range while achieving cumulative bandwidth above 10 Gbps with a market share of 20% and only one terabyte of storage per vehicle.

6.4.2. Energy Consumption in Communication Networks

Providing high data rates with minimum energy consumption is a crucial challenge for next generation wireless networks. There are few papers in the literature which combine these two issues. The work we propose in [10] focuses on multi-hop wireless mesh networks using a MAC layer based on S-TDMA (Spatial Time Division Multiple Access). We develop an optimization framework based on linear programming to study the relationship between throughput and energy consumption. Our contributions are twofold. First, we formulate and solve, using column generation, a new MILP to compute offline energy-throughput tradeoff curve. We use a physical interference model where the nodes can perform continuous power control and can use a discrete set

of data rates. Second, we highlight network engineering insights. We show, via numerical results, that power control and multirate functionalities allow optimal throughput to be reached, with lower energy consumption, using a mix of single hop and multihop routes.

Another strategy with regard to energy consumption is switching off some network nodes that are not carrying any data or control traffic. In [37], we tackle the problem of on-grid energy saving in cellular networks based on switch-on/off techniques for base stations and the usage of renewable energy. We aim to evaluate how much power can be saved in the network and dimension the renew able energy system according to the consumptions in real-world networks.

6.4.3. Service Level Agreements

The era of the Internet of Things (IoT) brings complexity and deployment costs in smart cities, particularly in WSNs. Utilities such as gas or water providers are keen on delegating the management of the communications to specialized firms, namely WSN Operators, that will share the WSN resource among their various clients. For this reason, in [34] we provide a guideline to write Service Level Agreements (SLAs) for IoT operation, borrowing a well studied concept from the web services domain. We extend the SLA definition with specific items that integrate the WSN constraints, and we facilitate the construction of complex metrics that express the performance of the WSN.

Furthermore. WSN operators will need a robust and reliable technology in order to guarantee QoS constraints in a wireless environment, as in the industrial world. IEEE 802.15.4e Time Slotted Channel Hopping (TSCH) is one good candidate. Moreover, the IETF experience in IP networks management is an important input for monitoring and QoS control over WSNs. In [19], we give formal guidelines for the implementation of a SLA architecture for operated WSNs. We distinguish the various formal algorithms that are necessary to operate a WSN according to SLAs, and determines which functional entities are necessarily technology-dependent. Detailed examples of such entities are developed in an IPv6 over IEEE 802.15.4e TSCH context, such as advocated in the IETF 6TiSCH Working Group.