### Distributed Systems and Services

<table>
<thead>
<tr>
<th>Project-Team</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACES</td>
<td>5</td>
</tr>
<tr>
<td>ADAM</td>
<td>9</td>
</tr>
<tr>
<td>AMAZONES</td>
<td>11</td>
</tr>
<tr>
<td>ARLES</td>
<td>15</td>
</tr>
<tr>
<td>ASAP</td>
<td>24</td>
</tr>
<tr>
<td>ASCOLA</td>
<td>34</td>
</tr>
<tr>
<td>ATLANMOD</td>
<td>41</td>
</tr>
<tr>
<td>CIDRE</td>
<td>43</td>
</tr>
<tr>
<td>FOCUS</td>
<td>48</td>
</tr>
<tr>
<td>INDES</td>
<td>53</td>
</tr>
<tr>
<td>MYRIADS</td>
<td>58</td>
</tr>
<tr>
<td>OASIS</td>
<td>63</td>
</tr>
<tr>
<td>PHOENIX</td>
<td>67</td>
</tr>
<tr>
<td>POPS</td>
<td>69</td>
</tr>
<tr>
<td>REGAL</td>
<td>77</td>
</tr>
<tr>
<td>RMOD</td>
<td>85</td>
</tr>
<tr>
<td>SARDES</td>
<td>90</td>
</tr>
<tr>
<td>SCORE</td>
<td>97</td>
</tr>
<tr>
<td>TRISKELL</td>
<td>101</td>
</tr>
<tr>
<td>ALGORILLE</td>
<td>104</td>
</tr>
<tr>
<td>CEPAGE</td>
<td>112</td>
</tr>
<tr>
<td>GRAAL</td>
<td>119</td>
</tr>
<tr>
<td>GRAND-LARGE</td>
<td>133</td>
</tr>
<tr>
<td>HIEPACS</td>
<td>143</td>
</tr>
<tr>
<td>KERDATA</td>
<td>147</td>
</tr>
<tr>
<td>MESCAL</td>
<td>154</td>
</tr>
<tr>
<td>MOAIS</td>
<td>158</td>
</tr>
<tr>
<td>RUNTIME</td>
<td>160</td>
</tr>
</tbody>
</table>

### Distributed and High Performance Computing

<table>
<thead>
<tr>
<th>Project-Team</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIONYSOS</td>
<td>166</td>
</tr>
<tr>
<td>DISTRIBUTCOM</td>
<td>173</td>
</tr>
<tr>
<td>DNET</td>
<td>182</td>
</tr>
<tr>
<td>GANG</td>
<td>184</td>
</tr>
<tr>
<td>HIPERCOM</td>
<td>192</td>
</tr>
<tr>
<td>MADYNES</td>
<td>202</td>
</tr>
<tr>
<td>MAESTRO</td>
<td>210</td>
</tr>
<tr>
<td>MASCOTTE</td>
<td>221</td>
</tr>
<tr>
<td>PLANETE</td>
<td>234</td>
</tr>
<tr>
<td>RAP</td>
<td>249</td>
</tr>
</tbody>
</table>

### Networks and Telecommunications
<table>
<thead>
<tr>
<th>No.</th>
<th>Team Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>39.</td>
<td>RESO Project-Team</td>
<td>254</td>
</tr>
<tr>
<td>40.</td>
<td>SWING Team</td>
<td>259</td>
</tr>
<tr>
<td>41.</td>
<td>TREC Project-Team</td>
<td>267</td>
</tr>
</tbody>
</table>
4. New Results

4.1. Spatial Computing approach and RFIDs

Participants: Michel Banâtre, Paul Couderc [contact], Yann Glouche, Arnab Sinha.

In the line of our previous research in pervasive computing, we are working on spatial computing approaches in the context of RFID. Spatial computing consists in data structures and computing processes directly supported by physical objects. RFID is an attractive technology for supporting spatial computing, enabling any object to interact in a smart environment. Traditional RFID solutions use a logical model, where the RFID tags are simple identifiers referring to data in a remote information system. In our approach, we use the memory of the tags to build self-contained data structures and self-describing objects. While featuring interesting properties, such as autonomous operation and high scalability, this approach also raises difficult challenges: the memory capacity of the tags is very limited, requiring compact and efficient data structures. Some results have been achieved for security applications, where we contributed to efficient integrity checking solution for coupled objects. Integrity checking based on objects group can also be used to provide reliable inventory protocol for RFID, which current readers are lacking. A paper has just been submitted on this aspect.

An applicative project (see 5.2.1) in the context of domestic waste management is broadly investigating the use of RFID at item level to provide early waste sorting, to avoid incompatible mix of waste and to prevent hazards. An ontology based system has been proposed to determine the possible interactions of tagged products based on their properties and the external conditions.

4.2. Integrity checking with coupled objects

Participants: Michel Banâtre [contact], Paul Couderc, Jean-Francois Verdonck.

Integrity checking is an important concern in many activities, such as logistic, telecommunication or even day to day tasks such as checking for someone missing in a group. While the computing and telecommunication worlds commonly use digital integrity checking, many activities from the real world do not benefit from automatic integrity control mechanisms. RFID technology offers promising perspectives for facing this problem, but also raises strong privacy concerns as most of the RFID-based systems rely on global identification and tracking. Previously we have already designed Ubi-Check to provide an approach aiming at coupling physical objects and enabling integrity control built on local interactions, without the support of a global information system. Ubi-Check led to the development of various novel applications running quite on the same technology. Most of the partners showed a high interest in defining coupled object-based security solutions, but they were lacking the possibility of defining hierarchical couplings. This is that we have studied and implemented this year. We have designed the Ubi-Tree software which strives to deal with those new requirements.

4.2.1. Hierarchical physical object coupling

Ubi-Tree is a new solution/software designed at INRIA aiming at setting-up and reading hierarchical couplings. It relies on a structure in which physical objects (also called fragments) are seen as external nodes of a tree that we call coupling tree. External nodes of a tree are called leaves. In the system, internal nodes are called coupling nodes. Each fragment embeds an RFID tag supporting coupling data. Coupling data stores the coupling tree. Each internal node can be checked, which means a lacking, illegally forged or corrupted node can be detected at any depth of a coupling. Ubi-Tree proposes a new API to create and check hierarchical couplings and an interactive editing GUI is under development.
New algorithms and structure to store and read hierarchical couplings trees in its leaves (RFID tags) have been developed, making it possible to use multiple coupling levels. Let’s take an example: given three physical objects $o_1, o_2, o_3$. A user can couple $o_1$ and $o_2$ together. Let $N_1$ be the parent coupling of $o_1$ and $o_2$: $N_1 = o_1, o_2$. Then he couples this coupling with $o_3$ to create the $N_2 = o_1, o_2, o_3$ coupling node. Figure 2 gives an illustration of the described coupling.

The key idea of the coupling structure is that coupling data are spread in a way that only descendant leaves of a coupling node are required to read it and process its integrity control. This way, $N_1$ only needs $o_1$ and $o_2$ to be read and / or checked as $o_1$ and $o_2$ are descendants from $N_1$ whereas $o_3$ is not. This choice enables to process integrity controls at multiple coupling levels. It is very convenient, as an example, if $o_3$ is physically separated from $o_1$ and $o_2$. So if $o_1, o_2$ and $o_3$ are brought into the field of the RFID reader, $N_1$ and $N_2$ can be written, read and checked. If $o_3$ is not present in its field, $N_2$ will not be recovered from the read data but $N_1$ can still be read and checked. If checked, $N_1$ will notify it was not the root of the coupling tree when the coupling tree was written. This way, the user knows it did not read the whole structure $o_1$ and $o_2$ are part of.

Indeed, coupling nodes can have the following status:

- **Valid**: the set of detected tags enabled to decode a structure in which the node has the children and the parent it is supposed to have.
- **Partial**: same as the valid status except the node should have a parent that could not be read from detected tags.
- **Incomplete**: the node is missing some of its children.

### 4.2.3. Ongoing work

Today the management algorithms of graphs are located on the memory available RFID reading and writing. However, in the various applications envisaged, only a subset of RFID memory are read / write, others are only accessible in read-only. Currently we are working on the development of our algorithms that take into account this kind of configuration. The other problem we are working on today is the interface provided to users to be sure that the association between RFID tag and physical object is the one that is perceived by our coupling software. The idea is to be able to identify in the right way the RFID tag associated to a physical object when we place one physical object $O$ onto the support of the antenna linked to the RFID reader. The position of $O$, and the tag associated to $O$, in the physical space is determined using a camera coupled with an image recognition algorithm. The result is displayed onto a touch screen. In that way, when we want to couple a set of physical objects $o_1, o_2, ...$ we place sequentially all these objects onto the support of the antenna, and
from the image of these objects displayed onto the touch screen we touch those we want to couple and activate the coupling operation. This solution is under development.

4.3. Pervasive support for Smart Homes

**Participants:** Minh Tuan Ho, Michele Dominici, Bastien Pietropaoli, Frédéric Weis [contact].

Pervasive computing involves tight links between real-world activities and computing processes. While the perception of the real-world events can be handled entirely by the application, we think that ad hoc approaches have limitations, in particular the complexity and the difficulty to re-use the code between applications. Instead, we promote the use of system-level abstraction that leverage on tangible structures and processes. Important properties of this approach is that applications are, by design, operating in an implicit way ("in the background" of physical processes). They also often exhibit simpler architectures, and "natural" scalability in the sense that being build upon existing real-world process, they are strongly distributed design that relies essentially on local interactions between physical entities. We are applying this approach to "Smart Homes". A Smart Home is a residence equipped with computing and information technology devices conceived to collaborate in order to anticipate and respond to the needs of the occupants, working to promote their comfort, convenience, security and entertainment while preserving their natural interaction with the environment.

4.3.1. Definition of a system architecture

In a classical "logical" approach, all the intelligence of the Smart Home is condensed in a single entity that takes care of every device in the house. The sensors distributed in the environment have to send back all the gathered data to the central entity, that takes all the burden of parsing the sensitive information and infer the policies to be implemented. Our architecture is instead focused on a physical approach, where every device carries a part of the global intelligence: every single entity can analyze the part of information sensitive for its goal, derive useful data, and communicate meaningful information to the other devices.

![Four-layer model](image-url)

*Figure 3. Four-layer model*

Our work is based on a four-layer model [8], as showed in Figure 3. The first layer of our system should be simply composed of sensors, but some constraints have to be fitted. In order to reduce the global system cost and to protect the inhabitants' privacy, the number of sensors dispatched in the environment has to be reduced as much as possible. However, a huge number of different sensors are required to sense context pieces and redundancy can significantly increase the reliability of the sources. With this idea in mind, the sensors are grouped in nodes. These nodes are able to preprocess the data with simple computation such as minimum, maximum and average. They also enable the sensors to communicate, using, for instance, 6LowPAN (IPv6 over LoW Power wireless Area Networks).
In the second layer, the raw data are processed to obtain more abstract data about context and occurring situations. It could be, for instance, a presence in a room, the number of people in this room or the posture of someone. The aggregation of raw data is realized thanks to a data fusion algorithm. The one we adopted is called the belief functions theory or theory of evidence \[6\].

The bridge between the second and the third layer is realized integrating the results of sensor data fusion into a context model called *Context Spaces*. This model uses geometrical metaphors to describe context and situations, relying on the following concepts: the context attributes, the application space, the situation spaces and the context state. The context attributes are information types that are relevant and obtainable by the system; in our case, the context attribute values are provided by the perception layer, together with a degree of confidence on them, needed to cope with the intrinsic uncertainty of sensing systems in real world scenarios. In the situation and context identification layer, the context state provided by the perception layer is analyzed to infer the ongoing situation spaces (representing real-life situations) and also produce a measure of confidence in their occurrence. As the same context state can correspond to several different situation spaces (and vice versa), reasoning techniques are needed to discern the actual ongoing real-life situations in spite of uncertainty \[5\].

![System architecture](image)

### 4.3.2. Experimentation

The computations required by the second and the third layers to obtain abstract data and to analyze context and situations are too heavy for our nodes to be processed on. To remedy to this problem, more powerful nodes acting like sinks are used. These nodes are small “plug and play” computers called plug computers. Their role is to gather data from sensor nodes and to perform data fusion, required to produce the context attributes, and context space reasoning, used to identify ongoing situations.

The figure 4 gives an overview of our system architecture. The latter has been demonstrated by ACES team at EDF R&D in November.
6. New Results

6.1. Adaptive Middleware


In 2011, we pursued our goal to demonstrate that general and high level concepts and solutions can be proposed to design multi-scale middleware systems. The multi-scale aspect has particularly been put forward and we obtained several interesting results: we showed that the concepts of service, component, and software architecture can be successfully used, in the small for wireless sensor middleware platforms [19], [43] with applications to the Internet of Things (IoT) [27] and for embedded systems [14], in mid-size distributed environments such as digital home networks [17], and in the large in cloud computing platforms [30]. We focus below on two achievements which are illustrative in the sense that they address both ends of the targeted spectrum of sizes.

At the scale of small systems, we proposed the REMORA platform [19], [43], [27] which defines a lightweight event-based programming model for wireless sensor networks. A C-like language for component implementation and an extension of the state-of-the-art Service Component Architecture (SCA) standard for service-oriented systems are proposed. The platform has been successfully deployed on the Contiki operating system. We showed that despite the characteristics of such resource-constrained environments, we are still able to obtain reconfigurability and adaptability properties for the deployed systems.

At the scale of very large systems, we showed in [30] first results that illustrate the fact that the FraSCAti platform [18] can be used to achieve interoperability between applications deployed on heterogeneous cloud platforms. The experiment is currently deployed on 13 public IaaS and PaaS cloud infrastructures. The very same concepts of service and software architecture that are used at smaller scales are put into practice here. Furthermore, we benefit from the same adaptability properties to address the heterogeneity of concepts needed to fit these very large scale infrastructures.

6.2. Context-awareness and Ambient Intelligence Software

Participants: Laurence Duchien, Sébastien Mosser, Clément Quinton.

Context-aware applications are applications that can react to changes on their environment. To achieve such reacting behavior, several challenges have to be faced in terms of: context management, support for dynamic reconfiguration, automation of development, and a consistent development process. One possible way to face those challenges is to use the principles of Software Product Line (SPL) and specifically dynamic SPL (DSPL). DSPLs focus on variability management and aim at deriving different products from a same product family. Additionally, DSPLs allow for products to be derived both at design and at runtime. This enables applications to be adapted during execution and dynamically fit new requirements or resource changes. In [16] we have proposed an approach to unify adaptation at design and at runtime based on Aspect Oriented Modeling. Our approach proposes a unified aspect metamodel and a platform that realizes two different weaving processes to achieve design and runtime adaptations. This approach is used in a Dynamic Software Product Line which derives products that can be configured at design time and adapted at runtime in order to dynamically fit new requirements or resource changes. Such products are implemented using the Service Component Architecture and Java. Finally, we have illustrated the use of our approach based on an adaptive e-shopping scenario. This work corresponds to Carlos Parra’s PhD thesis [11] and is partially funded by the CAPPUCINO project. Finally, in [40], we have proposed to develop an application for mobile devices using Software Product Lines (SPL). Considering variation factors, SPL allows the conception and the development of a software products family minimizing realization cost and time. The result is the APPLIDE framework, which provides SPL for smartphones, and we show how it works with a short demonstration.
6.3. SCeSAME: Formal Definition of Software Architecture Adaptation

Participants: Rubby Casallas, Laurence Duchien, Nohra Villegas, Gabriel Tamura.

In order to define properties on adaptation process, we need to formally model the architecture reconfiguration of a component-based (CB) system as an action performed by itself. These actions are performed in response to the disruption of Quality of Service (QoS) contracts, in the spirit of the Effeil’s rescue clause in object-oriented programming. By doing this, we aim to develop on the vision of the CBSE as a sound base to produce software systems enabled to automatically and safely reconfigure themselves by reconfiguring their abstract (reflection) architectures at runtime. For such structural reconfigurations, a system architect may reuse design patterns from other disciplines with the purpose of restoring QoS contracts, thus preserving them.

Our approach, named SCeSAME for “A Safe Contract-based Self-Adaptive Framework to Preserve QoS Properties on Mobile Devices” is built on the theory of extended graph (e-graph) rewriting proposed in e-graph [60], as a formalism to represent QoS contracts on component. We have given a formal definition of component-based structure systems, QoS contracts, and architecture reconfiguration rules. Based on these definitions, we built a framework that enables a component-based system to preserve its QoS contracts through architecture self-reconfiguration as a responding action to QoS contract violations. Our approach extends a theory of graph rewriting and defines a process calculus as formalisms to model the structure and reconfiguration process of architecture reconfiguration. The reconfiguration process, once parameterized with reconfiguration rules, can be verified as safe, i.e., component structural-compliant, terminating and confluent. This result is a part of Gabriel Tamura’s PhD and the results have been published in [36], [44].
6. New Results

6.1. SouthBound results

6.1.1. L4 micro kernels

As part of our investigations about what software architectures were the best candidates to base our Ambient Middleware Stack upon, we studied different micro-kernel operating systems such as CodeZero [33], OKL4, and L4/Fiasco. The objective here is to try and quantify the development effort that would be needed before being able to execute a Java application on top of a micro-kernel. These studies included, in addition to a lot of bibliographic research, several technical experiments such as booting each of these various micro-kernel systems in QEmu, as well as on real hardware. We use a BeagleBoard as a representative example of the kind of hardware platforms typically encountered in Ambient Intelligence scenarios.

6.1.2. Virtual machines

In parallel to our study of micro-kernel architectures, we worked on virtual machines as well, in the perspective of bridging the gap between the two. The basic question here is: what does it takes to to cut down a Java virtual machine into pieces so as to run each of these pieces as a separate software component in the system. We ran two actions in order to investigate this question. First, we ported the JamVM virtual machine to run on top of the Genode operating system framework [37]. This provided us with better understanding of what are the real requirements of a Java virtual machine in terms of underlying operating systems support. Second, we focused on one particular service of the virtual machine, the garbage collector, and we precisely identified and studied the coupling between this component and all other parts of the virtual machine (bytecode interpreter, scheduler, etc). This work was done as part of a student summer internship (Yann Chevalier, INSA-Lyon 3IF). Removing a garbage collector at runtime, and “plugging in” another one dynamically proved to be vastly harder than expected. Still, this work provided us with great insights about the coupling relationships between different OS components.

6.1.3. HiKoB

Antoine Fraboulet (Amazone team), Guillaume Chelius (D-NET team) and Christophe Braillon (INRIA SED) started a new company called HiKob http://openlab.hikob.com/ in July 2011. HiKoB is a development project following several successful research projects completed these last 6 years at INSA Lyon and INRIA. HiKoB hardware and software products help in building complex, large-scale and distributed applications in the domains of: motion capture, biomechanical study, biologging study, building instrumentation and many more applications targeting wireless sensor network solutions for distributed and embedded measurement. HiKoB business model is built on two major directions: complete solutions for industrial applications and software and hardware tools for research and innovation in the fields of sensor networking and embedded wireless measure. HiKoB is supported by IT-Translation and INSAValor.

6.1.4. Service-Oriented Tainted Object Propagation

Many Java technologies allow the execution of code provided by multiple parties. Service-oriented platforms based on components such as OSGi are good examples of such a scenario. Those extensible component-based platforms are service-oriented, as components may directly interact with each other via the services they provide. However, even robust languages such as Java were not designed to handle safely code interaction between trusted and untrusted parties.

In [38], we show how basic Java interactions can break encapsulation or execution safety and why the Java security layers’ threat coverage is incomplete. We also review flaws in the Java access control design that can allow untrusted code to bypass restrictions by exploiting vulnerabilities in trusted code.
As component-based platforms become more and more integrated to our daily life, we improved our Service-Oriented Tainted Object Propagation technique to find such vulnerabilities and used it on several open-source components to further demonstrate the real exposure that those vulnerabilities bring to the fore.

6.2. NorthBound results

Another key issue in the Amazones architecture was to bring together formal methods and service oriented programming such as OSGi/Java approach. We developed the Logos framework that observes and records communications that occurs between an OSGi client and a corresponding server. This architecture is developed in the LISE ANR project, and we made various improvements to the architecture.

6.2.1. Amazones Protocol

It aims at building automata from running and observing applications. The logos framework observes a running application at builds at run-time an automata that represents the application behavior. Julien Ponge wrote the corresponding Scala code.

6.2.2. Monitored oriented programming

Another Logos extension integrates Monitored Oriented architectures such as JavaMOP and Larva. We are currently using and working with the larva people. Each time we intercept a call, it is transferred to the Larva automata manager.

6.2.3. Real time SOA

Admission control for service oriented application in real time infrastructure. This work led by Lionel Morel tries to bring together real time architecture configuration and component based architectures. The deal is to find a better way of managing the dynamicity of applications in real time context.

6.2.4. B Model Slicing to Generate Tests

In a model-based testing approach as well as for the verification of properties, B models provide an interesting modelling solution. However, for industrial applications, the size of their state space often makes them hard to handle. To reduce the amount of states, an abstraction function can be used. The abstraction is often a domain abstraction of the state variables that requires many proof obligations to be discharged, which can be very time consuming for real applications.

we propose a contribution to this problem that complements an approach based on domain abstraction for test generation, by adding a preliminary syntactic abstraction phase, based on variable elimination. We define a syntactic transformation that suppresses some variables from a B event model, in addition to three methods that choose relevant variables according to a test purpose. This way, we propose a method that computes an abstraction of a source model $M$ according to a set of selected relevant variables. Depending on the method used, the abstraction can be computed as a simulation or as a bi-simulation of $M$. With this approach, the abstraction process produces a finite state system. We apply this abstraction computation to a Model Based Testing process. We evaluate experimentally the impact of the model simplification by variables elimination on the size of the models, on the number of proof obligations to discharge, on the precision of the abstraction and on the coverage achieved by the test generation.

This work is based on a B model approach. However, in the context of AMAZONES, one of our objectives is to extend it, in order to consider models automatically generated from the usage a the tested service on a particular context.

6.2.5. Distributed Data Centric Programs Verification

Netlog is a language designed to describe distributed programs. It has a precise semantics, provides a high-level of abstraction thanks to its datalog flavor and benefits from an efficient implementation. This makes it a very interesting target language for proofs of distributed programs. In [34], with the Coq proof assistant, we formalized the distributed computation model based on message passing with either synchronous or
asynchronous behaviors; built the translation of Netlog programs; modeled the embedded machine evaluating Netlog programs, and thus established a framework to formally verify properties of distributed programs in Netlog. To test the framework, we proved the correctness of a concrete distributed program for constructing spanning trees over connected graphs.

6.2.6. Managing dynamic service substitution at runtime

The service oriented approach is a paradigm allowing the introduction of dynamicity in developments. If there are many advantages with this approach, there are also some new problems associated to service disappearance. The particular case of service substitution is often studied and many propositions exist. However, proposed solutions are mainly server-side in the context of web-services.

In this work, we propose a client side API-based approach to allow service substitution without any restart of the client and without any assumption on external services. Our proposition is based on a transactional approach, defined to authorized substitutions of services dynamically, by preserving the current run and collected data.

We designed a framework organized by Julien Ponge [14].

6.3. Application domain results

An emerging trend into Amazones team is to apply our northBound/southBound approach to the InternetOfThing wave. We try to apply our architectures to the IoT application domain.

6.3.1. Data Centric Applications Distribution

Peer to peer systems have been widely used to alleviate the burden of servers by transferring to peers in a network tasks that do not require a centralization of the information. A wide range of applications are now emerging over peer-to-peer, such as social networking, multiplayer games, mobile messaging, etc. Most of these applications are essentially data-centric, they rely on exchange of data between peers, and can be expressed by queries over the database.

We propose a tool that allows for such applications, programmed as a collection of queries over a database, to be ported seamlessly without changing the initial queries from a client/server system to peer-to-peer system. The distribution is done with overlays network defined by declarative data centric programs specified in the Netlog language, thus resulting in a fully data centric modeling of the peer-to-peer application. The communication between peers relies on implicit addresses which can be evaluated on the fly to ensure the persistence of data.

We demonstrate the technique on a multiplayer online game, written in SQL, with players who connect to a mobile ad hoc network through their portable devices. The overlay is defined by a combination of an ad hoc routing protocol, DSDV, together with a DHT. The application runs on the QuestMonitor system, which allows to monitor the communication between peers, the evolution of the local data stores, as well as the execution of the declarative code.

6.3.2. Service Deployment in Disrupted Networks

OLD / REMOVE? Ambient environments classically use wireless connections that suffers from frequent disconnections. The hard research point is to ensure service continuity. This disconnection problem has been widely tackled for application data with proxy and prefetching approaches. For services, disconnections are more difficult to anticipate, since service calls are only solved at run-time.

We are currently working on service deployment and invocations in disrupted networks with a network coding approach. This research is a joint work with the Swing team and with Aline Viana (INRIA Saclay @ TU Berlin). The main idea is to study how social-oriented applications, that need inter-dependent services and updates to be distributed to all or part of the mobile users community, could benefit from a network coding approach. The project aims at assessing for the first time the performance, in terms of latency, energy efficiency and capacity, of standard network coding techniques in presence of realistic user mobility and service demands.
Building on these results, we plan to propose original social-aware network coding techniques that take advantage of the heterogeneous nature of the opportunistic network to reduce delays and energy consumption, in presence of multiple concurrent service flows targeting either all users or specific groups of interests.

These performance issues tackle at the same time the overall network capacity optimizations, as well as the overall software stack optimizations of a device with local and autonomous network coding strategies.

An INRIA ARC project proposal, entitled SoCool, has been submitted jointly with INRIA AMAZONES, INRIA SWING, INRIA MAESTRO, INRIA Saclay, University of Nice, TU Berlin and Fordham University.
6. New Results

6.1. Introduction

The ARLES project-team investigates solutions in the forms of languages, methods, tools and supporting middleware to assist the development of distributed software systems, with a special emphasis on mobile distributed systems enabling the ambient intelligence/pervasive computing vision. Our research activities in 2011 have focused on the following areas:

- Dynamic interoperability among networked systems toward making them eternal, by way of on-the-fly generation of connectors based on adequate system models (§ 6.2);
- Pervasive service-oriented software engineering, focusing on supporting service composition in an increasingly heterogeneous and dynamic networking environment, while enforcing quality of service (§ 6.3);
- Service oriented middleware for the ultra large scale future Internet of Things (§ 6.4);
- Abstractions for enabling domain experts to easily compose applications on the Internet of Things (§ 6.5); and
- System-level support for application development in the context of mobile social ecosystems, while taking into account privacy, performance, and data interoperability (§ 6.6).

6.2. Emergent Middleware Supporting Interoperability in Extreme Distributed Systems

Participants: Emil Andriescu, Nelly Bencomo, Amel Bennaceur, Luca Cavallaro, Nikolaos Georgantas, Sneha-Sham Godbole, Valérie Issarny, Rachid Saadi, Daniel Sykes.

Interoperability is a fundamental challenge for today’s extreme distributed systems. Indeed, the high-level of heterogeneity in both the application layer and the underlying infrastructure, together with the conflicting assumptions that each system makes about its execution environment hinder the successful interoperation of independently developed systems. A wide range of approaches have been proposed to address the interoperability challenge [31]. Solutions that require performing changes to the systems are usually not feasible since the systems to be integrated may be legacy systems, COTS (Commercial Off-The-Shelf) components or built by third parties; neither are the approaches that prune the behavior leading to mismatches since they also restrict the systems’ functionality. Therefore, many solutions that aggregate the disparate systems in a non-intrusive way have been proposed. These solutions use intermediary software entities, called mediators, to interconnect systems despite disparities in their data and/or interaction models by performing the necessary coordination and translations while keeping them loosely-coupled. However, creating mediators requires a substantial development effort and a thorough knowledge of the application-domain, which is best understood by domain experts. Moreover, the increasing complexity of today’s distributed systems, sometimes referred to as Systems of Systems, makes it almost impossible to develop ‘correct’ mediators manually. Therefore, formal approaches are used to synthesize mediators automatically.
In light of the above, we have introduced the notion of *emergent middleware* for realizing mediators. Our research on enabling emergent mediators is done in collaboration with our partners of the CONNECT project (§ 7.1.1). Our work during the year has more specifically focused on:

- **Supporting architecture.** We have been working together with our partners in the CONNECT project on the refinement of an overall architecture supporting emergent middleware, from the discovery of networked systems to the learning of their respective behavior, and synthesis of emergent middleware enabling them to interoperate [30].

- **Affordance inference.** We have proposed an ontology-based formal model of networked systems based on their affordances, interfaces, behavior, and non-functional properties, each of which describes a different facet of the system [2]. However, legacy systems do not necessarily specify all of the aforementioned facets. Therefore, we are currently exploring techniques to infer the affordance by using textual descriptions of the interface of networked systems. More specifically, we rely on machine learning techniques to automate the inference of the affordance from the interface description by classifying the natural-language text according to a predefined ontology of affordances [17].

- **Mediator synthesis for emergent connectors.** We focus on systems that have compatible functionality, i.e., semantically matching affordances, but are unable to interact successfully due to mismatching interfaces or behaviors. We propose two approaches to enable communication between such systems:
  
  1. A *mapping based* approach, whose goal is to automatically synthesize a mediator model that ensures their safe interaction, i.e., deadlock-freedom and the absence of unspecified receptions. Our approach combines semantic reasoning and constraint programming to identify the semantic correspondence between networked systems’ interfaces, i.e., interface mapping. Unlike existing approaches that only tackle the one-to-one correspondence between actions, this approach handles the more general cases of one-to-many and many-to-many mappings.
  
  2. A *goal based* approach, which enables the communication of two networked systems, so that the communication satisfies a given user goal. It aligns their actions using ontology matching. The aligned processes as well as the user goal are encoded as a satisfiability problem. It relies on model checking to determine if a feasible communication trace exists that satisfies the user goal. The model checking process is reiterated so as to discover all the feasible satisfying traces, which are finally concatenated to build the mediator.

The feasibility of both of our approaches has been demonstrated through prototype tools and real-world scenarios involving heterogeneous systems.

- **Mediator synthesis for streaming connectors.** In the context of dynamic mediator synthesis, we have targeted the domain of mobile multimedia streaming, resulting in a first step that statically solves the hard problem of streaming interoperability across heterogeneous smartphone multimedia platforms. With the recent evolution of mobile phones, multimedia streaming is now commonly used in smartphones for purposes such as video broadcast, video conferencing and place shifting, which in turn highlights the importance of multimedia enabled applications. However, peer-to-peer solutions are difficult to implement because of increased node heterogeneity and their low processing power. Furthermore, existing mobile platforms such as Android, iOS, Blackberry and Windows Phone 7 support multimedia streaming (as resource consumers) either through platform specific APIs or system services. However, they use heterogeneous protocols and data formats, thus compromising interoperability.

Given the challenges above, we designed AmbiStream [11], a lightweight middleware for heterogeneous mobile devices, capable of “on the fly” adaptation. AmbiStream relies on the highly-optimized multimedia software stacks provided by smartphone platforms and adds the necessary
layers to solve interoperability. More specifically, the middleware targets: a) Streaming of prerecorded or live audio/video using an intermediary real-time protocol; b) Managing streaming protocol translation and multimedia container format adaptation to the ones supported natively by each device; and c) Extensibility in order to support new multimedia streaming protocols and multimedia container formats given its plug-in based architecture. We have used a model-driven approach to generate multi-platform plug-ins from higher level descriptions in the form of a Domain Specific Language (DSL). The defined DSL takes into account multimedia specific operations such as timing, fragmenting, multiplexing, congestion control and buffering.

- **Models@run.time.** We have recently integrated the notion of Models@run.time \(^{11}\) in our research towards emergent middleware. We use Models@run.time to extend the applicability of models and abstractions to the runtime environment. As is the case for software development models, a run-time model is often created to support reasoning. However, in contrast to development models, run-time models are used to reason about the operating environment and runtime behavior, and thus these models must capture abstractions of runtime phenomena. Different dimensions need to be balanced, including resource-efficiency (time, memory, energy), context-dependency (time, location, platform), as well as personalization (quality-of-service specifications, profiles). The hypothesis is that because Models@run.time provide meta-information for these dimensions during execution, run-time decisions can be facilitated and better automated. Thus, we anticipate that Models@run.time will play an integral role in the management of extremely distributed systems. Our work on the use of Models@run.time has two aspects:

  - We have used Models@run.time to tackle the crucial problem of uncertainty in extremely distributed systems that are aware of their own requirements. Requirements awareness helps optimize requirements satisfaction when factors that were uncertain at design time are resolved at runtime. Using our approach, we are able to maintain goal-based models in memory while the system is running. The executing system, therefore, is able to introspect and consult it goals during runtime. Crucially, at runtime we use the notion of claims to represent assumptions that cannot be verified with confidence at design time. Such claims are attached to the goal-based runtime models. By monitoring claims at runtime, their veracity can be tested. If falsified, the effect of claim negation can be propagated to the system’s goal model and an alternative means of goal realization can be selected automatically, allowing the dynamic adaptation of the system to the prevailing environmental context \(^{[14]}\), \(^{[15]}\), \(^{[16]}\).

  - In a complementary way to the mediator synthesis approaches discussed above, we further promote the use of Models@run.time to support the runtime synthesis of software that will be part of the executing system. Specifically, we focus on the use of runtime models to support the realization of emergent middleware, i.e., the synthesis of mediators that define a sequences of actions to translate semantic actions of one system developed using a particular middleware protocol to the semantic actions of another system developed using an alternate middleware built with no prior knowledge on the former. Discovery and learning enablers capture the required knowledge of the context and environment during runtime. Supported by that knowledge, a runtime model of the mediator-to-be is reified. Reification means that the knowledge is explicitly formulated and made available for computational manipulation. The form of the runtime models is based on labeled transition systems (LTSs) which offer the behavioral semantics needed to model the interaction protocols. Ontologies complement the LTSs providing semantic reasoning about the mapping between protocols. Specifically the LTS of each protocol is annotated using ontologies to support the subsequent mapping between the protocols. From the LTS-based runtime models, mediators are synthesized.

\(^{11}\)Models@run.time Dagstuhl Seminar, [http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=11481](http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=11481)
6.3. Revisiting the Abstractions of Service Oriented Computing for the Future Internet

Participants: Mohammad Ashiqur Rahaman, Dionysis Athanasopoulos, Sandrine Beauche, Nebil Ben Mabrouk, Nikolaos Georgantas, Valérie Issarny.

A software architecture style characterizes, via a set of abstractions, the types of: components (i.e., units of computation or data stores), connectors (i.e., interaction protocols) and possibly configurations (i.e., system structures) that serve to build a given class of systems. As such, the definition of a software architectural style is central toward eliciting appropriate design, development and runtime support for any family of systems. The service oriented architecture style may then be briefly defined as follows: (1) components map to services, which may be refined into consumer, producer or prosumer services; (2) connectors map to traditional client-service interaction protocols; and (3) configurations map to compositions of services through (service-oriented) connectors, e.g., choreography and orchestration structures. While the service-oriented architecture style is well suited to support the development of Internet-based distributed systems, it is largely challenged by the Future Internet that poses new demands in terms of sustaining qualities such as scalability, heterogeneity, mobility, awareness & adaptability that come in extreme degrees compared to the current Internet. Therefore, we have been working on eliciting software architectural abstractions for the Future Internet by building upon the service-oriented architecture style, as well as on applying them to system design, development and execution.

Complex distributed applications in the Future Internet will be to a large extent based on the open integration of extremely heterogeneous systems, such as lightweight embedded systems (e.g., sensors, actuators and networks of them), mobile systems (e.g., smartphone applications), and resource-rich IT systems (e.g., systems hosted on enterprise servers and Cloud infrastructures). These heterogeneous system domains differ significantly in terms of interaction paradigms, communication protocols, and data representation models, provided by supporting middleware platforms. Specifically considering interaction paradigms, the client/server (CS), publish/subscribe (PS), and tuple space (TS) paradigms are among the most widely employed ones today, with numerous related middleware platforms. In light of the above, we have aimed at eliciting abstractions that (i) leverage the diversity of interaction paradigms associated with today’s and future complex distributed systems, as well as (ii) enable cross-paradigm interaction to sustain interoperability in the highly heterogeneous Future Internet [19].

Existing cross-domain interoperability efforts are based on bridging communication protocols, wrapping systems behind standard technology interfaces, and/or providing common API abstractions. In particular, such techniques have been applied by the two widely established system integration paradigms, that is, service oriented architecture (SOA) and enterprise service bus (ESB). However, state of the art interoperability efforts do not or only poorly address interaction paradigm interoperability. Indeed, systems integrated via SOA and ESB solutions have their interaction semantics transformed to the CS paradigm. Then, potential loss of interaction semantics can result in suboptimal or even problematic system integration. To overcome the limitation of today’s ESB-based connectors for cross-domain interoperability in the Future Internet, we introduce a new connector type, called GA connector, which stands for “Generic Application connector”. The proposed connector type is based on the service bus paradigm in that it achieves bridging across heterogeneous connector types. However, the behavior of the GA connector type differs from that of classical ESB connectors by bridging protocols across heterogeneous paradigms, which is further realized by paying special attention to the preservation of the semantics of the composed protocols. Indeed, the GA connector type is based on the abstraction and semantic-preserving merging of the common high-level semantics of base interaction paradigms.

Eliciting Interaction Paradigm Abstractions: We introduce a systematic abstraction of interaction paradigms with the following features:

- First, we introduce base CS, PS and TS connector types, which formally characterize today’s core interaction paradigms. The proposed types comprehensively cover the essential semantics of the considered paradigms, based on a thorough survey of the related literature and representative
Then, we further abstract these connector types into a single higher-level one, the GA connector type. GA is a comprehensive connector type based on the abstract union of CS, PS, and TS, where precise identification of the commonalities or similarities between the latter has enabled the optimization of the former. Further, GA preserves by construction the semantics of CS, PS, and TS.

In more detail, connector types are formally specified in terms of: (i) their API (Application Programming Interface), and (ii) their roles, i.e., the semantics of interaction of the connected component(s) with the environment via the connector. Regarding the latter, the behavioral specification of roles from a middleware perspective relates to specifying the production and consumption of information in the network, while the semantics of the information are abstracted and dealt with at the application layer. The behaviors of the connector roles are then specified using Labeled Transition Systems (LTS). We precisely define the mapping of the roles implemented by the base connector types to/from the corresponding roles of the GA connector type.

For both the above abstraction transformations, we provide counterpart concretizations, which enable transforming GA connector primitives to CS, PS, or TS connector primitives and then to concrete middleware platforms primitives.

Furthermore, based on the GA abstraction, we introduce mapping transformations between any pair from the set \{CS, PS, TS\} via GA. The fine knowledge of CS, PS, and TS semantics, as embedded in GA, enables these mappings to be precise: differing semantics are mapped to each other in such a way that loss of semantics is limited to the minimum. These mappings relate to the definition of the glue process implemented by the GA connector, which defines how a pair of producer and consumer roles coordinates in the environment. The GA glue reconciles consumer and producer roles that may differ with respect to time and space coupling as well as scoping. Hence, GA connectors support interactions among highly heterogeneous services of the Future Internet, and especially across domains.

eXtensible Service Bus: We apply the above connector abstractions to introduce an enhanced bus paradigm, the eXtensible Service Bus (XSB). XSB features richer interaction semantics than common ESB implementations to deal effectively with the increased Future Internet heterogeneity. Moreover, from its very conception, XSB incorporates special consideration for the cross-integration of heterogeneous interaction paradigms. When mapping between such paradigms, special attention is paid to the preservation of interaction semantics. XSB has the following features:

- XSB is an abstract bus that prescribes only the high-level semantics of the common bus protocol. The XSB common bus protocol features GA semantics.
- Heterogeneous systems can be plugged into the XSB by employing binding components that adapt between the native middleware of the deployed system and the common bus protocol. This adaptation is based on the systematic abstractions and mappings discussed above
- XSB, being an abstract bus, can have different implementations. This means that it needs to be complemented with a substrate which at least supports: (1) deployment (i.e., plugging) of various systems on the bus, and (2) a common bus protocol implementing GA semantics. With respect to the latter, we envision that a GA protocol realization may either be designed and built from scratch (still supposing at least an IP-based transport substrate) or be implemented by conveying GA semantics on top of an existing higher-level protocol used as transport carrier. The latter solution can be attractive, as it facilitates GA protocol realizations in different contexts and domains.

We have carried out an early realization of XSB on the PEtALS ESB. In particular, we addressed the workflow-based orchestration of heterogeneous systems, which is a preliminary step before dealing with peer-wise system integration. This work already provides a successful feasibility study of the XSB concept. This work comprises: (i) extending the BPEL workflow language with GA API primitives; (ii) introducing transformation between the GA-extended BPEL and the standard BPEL, which consists in encapsulating GA primitives into standard BPEL primitives and enables conveying GA semantics on top of BPEL primitives.
6.4. Service Oriented Middleware facing the Challenges of the Internet of Things

Participants: Benjamin Billet, Nikolaos Georgantas, Sara Hachem, Valérie Issarny, Roberto Speicys Cardoso, Thiago Valladares Sabino Teixeira.

Over the years, the Internet has become the most important networking infrastructure, providing an integrated entity enabling sharing, contributing, creating, using, collaborating and integrating information and knowledge by all. As a result, the Internet is changing at fast pace and is expected to evolve into the Future Internet, i.e., service-aware and self-aware federated networks that provide built-in and integrated capabilities such as: contextualization, reliability, robustness, mobility, security, service support, and self-management of communication resources and services. In our vision, The Future Internet can be defined as the union and cooperation of the Internet of Content, Internet of Services, Internet of Things, and 3D interactive Internet, supported by an expanding network infrastructure foundation. In ARLES, we chose to pay special attention to the Internet of Things (IoT). IoT is characterized by the integration of large numbers of real-world objects (or “things”) onto the Internet, with the aim of turning high-level interactions with the physical world into a matter as simple as is interacting with the virtual world today. As such, two devices that will play a key role in the IoT are sensors and actuators. In fact, such devices are already seeing widespread adoption in the highly localized systems within our cars, mobile phones, laptops, home appliances, etc. In their current incarnation, however, sensors and actuators are used for little more than low-level inferences and basic services. This is partly due to their highly specialized domains (signal processing, estimation theory, robotics, etc.), which demand application programmers to also be domain experts, and partly due to a glaring lack of interconnectivity between all the different devices. Our work within that domain was focused on three related directions:

- **Challenges related to IoT:** To prepare the ground for our research on middleware for the Internet of Things, we identified the set of challenges in the IoT, namely [10]: the large scale of the Internet of Things, heterogeneity of things, unknown and dynamic network topology, unknown data-point availability, incomplete or inaccurate metadata, and conflict resolution. The scale issue arises with the millions of devices, millions of users, large amounts of data to share and services to request. The heterogeneity of the IoT is due to the fact that the network will be composed of different types of devices from different vendors with varying sensing/actuating characteristics. The unknown dynamic network topology results from the fact that devices will be mostly mobile and their availability is unknown. A related challenge is the unknown data-point availability as things, which provide the desired measurements, may leave the network or malfunction at any time. A data point is measurement of an entity of interest at a specific time. As for metadata inaccuracy, this issue is a direct result of humans, who are prone to making errors, being the main source of metadata specification. Last but not least, conflict resolution is due to the multiple stakeholders involved in the Internet of Things.

- **Middleware Requirements for the Internet of Things:** The middleware we plan on implementing should abstract things (IoT devices) as services and support dynamic service composition. To handle the IoT challenges, the middleware should also support a probabilistic discovery approach where only a subset, instead of a whole set, of devices is selected in a way that provides a good enough answer that satisfies an application’s request [10]. However, and prior to designing the middleware architecture, we extensively surveyed the literature in order to identify research challenges for service-oriented middleware design, therefore investigating service description, discovery, access and composition in the Future Internet of services [7].

- **Ontologies for the Internet of Things:** As part of our middleware architecture, we specified a set of ontologies [20] that model real-world entities as physical concepts, along with things that measure...
those entities. Further, to support a smarter service composition, we also modeled mathematical formulas and physics relations as services to substitute missing thing-based services. Those services will instead compute the value of a desired measurement of an entity of interest. Finally, we also specified an ontology that describes estimation models that can be used to estimate the value of a measurement in case of a missing data point or a missing data source. Estimation models can further be used to define probabilistic discovery functions that will be executed by the middleware.

6.5. Composing Applications in the Internet of Things

Participants: Iraklis Leontiadis, Pankesh Patel, Animesh Pathak.

As introduced above, the Internet of Things (IoT) integrates the physical world with the existing Internet, and is rapidly gaining popularity, thanks to the increased adoption of smart phones and sensing devices. Several IoT applications have been reported in recent research, and we expect to see increased adoption of IoT concepts in the fields of personal health, inventory management, and domestic energy usage monitoring, among others.

An important challenge to be addressed in the domain of IoT is to enable domain experts (health-care professionals, architects, city planners, etc.) to develop applications in their fields rapidly, with minimal support from skilled computer science professionals. Similar challenges have already been addressed in the closely related fields of Wireless Sensor and Actuator Networks (WSANs) and Pervasive/Ubiquitous computing. While the main challenge in the former is the extremely large scale of the systems (hundreds to thousands of largely similar nodes, sensing and acting on the environment), the primary concern in the latter has been the heterogeneity of nodes and the major role that the user’s own interaction with these nodes plays in these systems (cf. the classic “smart home” scenario where the user interacts with a smart display which works together with his refrigerator and toaster). The upcoming field of IoT includes both WSANs as well as smart appliances, in addition to the elements of the “traditional” Internet such as Web and database servers, exposing their functionalities as Web services etc. Consequently, an ideal application development abstraction of the IoT will allow (domain expert) developers to intuitively specify the rich interactions between the extremely large number of disparate devices in the future Internet of Things.

The larger goal of our research is to propose a suitable application development framework which addresses the challenges introduced above. This will most likely be achieved by a domain specific language (DSL) that exposes specific functionalities to the domain experts. The first logical step was to construct a domain model. Towards that end, we took advantage of the CRC — Classes, Responsibility, Collaboration — technique, defining the main abstract concepts, their responsibilities, and associations that represent their relationship with each other in the IoT. Specifically, we used this technique to propose a domain model [22] that addresses the following challenges:

- **Creation of common understanding.** The different terms used by different people in the IoT domain can lead to confusion, which can be alleviated by the usage of a common lexicon, as provided by a domain model. This lexicon can then be used by researchers, system programmers, as well as domain experts.

- **Modeling invariant properties.** The domain model represents the invariant properties of the domain — concepts and relationships which do not change from one application to the other. An instance of this in the IoT domain can be the notion of a sensor attached to a device. Depending on the specific applications, the type of sensors and devices can change (e.g., a light sensor attached to a smart phone), but the inherent relationship between the types of entities they represent does not.

- **Enabling modular design.** Application needs often tend to arrive in terms of behavior, which needs to be broken down and divided among the entities in the system. A good domain model aids in this process, since the capabilities of each type of entity are clearly identified. E.g., the application requirement of “the system senses the temperature of a room and keeps it steady” can be easily broken down into an application consisting of temperature sensors, computational components, and HVAC actuators, each performing its well-known role in this sense-compute-actuate loop.
As part of a related effort with a narrower focus on the domain of sensor network *macroprogramming* — a technique that aims to aid the wide adoption of networked sensing by providing the domain expert the ability to specify their applications at a high level of abstraction — we have explored techniques to bring *Web services* in the gamut of sensor network macroprogramming. Our research addresses the challenges faced by developers of systems where sensors (e.g., RFID badge sensors in an office) interact with pre-existing larger software components exposed as Web services (e.g., the office personnel access control database). As part of our work, we have proposed extensions to the data-driven ATaG macroprogramming language using which developers can easily incorporate existing Web services in their applications.

We have incorporated our continued research 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.

### 6.6. Addressing Middleware Challenges in Large Scale Mobile Social Networks of the Future

**Participants:** Sara Hachem, Valérie Issarny, Animesh Pathak, Amir Seyedi.

With the increased prevalence of advanced mobile devices (the so-called “smart” phones), interest has grown in *Mobile Social Ecosystems* (MSE), where users not only access traditional on-line Web-based social networks using their mobile devices, but are also able to use the context information provided by these devices to further enrich their interactions. In complex mobile social ecosystems of the future, the heterogeneity of software platforms on constituent nodes, combined with their intrinsic distributed nature and heterogeneity in representation of data and context, as well as user’s privacy and trust concerns, raises the need for middleware support for the development of mobile social applications. We believe that the development of mobile social applications can be greatly simplified by the presence of middleware support. To that end, we have been working on addressing the following challenges:

- **Semantic models for mobile social ecosystems.** In order to enable re-use of data between different social applications run by the same user, we have proposed an expressive and extensible model using semantic techniques to represent MSE and the interactions possible in them. This supports semantic interoperability between separately developed applications and minimizes resource-consuming operations such as data mapping and replication.

- **Efficient decentralized storage of social data.** Instead of storing the social knowledge of the whole world with a single provider — a practice performed today by common social networks such as Facebook — which can lead to privacy issues, our research endeavors to propose a middleware using which users can store their personal knowledge in a distributed manner on the devices owned by them (e.g., smart phone, home desktop, laptop). This also allows users to provide selective access to other users based on semantically defined access control policies.

- **Socially aware policies for access control.** Since social data is private and sensitive in nature, we have proposed a policy framework [21] where the user can specify both the data to be protected as well as the relevant set of peers with access to that data in a socially-aware manner (e.g., “only let my colleagues know my location during weekdays from 9 – 5”). This policy framework can be used as a guard around the user’s knowledge base, allowing access only to authorized peers. We are also working on providing end-users an easy to use editor so as to be able to specify these socially-aware policies easily.

- **Social data extraction from existing sources.** Our research includes work in enabling users to populate their social knowledge base by extracting data from their existing repositories. We have identified two types of sources of such data. The first already contain social links such as “friendship” in addition to general information, while the second do not contain social links, but may contain information which can be correlated to infer social links (e.g., call and SMS logs). We are working on a framework where adapters can be written for the former using their API to import their data; while for the latter, inference algorithms can be used to correlate data and guess/recommend social links.
• **Inferring trust from proximity.** In mobile social network, highly sensitive private data is at risk of being shared with unwanted peers, since users may not have any knowledge about the users they socially connect with. Trust management then appears as a promising decision support for mobile users in establishing social links. However, while the literature is rich in trust models, most approaches lack appropriate trust bootstrapping, i.e., the initialization of trust values. In [24], we address this challenge by introducing proximity-based trust initialization based on the users’ behavioral data available from their mobile devices or other types of social interactions. The proposed approach is further assessed in the context of mobile social networking using users behavioral data collected by the MIT reality mining project. Results show that the inferred trust values correlate with the self-reported survey of users relationships.

We have incorporated our research in the above areas into Yarta [25], a middleware for mobile social applications. Our prototype middleware, as discussed in §5.6, currently supports application development for laptops as well as Android-powered smart phones, providing distributed storage of semantically-modeled social knowledge guarded by a rich policy framework.
6. New Results

6.1. Models and abstractions for distributed systems

This section summarizes the major results obtained by the ASAP team that relate to the foundations of distributed systems.

6.1.1. The weakest failure detector to implement a register in asynchronous systems with hybrid communication

Participants: Damien Imbs, Michel Raynal.

This work introduces an asynchronous crash-prone hybrid system model. The system is hybrid in the way the processes can communicate. On the one side, a process can send messages to any other process. On another side, the processes are partitioned into clusters and each cluster has its own read/write shared memory. In addition to the model, a main contribution of the work concerns the implementation of an atomic register in this system model. More precisely, a new failure detector (denoted $M \Sigma$) is introduced and it is shown that, when considering the information on failures needed to implement a register, this failure detector is the weakest. To that end, the work presents an $M \Sigma$-based algorithm that builds a register in the considered hybrid system model and shows that it is possible to extract $M \Sigma$ from any failure detector-based algorithm that implements a register in this model. The work also (a) shows that $M \Sigma$ is strictly weaker than $\Sigma$ (which is the weakest failure detector to implement a register in a classical message-passing system) and (b) presents a necessary and sufficient condition to implement $M \Sigma$ in a hybrid communication system.

This work has been published in SSS 2011 [38].

6.1.2. The universe of symmetry breaking tasks

Participants: Damien Imbs, Michel Raynal.

Processes in a concurrent system need to coordinate using a shared memory or a message-passing subsystem in order to solve agreement tasks such as, for example, consensus or set agreement. However, coordination is often needed to “break the symmetry” of processes that are initially in the same state, for example, to get exclusive access to a shared resource, to get distinct names or to elect a leader.

This work introduces and studies the family of generalized symmetry breaking (GSB) tasks, that includes election, renaming and many other symmetry breaking tasks. Differently from agreement tasks, a GSB task is “inputless”, in the sense that processes do not propose values; the task only specifies the symmetry breaking requirement, independently of the system’s initial state (where processes differ only on their identifiers).

Among various results characterizing the family of GSB tasks, it is shown that (non adaptive) perfect renaming is universal for all GSB tasks.

This work was done in collaboration with Sergio Rajsbaum from the Universidad Nacional Autonoma de Mexico and was published in SIROCCO 2011 [36].

6.1.3. Read invisibility, virtual world consistency and probabilistic permissiveness are compatible

Participants: Tyler Crain, Damien Imbs, Michel Raynal.
The aim of a Software Transactional Memory (STM) is to discharge the programmers from the management of synchronization in multiprocess programs that access concurrent objects. To that end, an STM system provides the programmer with the concept of a transaction. The job of the programmer is to design each process the application is made up of as a sequence of transactions. A transaction is a piece of code that accesses concurrent objects, but contains no explicit synchronization statement. It is the job of the underlying STM system to provide the illusion that each transaction appears as being executed atomically. Of course, for efficiency, an STM system has to allow transactions to execute concurrently. Consequently, due to the underlying STM concurrency management, a transaction commits or aborts.

This work studies the relation between two STM properties (read invisibility and permissiveness) and two consistency conditions for STM systems, namely, opacity and virtual world consistency. Both conditions ensure that any transaction (be it a committed or an aborted transaction) reads values from a consistent global state, a noteworthy property if one wants to prevent abnormal behavior from concurrent transactions that behave correctly when executed alone. A read operation issued by a transaction is invisible if it does not entail shared memory modifications. This is an important property that favors efficiency and privacy. An STM system is permissive (respectively probabilistically permissive) with respect to a consistency condition if it accepts (respectively accepts with positive probability) every history that satisfies the condition. This is a crucial property as a permissive STM system that implements opacity while ensuring read invisibility. It then shows that read invisibility, probabilistic permissiveness and virtual world consistency are compatible. To that end the work describes a new STM protocol called IR_VWC_P. This protocol presents additional noteworthy features: it uses only base read/write objects and locks which are used only at commit time; and, in favorable circumstances, the cost of a read operation is $O(1)$.

This work has been published in ICA3PP 2011 [29].

6.1.4. Towards a universal construction for transaction-based multiprocess programs

Participants: Tyler Crain, Damien Imbs, Michel Raynal.

The aim of a Software Transactional Memory (STM) system is to discharge the programmer from the explicit management of synchronization issues. The programmer’s job resides in the design of multiprocess programs in which processes are made up of transactions, each transaction being an atomic execution unit that accesses concurrent objects. The important point is that the programmer has to focus her/his efforts only on the parts of code which have to be atomic execution units without worrying on the way the corresponding synchronization has to be realized.

Non-trivial STM systems allow transactions to execute concurrently and rely on the notion of commit/abort of a transaction in order to solve their conflicts on the objects they access simultaneously. In some cases, the management of aborted transactions is left to the programmer. In other cases, the underlying system scheduler is appropriately modified or an underlying contention manager is used in order that each transaction be (“practically always” or with high probability) eventually committed.

This work paper proposed a deterministic STM system in which (1) every invocation of a transaction is executed exactly once and (2) the notion of commit/abort of a transaction remains unknown to the programmer. This system, which imposes restriction neither on the design of processes nor or their concurrency pattern, can be seen as a step towards the design of a deterministic universal construction to execute transaction-based multiprocess programs on top of a multiprocessor. Interestingly, the proposed construction is lock-free (in the sense that it uses no lock).

This work has been published in ICDCN 2012 [30].

6.1.5. A transaction friendly binary search tree

Participants: Tyler Crain, Michel Raynal.
Transactions, which provide optimistic synchronization by avoiding the use of blocking, greatly simplify multicore programming. In fact, the programmer has simply to encapsulate sequential operations or existing critical sections into transactions to obtain a safe concurrent program. Programmers have thus started evaluating transactional memory using data structures originally designed for pessimistic (i.e., non-optimistic) synchronization, whose prominent example is the red-black tree library developed by Oracle Labs that is part of STAMP and microbench distributions. Unfortunately, existing data structures are badly suited for optimistic synchronization as they rely on strong structural invariants, like logarithmic tree depth, to bound the step complexity of pessimistically synchronized accesses. By contrast, this complexity does not apply to optimistically synchronized accesses thus making the invariants overly conservative. More dramatically, guaranteeing such invariants tends to increase the probability of aborting and restarting the same access before it completes. We introduced a concurrent binary search tree that breaks transiently its balance structural invariants for efficiency, a property we call transaction-friendly. This new tree outperforms the existing transaction-based version of the AVL and the red-black trees. Its key novelty stems from the decoupling of update operations: they are split into one transaction that modifies the abstraction state and multiple ones that restructure its tree implementation. The resulting transaction-friendly library trades aborts for few additional access steps and, in particular, it speeds up a transaction-based travel reservation application by up to 3.5X. This work was done in collaboration with Vincent Gramoli from EPFL Lausanne, and is described in [52].

6.1.6. Relations linking failure detectors associated with k-set agreement in message-passing systems

Participants: Achour Mostefaoui, Michel Raynal, Julien Stainer.

The k-set agreement problem is a coordination problem where each process is assumed to propose a value and each process that does not crash has to decide a value such that each decided value is a proposed value and at most k different values are decided. While it can always be solved in synchronous systems, k-set agreement has no solution in asynchronous send/receive message-passing systems where up to t ≥ k processes may crash.

A failure detector is a distributed oracle that provides processes with additional information related to failed processes and can consequently be used to enrich the computability power of asynchronous send/receive message-passing systems. Several failure detectors have been proposed to circumvent the impossibility of k-set agreement in pure asynchronous send/receive message-passing systems. Considering three of them (namely, the generalized quorum failure detector Σ_k, the generalized loneliness failure detector ζ_k and the generalized eventual leader failure detector Ω_k) this work investigates their computability power and the relations that link them. There are three main contributions: (a) it shows that the failure detector Ω_k and the eventual version of ζ_k have the same computational power; (b) it shows that ζ_k is realistic if and only if k ≥ n/2; and (c) it gives an exact characterization of the difference between ζ_k (that is too strong for k-set agreement) and Σ_k (that is too weak for k-set agreement). This work was published at SSS 2011 [45].

6.1.7. The price of anonymity: optimal consensus despite asynchrony, crash and anonymity

Participant: Michel Raynal.

This work [23], done in collaboration with François Bonnet, from JAIST, Japan, addresses the consensus problem in asynchronous systems prone to process crashes, where additionally the processes are anonymous (they cannot be distinguished one from the other: they have no name and execute the same code). To circumvent the three computational adversaries (asynchrony, failures and anonymity) each process is provided with a failure detector of a class denoted ψ, that gives it an upper bound on the number of processes that are currently alive (in a non-anonymous system, the classes ψ and P—the class of perfect failure detectors—are equivalent).

The first part presents a simple ψ-based consensus algorithm where the processes decide in 2t + 1 asynchronous rounds (where t is an upper bound on the number of faulty processes). It then shows one of its main results, namely, 2t + 1 is a lower bound for consensus in the anonymous systems equipped with ψ. The second contribution addresses early-decision. The paper presents and proves correct an early-deciding algorithm where the processes decide in min(2f + 2, 2t + 1) asynchronous rounds (where f is the actual number
of process failures). This leads to think that anonymity doubles the cost (wrt synchronous systems) and it is conjectured that min(2f + 2, 2t + 1) is the corresponding lower bound.

The work finally considers the $k$-set agreement problem in anonymous systems. It first shows that the previous $\psi$-based consensus algorithm solves the $k$-set agreement problem in $R_{t} = 2 \left\lfloor \frac{t}{k} \right\rfloor + 1$ asynchronous rounds. Then, considering a family of failure detector classes $\{\psi_{t}\}_{0 \leq t < k}$ that generalizes the class $\psi(= \psi_{0})$, the paper presents an algorithm that solves the $k$-set agreement in $R_{t,\ell} = 2 \left\lfloor \frac{1}{\ell - t} \right\rfloor + 1$ asynchronous rounds. This last formula relates the cost ($R_{t,\ell}$), the coordination degree of the problem ($k$), the maximum number of failures ($t$) and the the strength ($\ell$) of the underlying failure detector.

### 6.1.8. On the road to the Weakest Failure Detector for $k$-Set Agreement in Message-passing Systems

**Participant:** Michel Raynal.

In the $k$-set agreement problem, each process (in a set of $n$ processes) proposes a value and has to decide a proposed value in such a way that at most $k$ different values are decided. While this problem can easily be solved in asynchronous systems prone to $t$ process crashes when $k > t$, it cannot be solved when $k \leq t$. Since several years, the failure detector-based approach has been investigated to circumvent this impossibility. While the weakest failure detector class to solve the $k$-set agreement problem in read/write shared-memory systems has recently been discovered (PODC 2009), the situation is different in message-passing systems where the weakest failure detector classes are known only for the extreme cases $k = 1$ (consensus) and $k = n - 1$ (set agreement).

This work [22], done in collaboration with François Bonnet, from JAIST, Japan, has four contributions whose aim is to help pave the way to discover the weakest failure detector class for $k$-set agreement in message-passing systems. These contributions are the following. (a) The first is a new failure detector class, denoted $\Pi_{k}$, that is such that $\Pi_{1} = \Sigma \times \Omega$ (the weakest class for $k = 1$), and $\Pi_{n-1} = \mathcal{L}$ (the weakest class for $k = n - 1$). (b) The second is an investigation of the structure of $\Pi_{k}$ that shows that $\Pi_{k}$ is the combination of two failures detector classes $\Sigma_{k}$ (that is new) and $\Omega_{k}$ (they generalize the previous “quorums” and “eventual leaders” failure detectors classes, respectively). (c) The third contribution concerns $\Sigma_{k}$ that is shown to be necessary requirement (as far as information on failure is concerned) to solve the $k$-set agreement problem in message-passing systems. (d) Finally, the last contribution is a $\Pi_{n-1}$-based algorithm that solves the $(n-1)$-set agreement problem. This algorithm provides us with a new algorithmic insight on the way the $(n-1)$-set agreement problem can be solved in asynchronous message-passing systems. It is hoped that these contributions will help discover the weakest failure detector class for $k$-set agreement in message-passing systems.

### 6.1.9. A non-topological proof for the impossibility of $k$-set agreement.

**Participant:** Armando Castañeda.

This work was done in collaboration with Hagit Attiya, from Technion, Haifa, Israel. In the $k$-set agreement task each process proposes a value, and each correct process has to decide a value which was proposed, so that at most $k$ distinct values are decided. Using topological arguments it has been proved that $k$-set agreement is unsolvable in the asynchronous wait-free read/write shared memory model, when $k < n$, the number of processes.

This work [34] focuses on a simple, non-topological impossibility proof of $k$-set agreement. The proof depends on two simple properties of the immediate snapshot executions, a subset of all possible executions, and on the well known handshaking lemma stating that every graph has an even number of vertices with odd degree.

The paper was presented in the 13th Int’l Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS’11) in Grenoble, France. The journal version of the paper was submitted to Theoretical Computer Science.
6.1.10. **Enriching the reduction map of sub-consensus tasks**

**Participants:** Armando Castañeda, Damien Imbs, Michel Raynal.

This work [51] was done in collaboration with Sergio Rajsbaum from the Universidad Nacional Autonoma de Mexico.

Understanding the relative computability power of tasks, in the presence of asynchrony and failures, is a central concern of distributed computing theory. In the wait-free case, where the system consists of \( n \) processes and any of them can fail by crashing, substantial attention has been devoted to understanding the relative power of the subconsensus family of tasks, which are too weak to solve consensus for two processes. The first major results showed that set agreement and renaming (except for some particular values of \( n \)) cannot be solved wait-free in read/write memory. Then it was proved that renaming is strictly weaker than set agreement (when \( n \) is odd).

This work considers a natural family of subconsensus tasks that includes set agreement, renaming and other generalized symmetry breaking (GSB) tasks. It extends previous results, and proves various new results about when there is a reduction and when not, among these tasks. Among other results, the work shows that there are incomparable subconsensus tasks.

6.1.11. **Byzantine Consensus Decidability**

**Participants:** Achour Mostefaoui, Michel Raynal.

Solving the consensus problem requires in one way or another that the underlying system satisfies synchrony assumptions. Considering a system of \( n \) processes where up to \( t < n/3 \) may commit Byzantine failures, we proposed in [26] a necessary and sufficient synchrony assumption to solve consensus.

Such a condition is formulated with the notions of a symmetric synchrony property and property ambiguity. A symmetric synchrony property is a set of graphs, where each graph corresponds to a set of bi-directional eventually synchronous links among correct processes. Intuitively, a property is ambiguous if it contains a graph whose connected components are such that it is impossible to distinguish a connected component that contains correct processes only from a connected component that contains faulty processes only. The paper connects then the notion of a symmetric synchrony property with the notion of eventual bi-source, and shows that the existence of a virtual \( \diamond [t + 1] \) bi-source is a necessary and sufficient condition to solve consensus in presence of up to \( t \) Byzantine processes in systems with bi-directional links and message authentication. Finding necessary and sufficient synchrony conditions when links are timely in one direction only, or when processes cannot sign messages, still remains open (and very challenging) problems.

6.1.12. **Solving \( k \)-set agreement in message-passing systems**

**Participants:** Achour Mostefaoui, Michel Raynal, Julien Stainer.

The \( k \)-set agreement problem is a coordination problem where each process is assumed to propose a value and each process that does not crash has to decide a value such that each decided value is a proposed value and at most \( k \) different values are decided. While it can always be solved in synchronous systems, \( k \)-set agreement has no solution in asynchronous send/receive message-passing systems where up to \( t \geq k \) processes may crash.

A failure detector is a distributed oracle that provides processes with additional information related to failed processes and can consequently be used to enrich the computability power of asynchronous send/receive message-passing systems. Several failure detectors have been proposed to circumvent the impossibility of \( k \)-set agreement in pure asynchronous send/receive message-passing systems. Considering three of them (namely, the generalized quorum failure detector \( \Sigma_k \), the generalized loneliness failure detector \( \mathcal{L}_k \) and the generalized eventual leader failure detector \( \Omega_k \)), we investigated their computability power and the relations that link them in [45]. It has three main contributions: (a) it shows that the failure detector \( \Omega_k \) and the eventual version of \( \mathcal{L}_k \) have the same computational power; (b) it shows that \( \mathcal{L}_k \) is realistic if and only if \( k \geq n/2 \); and (c) it gives an exact characterization of the difference between \( \mathcal{L}_k \) (that is too strong for \( k \)-set agreement) and \( \Sigma_k \) (that is too weak for \( k \)-set agreement).
6.1.13. Efficient Implementations of Concurrent Objects
Participants: Achour Mostefaoui, Michel Raynal.

As introduced by Taubenfeld, a contention-sensitive implementation of a concurrent object is an implementation such that the overhead introduced by locking is eliminated in the common cases, i.e., when there is no contention or when the operations accessing concurrently the object are non-interfering. In [44], we present a methodological construction of a contention-sensitive implementation of a concurrent stack. In a contention-free context a push or pop operation does not rest on a lock mechanism and needs only six accesses to the shared memory. In case of concurrency a single lock is required. Moreover, the implementation is starvation-free (any operation is eventually executed). The paper, that presents the algorithms in an incremental way, visits also a family of liveness conditions and important concurrency-related concepts such as the notion of an abortable object.

6.2. Large-scale and user-centric distributed system

This section summarizes the major results obtained by the team in 2011 in the context of large-scale distributed systems and social networks. This includes the results obtained within the GOSSPLE ERC project, which encompass two types of social networks: explicit and implicit.

Explicit networks connect users based on explicit social connections. In Facebook or MySpace, users issue and accept friendship requests. In Twitter, they decide that they wish to follow the tweets of specific users. In all cases, the topology of the resulting network reflects the choices of users and often consists of links that already exist between real people. Explicit networks are therefore very useful in reinforcing and exploiting existing connections but provide little support for discovering new content.

Implicit networks complement explicit ones by providing each user with a set of anonymous acquaintances that share similar interests, that visit similar websites or that have otherwise similar profiles. Different from explicit networks, implicit ones are naturally suited to support the discovery of new content. In previous work [1], we exploited this network to improve web navigation. In the following, we consider additional applications encompassing news dissemination, online transactions, and recommendation.

6.2.1. WhatsUp: P2P news recommender
Participants: Antoine Boutet, Davide Frey, Anne-Marie Kermarrec.

The main application in the context of GOSSPLE is WhatsUp, an instant news system designed for a large-scale network with no central authority. WhatsUp builds an implicit social network based on the opinions users express about the news items they receive (like-dislike). This is achieved through an obfuscation mechanism that does not require users to ever reveal their exact profiles. WhatsUp disseminates news items through a novel heterogeneous gossip protocol that biases the choice of its targets towards those with similar interests and amplifies dissemination based on the level of interest in every news item. WhatsUp outperforms various alternatives in terms of accurate and complete delivery of relevant news items while preserving the fundamental advantages of standard gossip: namely simplicity of deployment and robustness. This work has been carried out in collaboration with Rachid Guerraoui from EPFL and was demonstrated during the different local events.

6.2.2. Personalized top-k processing
Participant: Anne-Marie Kermarrec.

Another way to improve the experience of users on the web is to personalize top-k queries. In collaboration with Xiao Bai and Vincent Leroy from Yahoo! Research in Barcelona and Rachid Guerraoui from EPFL Lausanne we, therefore, introduced P4Q, a fully decentralized gossip-based protocol to personalize query processing in social tagging systems. P4Q dynamically associates each user with social acquaintances sharing similar tagging behaviors. Queries are gossiped among such acquaintances, computed on the fly in a collaborative, yet partitioned manner, and results are iteratively refined and returned to the querier. Analytical and experimental evaluations convey the scalability of P4Q for top-k query processing, as well its inherent ability to cope with users updating profiles and departing. The work appeared in the ACM transactions of database systems [12].
6.2.3. Social Market  
**Participants:** Davide Frey, Arnaud Jegou, Anne-Marie Kermarrec.

The ability to identify people that share one’s own interests is one of the most interesting promises of the Web 2.0 driving user-centric applications such as recommendation systems or collaborative marketplaces. To be truly useful, however, information about other users also needs to be associated with some notion of trust. Consider a user wishing to sell a concert ticket. Not only must she find someone who is interested in the concert, but she must also make sure she can trust this person to pay for it. Social Market (SM) solve this problem by allowing users to identify and build connections to other users that can provide interesting goods or information and that are also reachable through a trusted path on a explicit social network like Facebook. This convergence of implicit and explicit networks yields TAPS, a novel gossip protocol that can be applied in applications devoted to commercial transactions, or to add robustness to standard gossip applications like dissemination or recommendation systems.

This work has been published at SSS 2011 [33], and an extended version bringing better performances and strong privacy guarantees have recently been submitted for publication.

6.2.4. Member classification and party characteristics in Twitter  
**Participant:** Antoine Boutet.

In modern politics, parties and individual candidates must have an online presence and usually have dedicated social media coordinators. In this context, real time member classification and party characterization, taking into account the dynamic nature of social media, are essential to highlight the main differences between parties and to monitor their activities, influences, structures, contents and mood. This work [53] was done in collaboration with E. Yoneki from Computer Lab, Cambridge, UK.

6.2.5. Graph Drawing and Visual Recommendations  
**Participants:** Anne-Marie Kermarrec, Afshin Moin.

An important aspect of social network is their graph structure. In a collaboration with Vincent Leroy (Yahoo! Research) and Gilles Tredan (TU Berlin) [41], we started from this structure to propose a decentralized gossip-based algorithm called SoCS (Social Coordinate Systems). SoCS achieves efficient distributed social graph embedding using a force-based graph embedding technique to extract communities from a graph. SoCS (i) scales to large dynamic graph, aggregating the computing power of individual nodes and, (ii) avoids a central entity controlling users sensitive data such as relations and preferences. We evaluated SoCS using two different force-based models and compare them in the context of a generated Kleinberg small-world topology. More specifically, we showed that the SoCS graph embedding enables to clearly distinguish between short and long-range links. We also evaluate SoCS against a real DBLP data set, showing that removed links are correctly predicted.

Graph structures are also at the basis of our work on energy/force-based models for graph visualization. We applied visualization both to social network and in the context of recommendation systems. In particular we are working on an SVD-like algorithm for drawing precise 2-dimensional visual recommendations based on Principal Component Analysis (PCA) and Curvilinear Component Analysis (CCA).

6.2.6. Private Similarity Computation in Distributed Systems: from Cryptography to Differential Privacy  
**Participants:** Mohammad Alaggan, Anne-Marie Kermarrec.

The use of personal data in the context of social networks raises important concerns about privacy. In a collaboration [24] with Sébastien Gambs from the CIDre team, we addressed the problem of computing the similarity between two users (a key operation in an implicit social network [1]) while preserving their privacy in a fully decentralized system and for the passive adversary model. First, we introduced a two-party protocol for privately computing a threshold version of the similarity and applied it to well-known similarity measures such as the scalar product and the cosine similarity. The output of this protocol is only one bit
of information telling whether or not two users are similar beyond a predetermined threshold. Afterwards, we explored the computation of the exact and threshold similarity within the context of differential privacy. Differential privacy is a recent notion developed within the field of private data analysis guaranteeing that an adversary that observes the output of the differentially private mechanism, will only gain a negligible advantage (up to a privacy parameter) from the presence (or absence) of a particular item in the profile of a user. This provides a strong privacy guarantee that holds independently of the auxiliary knowledge that the adversary might have. More specifically, we designed several differentially private variants of the exact and threshold protocols that rely on the addition of random noise tailored to the sensitivity of the considered similarity measure. We also analyzed their complexity as well as their impact on the utility of the resulting similarity measure.

6.2.7. Constellation: Programming decentralized social networks

Participants: François Taiani, Anne-Marie Kermarrec.

As they continue to grow, social and collaborative applications (e.g. twitter, Facebook, digg) are increasingly calling for disruptive distributed solutions than can cater for the millions of users these applications serve daily, in hundreds of countries, over a wide variety of devices. To address these challenges, fully decentralized versions of social and collaborative applications are progressively emerging that seek to provide naturally scalable solutions to deliver their services. Gossip protocols in particular appear as a natural solution to implement these decentralized versions, as they intrinsically tend to be highly resilient, efficient, and scalable. Social applications based on gossip have however been limited so far to relatively homogeneous systems: They typically rely on one similarity measure to self-organize large amount of distributed users in implicit communities, and thus offer powerful means to search, mine, and serve personalized data in a distributed manner.

We posit in this work [54] that we now need to move to more complex gossip-based social applications that can cater for different types of data and similarity, organized in multiple levels of abstraction. Exploring, designing, and evaluating such novel approaches is unfortunately time-consuming and error-prone. To help in this task, we have started to design a new programming language, Constellation, that seeks to simplify the realization and experimentation with social gossip-based applications. Constellation is based on two central observations: (i) future decentralized social applications will need to handle heterogeneous forms of data and self-organization, and (ii) to offer more powerful services, these applications will need to move beyond physical nodes to encompass richer data structures organized in virtualized levels of abstractions.

6.2.8. Leveraging content interconnections for efficient data storage.

Participants: Anne-Marie Kermarrec, Konstantinos Kloudas, François Taiani.

Traffic generated by User Generated Content (UGC) sharing sites, such as YOUTUBE, accounts for a substantial fraction of today’s global Internet load. This success has however brought a number of key technical challenges, crucial for system sustainability and user experience. One of them is the need to place content close to consumers, so that user perceived latency is reduced and bandwidth utilization is minimized. In a joint work with Kevin Huguenin, we try to tackle this problem by leveraging the fact that content hosted by these sites is interconnected, forming a content graph that as shown by former works, has an important impact on a file’s view pattern. In our work titled “Recommended nearby in UGC delivery networks: leveraging geographical and content locality”, we focused on YOUTUBE and we studied how two types of locality previously analyzed in isolation in UGC systems, namely content locality (a.k.a graph locality, induced by the related video feature) and geographic locality, are in fact correlated. Leveraging the above finding, we proposed a novel algorithm for replica placement that tries to predict where future views for a video will come from based on the video’s related videos and places its replicas accordingly. This work has been submitted for publication.

6.2.9. Transparent Componentization: High-level (Re)configurable Programming for Evolving Distributed Systems

Participants: François Taiani, Marin Bertier, Anne-Marie Kermarrec.
This work was done in collaboration Component frameworks and high-level distributed languages have been widely used to develop distributed systems, and provide complementary advantages: Whereas component frameworks foster composability, reusability, and (re)configurability; distributed languages focus on behavior, simplicity and programmability. We argue that both types of approach should be brought together to help develop complex adaptive systems, and we propose an approach to combines both technologies without compromising on any of their benefits. Our approach, termed Transparent Componentization [43], automatically maps a high-level distributed specification onto a underlying component framework. It thus allows developers to focus on the programmatic description of a distributed system’s behavior, while retaining the benefits of a component architecture. As a proof of concept, we present WhispersKit, a programming environment for gossip-based distributed systems. Our evaluation shows that WhispersKit successfully retains the simplicity and understandability of high-level distributed language while providing efficient and transparent reconfigurability thanks to its component underpinnings.

6.2.10. **Efficient peer-to-peer backup services through buffering at the edge**

**Participants:** Anne-Marie Kermarrec, Alexandre Van Kempen.

The availability of end devices of peer-to-peer storage and backup systems has been shown critical for usability and for system reliability in practice. This has led to the adoption of hybrid architectures composed of both peers and servers. Such architectures mask the instability of peers thus approaching the performances of client-server systems while providing scalability at a low cost. In this work [31] - done in collaboration with Erwan Le Merrer, Serge Defrance, Nicolas Le Scouarnec and Gilles Straub from Technicolor, Rennes, France - we advocate the replacement of such servers by a cloud of residential gateways, as they are already present in users’ homes, thus pushing the required stable components at the edge of the network. In our gateway-assisted system, gateways act as buffers between peers, compensating for their intrinsic instability. This enables to offload backup tasks quickly from the user’s machine to the gateway, while significantly lowering the retrieval time of backed up data. We evaluate our proposal using real world traces including existing traces from Skype and Jabber as well as a trace of residential gateways for availability, and a residential broadband trace for bandwidth. Results show that the time required to backup data in the network is comparable to a server-assisted approach, while substantially improving the time to restore data, which drops from a few days to a few hours. As gateways are becoming increasingly powerful in order to enable new services, we expect such a proposal to be leveraged on a short term basis.

6.2.11. **Commutative Replicated Data Type for Semantic Stores**

**Participant:** Stéphane Weiss.

This work has been done in collaboration with Khaled Aslan (Université de Nantes - Lina), Pascal Molli (Université de Nantes - Lina) and Hala Skaf-Molli (Université de Nantes - Lina).

Web 2.0 tools are currently evolving to embrace semantic web technologies. Blogs, CMS, Wikis, social networks and real-time notifications, integrate ways to provide semantic annotations and therefore contribute to the linked data and more generally to the semantic web vision. This evolution generates a lot of semantic datasets of different qualities, different trust levels and partially replicated. This raises the issue of managing the consistency among these replicas. This issue is challenging because semantic data-spaces can be very large, they can be managed by autonomous participants and the number of replicas is unknown. A new class of algorithms called Commutative Replicated Data Type are emerging for ensuring eventual consistency of highly dynamic content on P2P networks. We define C-Set [25] a CRDT specifically designed to be integrated in Triple-stores. C-Set allows efficient P2P synchronization of an arbitrary number of autonomous semantic stores.

6.2.12. **Building large scale platform for chemical program**

**Participants:** Marin Bertier, Achour Mostefaoui.

This work [28] was done in collaboration with the Myriads project team.
Chemical programming is a promising paradigm to design autonomic systems. Within such a paradigm, computations can be seen as chemical reactions controlled by a set of chemical rules. In other words, data are molecules of a chemical solution, reacting together to produce new data. Reactions take place in an implicitly parallel, and autonomic fashion.

Our objective was to design a distributed chemical platform bringing such concepts. This platform should be adapted to large scale distributed system to benefit at his best the inherent distribution of chemical program.
6. New Results

6.1. Aspects

Participants: Rémi Douence, Abdelhakim Hannousse, Ismael Mejía, Jacques Noyé, Angel Núñez, Nicolas Tabareau, Mario Südholt.

We have provided results on three subject matters in the domain of aspect-oriented software development: the foundations of aspects, aspect languages and their applications, as well as the use of aspects for the manipulation of distributed systems.

As a side note on the form of this document, the reader should be aware that much of our work reported in this section is not exclusive to AOSD but also contributes to software composition issues in a larger sense. This applies, in particular, to the results cited in the subsections on aspect languages and distributed aspects.

6.1.1. Foundations of Aspects

In the domain of foundations of AOSD, we have mainly provided new results on the preservation of formal correctness properties in the presence of aspects and how AOP can be modeled using category theory.

- **Property preservation in the presence of aspects.** In general aspects can arbitrarily change the semantics of the base program. We have identified categories of aspects that preserve class of properties of the base programs [17]. This approach makes it possible to prove once and for all that a category of aspects preserves a class of properties. The categories are defined with respect to the semantics of the woven program as well as with restricted aspect languages. In this latter case, languages are defined by grammars hence checking for property preservation boils down to a syntactic check for aspects. Classes of properties are defined using a subset of temporal logic both for sequential and concurrent programs. Our framework is abstract in that it does not depend on the actual programming language but only on conditions on its small step semantics.

- **AOP and category theory.** Aspect-Oriented Programming (AOP) started ten years ago with the remark that modularization of so-called crosscutting functionalities is a fundamental problem for the engineering of large-scale applications. Originating at Xerox PARC, this observation has sparked the development of a new style of programming that is gradually gaining traction. However, AOP lacks theoretical foundations to clarify new ideas showing up in its wake. We have proposed to put a bridge between AOP and the notion of 2-category to enhance the conceptual understanding of AOP [36], [46]. Starting from the connection between the \(\lambda\)-calculus and the theory of categories, we have provided an internal language for 2-categories and shown how it can be used to define the first categorical semantics for a realistic functional AOP language, called MinAML.

We have later taken advantage of this new categorical framework to introduce the notion of computational 2-monads for AOP [37]. We have illustrated their conceptual power by defining a 2-monad for Éric Tanter’s execution levels—which constitutes the first algebraic semantics for execution levels—and then introducing the first exception monad transformer specific to AOP that gives rise to a non-flat semantics for exceptions by taking levels into account.

- **Membranes for AOP.** AOP aims to enhance modularity and reusability in software systems offering an abstraction mechanism to deal with crosscutting concerns. However, in most aspect-oriented languages aspects have a global view of computation that actually introduces a strong coupling between advised code and aspect code, hampering modularity. Proposals that address this problem can be classified in two categories: the first one focuses on controlling aspect scoping, i.e. the visibility of join points to aspects, while the second one focuses on protecting software units from aspects. As a new approach, we have proposed programmable membranes (inspired by the work of Boudol for distributed processes) to control aspect influence over software systems as a uniform framework that unifies and extends previous approaches [49].
Aspects and invertible program restructurations. When one chooses a main axis of structural decomposition of a problem, the other axes become secondary. This is known as the tyranny of the dominant decomposition and this hinders program maintenance of secondary concerns. We propose to use automatic program transformations built on top of refactors in order to solve this tyranny issue [47]. This offers a new approach to the expression problem by always providing the right view to the programmers.

6.1.2. Aspect Languages

This year we have provided major results on the integration of programming features for objects, events and aspects, and aspect execution infrastructures. Furthermore, we have investigated the application of aspect languages to the Java security model, component-based software development and software product lines.

Integrating OOP, AOP and EBP: Object-Oriented Programming (OOP) has become the de facto programming paradigm. Event-Based Programming (EBP) and Aspect-Oriented Programming (AOP) complement OOP, covering some of its deficiencies when building complex software. Today’s applications combine the three paradigms. However, OOP, EBP and AOP have not yet been properly integrated. Their underlying concepts are in general provided as distinct language constructs, whereas they are not completely orthogonal. This lack of integration and orthogonality complicates the development of software as it reduces its understandability, its composability and increases the required glue code. [16] proposes an integration of OOP, EBP and AOP leading to a simple and regular programming model. This model integrates the notions of class and aspect, the notions of event and join point, and the notions of piece of advice, method and event handler. It reduces the number of language constructs while keeping expressiveness and offering additional programming options. ECaesarJ had previously been developed based on these ideas. [16] introduces a simpler variant of it, EJava, a plain extension of Java. [24] shows that these ideas can also be easily applied to Scala, leading to EScala, and focuses on the idea that the declarative events provided by the model can be seen as object-oriented events, which, unlike global typed events, obey to the basic OO principles (OO modular reasoning, encapsulation and late-binding). An efficient implementation of EScala is described, based on the idea that an event should not be produced in the absence of at least a consumer. This is equivalent to what could be programmed by hand using variants of the observer pattern, except that all the necessary scaffolding is provided by the language compiler and runtime rather than by the programmer.

Prototyping and composing Aspect Languages: [11] presents CALI (Common Aspect Languages Interpreter), a framework for rapid prototyping and composition of aspect languages based on interpreters. In practice the common interpreter is actually a thin interpretative layer built on top of Java and implemented as an AspectJ aspect. This interpreter implements common aspect mechanisms and leaves holes to be defined when developing concrete languages. The approach has been validated by implementing prototypes of significant subsets of well-known general-purpose and domain-specific aspect languages (AspectJ, EAOP, COOL and a couple of other small domain-specific aspect languages) and exploring variants of AspectJ. Languages implemented with CALI, for instance AspectJ and COOL, can then be easily composed.

Application to security: It is inevitable that some concerns crosscut a sizeable application, resulting in code scattering and tangling. This issue is particularly severe for security-related concerns: it is difficult to be confident about the security of an application when the implementation of its security-related concerns is scattered all over the code and tangled with other concerns, making global reasoning about security precarious. In [21], we consider the case of access control in Java, which turns out to be a crosscutting concern with a non-modular implementation based on runtime stack inspection. We describe the process of modularizing access control in Java by means of AOP. We first show a solution based on AspectJ that must rely on a separate automata infrastructure. We then put forward a novel solution via dynamic deployment of aspects and scoping strategies. Both solutions, apart from providing a modular specification of access control, make it possible to easily express other useful policies such as the Chinese wall policy. However, relying on expressive scope
control results in a compact implementation, which, at the same time, permits the straightforward expression of even more interesting policies.

- **Aspects and software components**: The relationship of aspects and components, as well as their integration as part of real-world infrastructures and application is still subject to many open issues. We have studied, in particular, how crosscutting concerns naturally appear when several architectural views must be considered at the same time. In this context, we have proposed a domain-specific language to specify these architectures. We have proposed an implementation of composable controllers in Fractal as well as composition operators that makes it possible to solve aspects interferences [25]. We have also shown how to formally model the implementation in Uppaal in order to statically check aspects interference [26]. This work has principally been undertaken in the context of the PhD thesis of Abdelhakim Hannouse [12] that has been defended in Sep. 2011.

Another fundamental issue that relates components and aspects is the question whether behavioral protocols can be used to concisely define crosscutting concerns of component-based systems. Furthermore, the protocols should then be instrumental to enable (automatic) reasoning about composition properties of component systems. As part of her PhD thesis (defended in Oct. 2011) [14], Dong Ha Nguyen has defined a notion of aspects that allow crosscutting concerns to be expressed in terms of a class of non-regular behavioral protocols, so-called visibly-pushdown languages. She has developed, most notably, a constructive way of building correct component-based systems that respect such aspect-modified behavioral protocols. Furthermore, she has shown this year how composition properties can generally be verified using model checking techniques.

- **Aspects and software product lines**: In [41], we take a closer look at the difficulties of feature-oriented modularisation of product lines and demonstrate, using a case study in the domain of home automation, how a better modularisation can be achieved with the ECAesarJ programming language, through a type-safe and stable decomposition of a broad spectrum of software abstractions: classes, methods, events, and state machines, based on late binding and mixin composition. A nice property of this modularisation is that it directly captures, at the implementation level, the high-level decomposition in features, without requiring the user to resort on complex transformation technology (the technology is embedded in the compiler).

Furthermore, ASCOLA members have co-edited and contributed to a comprehensive book on aspects, model-driven engineering and product lines [42] (see Sec. 6.2 for details).

### 6.1.3. Aspects for Distributed Systems

In the field of aspects for distributed systems, we have mainly extended the AWED model for distributed aspects (see Sec. 5.1) in order to define and implement a gray-box distributed composition model. We have also worked on a notion of distributed crosscuts that enable causality relationships to be captured using logical clocks.

- **Gray-box distributed composition**: Existing composition and coordination techniques for distributed applications typically support only interface-level (black-box) composition or arbitrary access to the implementation (gray-box or white-box composition). In [20], we have presented a structured approach to the composition of complex distributed software systems that require invasive modifications. Concretely, we have provided three contributions:
  - We have introduced a small kernel composition language for structured gray-box composition using invasive distributed patterns.
  - We have motivated that gray-box composition approaches should be defined and evaluated in terms of the flexibility and control they provide; a notion of degrees of invasiveness has been introduced to help assess this trade-off.
  - We have applied and evaluated it in the context of several studies involving two real-world software systems: benchmarking of grid algorithms with NASGrid and transactional replication with JBoss Cache. As a main result, we have shown that gray-box composition using
invasive distributed patterns allows the declarative and modular definition of evolutions of real-world applications that need moderate to high degrees of invasive modifications. We have then provided a first framework-based implementation based on our AWED tool and applied it to evolution tasks in OpenMRS, a health information system [31]. The composition framework supports the definition of expressive pattern-based invasive compositions based on the skeleton paradigm for distributed patterns. Concretely, we have shown that the composition framework enables the concise definition of an evolution scenario of OpenMRS that supports the consolidation of patient data from different distributed instances. Finally, we have studied a model that integrates distributed aspects and actors and thus improves on the robustness properties of existing models for distributed aspects [32].

- **Causal aspects for distributed applications.** In [45], we have applied logical clocks à la Lamport in order to define causality between distributed join points. This enables us to declaratively define pointcuts in a distributed context such as web-based applications in JavaScript. The definition of advice can then be simplified, because they are executed only when necessary and do not have to maintain and check crosscutting information.

### 6.2. Software Composition

**Participants:** Jean-Claude Royer, Hervé Grall, Christine Louberry, Mario Südholt.

ASCOLA’s work on software composition addresses the foundations of software composition methods, the definition and implementation of concrete composition techniques and their application to different functionalities and application domains. This year we have contributed to the questions of how to identify components in legacy software, how to effectively deploy and reconfigure component-based software in pervasive environments, and how to apply software composition techniques to Cloud security as well as software product lines.

- **Component identification.** The communication integrity property is one of the major principles to implement software architectures, however, there is a lack of tooling for assessing the quality of components codes. To cope with this issue, we defined, in [23], a Java component model and a tool for identifying component types based on the communication integrity property. We apply it to several case studies and compare the result with the SOMOX component recovery tool.

- **QoS-driven deployment and reconfiguration of pervasive applications.** [28] presents the Kalimucho platform, a platform for the dynamic reconfiguration of applications on mobiles and constraints devices. First this article focuses on the heuristics implemented by Kalimucho. They support finding a configuration and a deployment matching two criteria: utility and durability. Moreover, we present a case study to experiment the approach of Kalimucho. It confirms that Kalimucho provides a satisfactory execution time for mobile devices. A second result we have contributed in this context is a two-dimensional QoS model for Kalimucho that supports the QoS-driven deployment of mobile applications [29]. After presenting the definitions of the context and the quality of service considered in this work, this article describes the QoS model and the process allowing finding the best configuration and deployment. Finally, we present several results concerning the execution time of the deployment process of Kalimucho with different devices.

- **Securing the Cloud: cross domain and multi-level concerns.** The evolution of new deployment architectures, as illustrated by the move towards mobile platforms and the Internet Of Services, and the introduction of new security regulations (imposed by national and international regulatory bodies, such as SOX4 or BASEL5) are important constraints in the design and development of service composition. In such a context, it is not sufficient to apply the corresponding adaptations only at the service orchestration or at the choreography level; there is also the need for controlling the impact of new security requirements on several architectural layers. In [35] we have presented a new service model that supports the clean modularization of such crosscutting security concerns.
• **Software composition and product lines.** As part of the AMPLE project (see [http://www.ample-project.net](http://www.ample-project.net)) we have co-edited a book [42] covering the main scientific solutions and techniques on new methods and techniques for software product lines that have been developed in the project. This project aims at improving traditional software product lines engineering using advanced software engineering namely model-driven and aspect-oriented engineering approaches. Software composition takes an important part in this book and it appears in several chapters with models, events, aspects and components.

Chapter [43] provides an overview of software product lines, model-driven engineering and aspect-oriented software development. The challenges to address and the expected benefits are drawn, this is concluded by an overview of the AMPLE approach and its tool chain.

Chapter [40] reviews the specificities of traceability in a product line context starting by identifying the challenges of maintaining traceability for traditional system development and for software product lines. This work defines the concepts that should guide the adoption of a traceability environment for product lines and illustrates these specifications with a concrete example of a traceability repository. It also provides examples of scenarios that use this traceability environment to solve concrete problems.

### 6.3. Cloud Computing, Virtualization and Data Centers

**Participants:** Frederico Alvares, Gustavo Bervian Brand, Thomas Chavrier, Fabien Hermenier, Adrien Lèbre, Thomas Ledoux, Guillaume Le Louët, Jean-Marc Menaud, Hien Nguyen Van, Rémy Pottier, Flavien Quesnel.

In the context of Cloud computing ASCOLA members have principally worked this year on capacity planning solutions for large scale distributed systems. Capacity planning is the process of planning, analyzing, sizing, managing and optimizing capacity to satisfy demand in a timely manner and at a reasonable cost.

Applied to distributed systems like the Cloud, a capacity planning solution must mainly provide necessary resources for the proper execution of applications and respect service-level agreements in a large distributed environment.

The main challenges in this context are: scalability, fault tolerance and reactivity of the solution in a large-scale distributed system; analyzing, sizing, and optimizing resources to minimize the cost (energy, human, hardware etc.); and profiling, adapting application to ensure the quality of service (throughput, response time, availability etc.).

Our solutions are mainly based on virtualized infrastructures. Our main results concern the management and the execution of applications by leveraging virtualization capabilities on cloud infrastructures and the investigation of solutions that aim at optimizing the trade-off between performance and energy costs of both applications and Cloud resources.

#### 6.3.1. Virtualization and Job Management

This year, in cooperation with the Myriads project-team from INRIA Rennes-Bretagne Atlantique, we have continued to address resource-management issues concerning the federation of very large scale platforms. We have completed our approach aiming at the automatic adaptation of both hardware and software resources to the needs of the applications through a unique method. For each application, scientists describe the requirements in terms of both hardware and software expectations through the definition of a Virtual Platform (VP) and a Virtual System Environment (VSE) [18].

In addition, we started to address the design and the implementation of a fully distributed cloud OS. Designing and implementing such models is a tedious and complex task. However, as well as research studies on OSes and hypervisors are complementary at the node level, we showed that virtualization frameworks can benefit from lessons learned from distributed operating system proposals [34]. Leveraging this preliminary result, we designed and developed a first proof-of-concept of a fully distributed scheduler [33]. This system makes it possible to schedule VMs cooperatively and dynamically in large scale distributed systems. Preliminary results showed that our scheduler was more reactive. This building block is a first element of a more complete
cloud OS, entitled DISCOVERY (DIStributed and COoperative mechanisms to manage Virtual EnviRonments autonominallyY) [30]. The ultimate goal of this system is to overcome the main limitations of the traditional server-centric solutions. The system, currently under development, relies on a peer-to-peer model where each agent can efficiently deploy, dynamically schedule and periodically checkpoint the virtual environments s/he manage.

We have contributed new results on the Entropy software and extended our solution VM that features dynamic consolidation. In [44] and [38] we extended our dynamic consolidation manager to take into account not only resource constraints but also the placement constraints of highly available (HA) applications. In fact, most previous dynamic consolidation systems optimize the placement of the VMs according to their resource usage but do not consider the application placement constraints that are required to achieve both high availability and scalability. Our approach provides efficient dynamic consolidation while guaranteeing to the application administrator that placement requirements will be satisfied and relieving the data center administrator of the burden of considering the constraints of the applications when performing maintenance.

In the same domain, Jean-Marc Menaud has defended his habilitation (HdR - Habilitation à diriger des recherches) [13] on Jun. 22, 2011. One part of this HDR focuses on dynamic adaptation strategies for cluster administration. We have proposed a dedicated language for virtual machines management and one particular feature of our solution is to use a constraint solver to provide an appropriate placement. Based on these results, our recent contributions address the problem of data center energy consumption.

Moreover, we have continued to analyze how energy concerns can be addressed in large scale distributed infrastructures.

### 6.3.2. Optimization of Energy Consumption in Data Centers

As a direct consequence of the increasing popularity of Cloud Computing solutions, data centers are growing at a fast rate and hence have to face difficult energy consumption issue. Available solutions rely on Cloud Computing models and virtualization techniques to scale up/down applications based on their performance metrics. Although those proposals can reduce the energy footprint of applications and, by transitivity, of cloud infrastructures, they do not consider the internal characteristics of applications to finely define a trade-off between QoS properties of applications and their energy footprint. In [22], we propose a self-adaptation approach that considers both application internals and system properties to reduce the energy footprint in cloud infrastructures. Each application and the infrastructure are equipped with their own control loop, which allows them to autonomously optimize their executions. Simulations show that the approach may lead to appreciable energy savings without interfering on application provider revenues.

### 6.3.3. Cloud Computing and SLA Management

Cloud computing is a paradigm for enabling remote, on-demand access to a set of configurable computing resources as a service. The pay-per-use model enables service providers to offer their services to customers at different Quality-of-Service (QoS) levels. These QoS parameters are used to compose service-level agreements (SLAs) between a service provider and a service consumer. A main challenge for a service provider is to manage SLAs for its service consumers, i.e., automatically determine the appropriate resources required from the lower layer in order to respect the QoS requirements of the consumers. In [27], we have proposed an optimization framework driven by consumer preferences to address the SLA dependencies problem across the different cloud layers as well as the need of flexibility and dynamicity required by the domain of Cloud computing. Our approach aims at selecting the optimal vertical business process designed using cross-layer cloud services and enforcing SLA dependencies between layers. Experimental results demonstrate the flexibility and effectiveness of our approach.

### 6.4. Foundations of program semantics

**Participants:** Nicolas Tabareau, Guilhem Jaber.

ASCOLA team members have contributed several results to the foundations of program semantics.
6.4.1. Program Equivalences

Reasoning about program equivalence is a major problem in semantics. This very old topic has many applications, e.g., program verification, compiler correctness or representation independence. It has been understood since the late 1960s that tools and structures arising in mathematical logic and proof theory can usefully be applied to the development of reasoning principles for program equivalence. In recent years, based on the seminal work of Pitts and Stark, the notion of logical relation appeared to be very fruitful for proving the equivalence of programs in the presence of features like general recursive types, general (higher-order) mutable references, and first-class continuations. We have studied the notion of logical relations for proving program equivalences of low level machine codes [39].

We have also developed a forcing theory inspired by Cohen’s forcing to increase the power of a semantics model just as the latter makes it possible to enrich a set theoretical universe. In this way, we can define a new generation of logical relations—that can be introduced modularly using forcing theory—to be used for proving program equivalence for low level languages, concurrent languages or domain specific languages. [48]
6. New Results

6.1. Model Transformation

Model transformation and in particular our ATL model transformation language continues playing a key role in our MDE strategy. During 2011 the new results in this area have been:

- The development of an execution algorithm for the lazy execution of ATL transformations [31]. The increasing adoption of Model-Driven Engineering in industrial contexts highlights scalability as a critical limitation of several MDE tools. When these tools are built around model-to-model (M2M) transformations, the efficiency of the transformation engine risks to become a performance bottleneck for the whole MDE environment. This new execution mode solves this problem by providing on-demand execution of model transformations. The computation required to generate a data element of the target model is only triggered once the user requests to read that data element.

- Towards a general semantics for transformation languages. In the mid-term, we would like to be able to achieve interoperability among existing transformation languages (e.g. to create mixed transformations or simply to compare alternative transformation solutions for the same scenario). In this direction, we have worked on a general composition semantics for rule-based transformation languages [32] and an extensive survey of inheritance support in existing languages [33]. Moreover, to make ATL closer to other transformation languages, in special graph transformations, we have developed a new ATL refining mode [43] that allows executing in-place transformations.

- Also relevant, this year AtlanMod has coorganized the main international conference on model transformations [41].

6.2. Model Representation

As part of our work on core techniques for the specification and internal representation of models we would like to highlight:

- Improved model typing. In order to represent metadata more accurately, we have worked on a functional typing system for megamodeling [18]. The basic idea is to consider transformations as functions, and to give them functional types that can facilitate the reasoning on them.

- Virtual EMF [21] for the transparent Composition, Weaving and Linking of Models. Virtual EMF facilitates the global manipulation of an heterogeneous but related set of models by providing the illusion of having a single and unified view of the modeled system.

- EMF Profiles [25], a mechanism to enable users annotate existing models without polluting them (i.e. without directly changing their content).

- Batch scripting support for the retrieval and manipulation of models from model repositories thanks to our new language MoScript [24].

6.3. Model Quality

Our work on model quality defends the idea that there is not a silver bullet technique to verify the quality of models and that a combined approach offers the best trade-off. Lightweight techniques can provide a quick feedback even if it is partial and, when necessary, more complete ones can be utilized to complement the results. In this sense, this year we have developed:

- Lightweight techniques for the verification of ATL transformations [27] and UML Executable models [26].
• Proposed a method for the automatic generation of correct dynamic models for a given domain [11]. Basic operations to modify the information the software needs to store and manipulate about the domain are generated. The generation process ensures that all generated operations are strongly executable.

• Created a brand new Eclipse Lab project aimed at facilitating the verification of any kind of (EMF-based) model with the EMFtoCSP tool (see the tool description in the tools section of this report).

• And applied some of these ideas to the verification and testing of web applications [15].

6.4. Domain Specific Language

There is a growing interest in the research community on the definition of a new generation of language workbenches to facilitate the definition of non-trivial domain specific languages. In this area, we are working on the definition of a set of quality properties that will help language designers to validate if their language is well-adapted to the needs of the language users. This is done by mining repositories containing a corpus of examples of the language.

To begin with we have been reusing the notion of software clones [16] [30], a well-known property in the programming domain and see if it is also meaningful for languages at the model level [29].

6.5. Reverse Engineering

We have continued our work on MoDisco, especially wrt the dissemination activities around the platform [19] [40] and the extension of the tooling support mainly through the spin-off of the EMFFacet Eclipse project, explained in the tools section of this report.

6.6. Systems interoperability

MDE can be used as an intermediate representation between two different technical spaces / platforms / tools to facilitate their interoperability. During this year, we have followed this approach in the following results:

• Automation of the interactions with APIs by automatically discovering and expressing them as models thanks to our API2MOL approach [14].

• Bridging the business and the technical domains, developing among others the Portolan tool [42] for the cartography of Information Systems.

• MDE itself can benefit from the work and techniques available in a different technical space. In this context we have combined MDE and constraint programming to see how the combination improves the solution of classical problems like the configuration of a set of components/plug-ins [22]. We have even organized a workshop on the topic of merging MDE and logic programming [20] to better understand how they can benefit each other.
6. New Results

6.1. Intrusion Detection

Metamorphic codes:

In [12], we have proposed an advance code obfuscation technique for metamorphic codes (i.e., codes that automatically recode themselves each time they propagate or are distributed). We have shown that the detection of such codes was a problem for classical nowadays static detection tools. In [25], we focus on a new dynamic detection approach which allows to detect variants produced by our metamorphic engine. In addition, our approach can detect unknown malware as long as their behavior approaches that of a known malware. For this, we propose to use a measure of similarity between program behaviors. This measure is obtained by lossless compression of execution traces in terms of system calls.

Intrusion Detection based on an Analysis of the Flow Control:

In [13], intrusion detection mechanisms based on the construct a model of normal behavior of the supervised entity are studied. Such a model is used during the detection phase to raise an alarm when a deviation is observed. This approach allows to detect unknown attacks.

The most common anomaly detection mechanisms at application level consist in detecting a deviation of the control-flow of a program. A popular method to detect such anomaly is the use of application sequences of system calls. However, such methods do not detect mimicry attacks or attacks against the integrity of the system call parameters. To enhance such detection mechanisms, we propose in [27] an approach to detect in the application the corruption of data items that have an influence on the system calls. This approach consists in building automatically a data-oriented behavior model of an application by static analysis of its source code. The proposed approach is illustrated on various examples, and an injection method is experimented to obtain an approximation of the detection coverage of the generated mechanisms.

Most of today’s MAC implementations can be turned into permissive mode, where no enforcement is performed but alerts are raised instead. This behavior is very close to an anomaly IDS except that the system is configured through a MAC policy. MAC implementations such as SELinux and AppArmor come with a default policy including real life and practical rules ready to be used as is or as a basis for a custom policy. In [30], we first propose an extension of an IDS based on information flow control. We address issues concerning programs execution and improve its expressiveness in terms of security policy. This extended model can be configured to reach a wide variety of different security goals. Particularly, it allows for information flow checking based on users and/or programs dependent policy rules. Furthermore, suspicious modification of binary programs can be detected to avoid malware execution. We also propose an algorithm for deriving an AppArmor MAC policy into an information flow policy, and thus get the advantage of having a ready to use policy offering good security. An integration within Android is described in [37].

Flow based Interpretation of Access Control Policies:

In [32], we introduce a formal property characterizing access control policies for which the interpretations of access control as mechanism over objects and as mechanism over information contained into objects are similar. This leads us to define both a flow based interpretation of access control policies and the information flows generated during the executions of a system implementing an access control mechanism. When these two interpretations are not equivalent, we propose to add a mechanism dedicated to illegal information flow detection to the mechanism of access control over objects. Such a mechanism is parameterized by the access control policy and is proved sound and complete. We also briefly describe two real implementations, at two levels of granularity, of our illegal flow detection mechanism: one for the Linux operating system and one for the Java Virtual Machine.

Intrusion Detection based on Invariants:
RRABIDS (Ruby on Rails Anomaly Based Intrusion Detection System) [40] is an application level intrusion detection system for applications implemented with the Ruby on Rails framework. The goal of this intrusion detection system is to detect attacks against data in the context of web applications. This anomaly based IDS focuses on the modeling of the normal application profile using invariants. These invariants are discovered during a learning phase. Then, they are used to instrument the web application at source code level, so that a deviation from the normal profile can be detected at run-time. On simple examples we show how the approach detects well known categories of web attacks that involve a state violation of the application, such as SQL injections. Finally, an assessment phase is performed to evaluate the accuracy of the detection provided by the proposed approach.

Alert Correlation:
Alert correlation is a crucial problem for monitoring and securing computer networks. It consists in analyzing the alerts triggered by intrusion detection systems (IDSs) and other security related tools in order to detect complex attack plans, discover false alerts, etc. The huge amounts of alerts raised continuously by IDSs and the impossibility for security operators to efficiently analyze them requires tools for eliminating false and redundant alerts on the one hand and prioritize them according to the detected activities? dangerousness and preferences of the analysts on the other hand. In [35], we describe an architecture that combines AI-based approaches for representing and reasoning with security operators’ knowledge and preferences. Moreover, this architecture allows to combines experts’ knowledge with machine learning and classifier based tools. This prototype collects the alerts raised by security related tools and analyzes them automatically.

Trust-Based IDS for the AODV Protocol:
Routing in ad hoc networks is based on mutual trust between collaborating nodes. Security problems arise when supposedly honest nodes lie deliberately to maximize their profit. In [11], we are interested in detecting misbehaving nodes within the ad hoc routing protocol AODV. We propose and implement a real-time intrusion detection system based on implicit trust relations: a node implementing this system collects its neighbors’ routing messages and reasons on them to decide on their honesty. We also evaluate our implementation, and, based on simulations, show that the system we have developed to detect dishonest behavior is efficient.

Modelization and Simulation of Zombies Behaviours:
In [26], we study the modelization and simulation of zombie machines for the evaluation of Network Intrusion Detection Systems (NIDS), used to detect botnets. We propose an automatic method to infer zombies behaviours through the analysis of messages exchanged with their masters. Once computed, a model provides a way to generate realistic and manageable traffic, which is mandatory for an NIDS evaluation. We propose to use a Stochastic Mealy Machine to model zombies behaviours, and an active inference algorithm to learn it. With our approach, it is possible to generate a realistic traffic corresponding to the communications of botnets while ensuring its controllability in the context of an NIDS evaluation.

6.2. Privacy
Computer privacy is a domain where the education and information of the general public is paramount. In this perspective, through [44] we have participated to the popularization effort in the area, by exposing a survey of accessible computing tools allowing users to better protect their online privacy.

Formal Privacy Policies and Logical Tools:
One of the obstacles to the improvement of the privacy level in distributed applications is the lack of expressiveness, usability and enforceability of the associated policies. This new research track aims at designing better privacy policies for complex systems, more adapted to the specific needs of personal data protection regulations and easier to enforce in a distributed fashion. Logical languages, in particular, are considered as interesting candidates because of the reasoning capabilities attached to the formalisms, allowing autonomous peers to perform efficient, privacy-aware planning. [18] is a contribution to the modal logics used to model formal norms, focusing on specific deadline-related temporal notions often encountered in privacy policies. In [39], we propose an ambitious, collaborative research project based on an epistemic view of
the privacy laws and regulations, which should lead to the design of several tools, including policy writing assistants and validation software. [24] is a generic work in the domain of formal policies, where we propose a logical model of various concepts of responsibility in an organizational framework featuring obligation delegation. This kind of framework is intended to model the handling of complex policies in real-life human institutions.

Privacy in Social Networking Sites:

Social Networking Sites (SNS), such as Facebook and LinkedIn, have become the established place for keeping contact with old friends and meeting new acquaintances. As a result, a user leaves a big trail of personal information about him and his friends on the SNS, sometimes even without being aware of it. This information can lead to privacy drifts such as damaging his reputation and credibility, security risks (for instance identity theft) and profiling risks. Another research challenge stems from the fact that in the digital world where it is possible to copy the information as often as desired, it is not easy to control how information is disseminated once it is out on the Internet. In an ongoing collaboration [23] with Ai Thanh Ho and Esma Aïmeur (Université de Montréal), we investigate tools that can help user to maintain the sovereignty of their data on the World Wide Web. We also introduce PrivacyMarker, an approach drawing on the concept of provenance and accountability to protect user privacy on SNS. More precisely, it is possible to imagine that by a combination of logs and techniques such as watermarking and traitor-tracing schemes, the dissemination of information can be (at least partially) controlled and that in case of a privacy breach, it is possible to identify which persons are potentially suspect because they have previously accessed this information.

Geo-privacy:

A geolocalised system generally belongs to an individual and as such knowing its location reveals the location of its owner, which is a direct threat against his privacy. To protect the privacy of users, a sanitization process, which adds uncertainty to the data and removes some sensible information, can be performed but at the cost of a decrease of utility due to the quality degradation of the data. In a joint work [16] with Marc-Olivier Killijian and Miguel Nunez del Prado (LAAS-CNRS), we describe GEPETO (for GEoPrivacy-Enhancing TOolkit), a flexible open source software which can be used to visualize, sanitize, perform inference attacks and measure the utility of a particular geolocalised dataset. We also introduce a mobility model that we coin as mobility Markov Chain, which can represent in a compact yet precise way the mobility behaviour of an individual. Finally, we describe an algorithm for learning such a structure from the mobility traces of an individual.

Geosocial networks are relatively new compared to the more “traditional” (i.e. non-geolocated) social networking sites such as Facebook or LinkedIn that have been around since more than 6 years, but they are currently growing relatively fast along with the widespread development of other geolocated applications and technologies. In a study [29] done in cooperation with Olivier Heen (Technicolor) and Christophe Potin, we provide a comparative analysis of some existing geosocial networks with respect to privacy in order to (1) highlight some of privacy issues that are raised by the fast development of these system and (2) propose recommendations that could be integrated in the design of these systems to enhance the privacy of their users based on this analysis.

Privacy in Distributed Systems:

In a joint work [19] with Anne-Marie Kermarrec and Mohammad Alaggan (team INRIA ASAP), we address the problem of computing the similarity between two users (according to their profiles) while preserving their privacy in a fully decentralized system and for the passive adversary model. First, we introduce a two-party protocol for privately computing a threshold version of the similarity and apply it to well-known similarity measures such as the scalar product and the cosine similarity. The output of this protocol is only one bit of information telling whether or not two users are similar beyond a predetermined threshold. Afterwards, we explore the computation of the exact and threshold similarity within the context of differential privacy, a recent notion developed that provides a strong privacy guarantee that holds independently of the auxiliary knowledge that the adversary might have. More specifically, we design several differentially private variants of the exact and threshold protocols and we also analyze their complexity as well as their impact on the utility
of the resulting similarity measure. Finally, we provide experimental results validating the effectiveness of the proposed approach on real datasets.

Other ongoing work tackles the problem of computing an aggregation function in a secure and scalable way in a distributed network [42] (joint work with Rachid Guerraoui, Hamza Harkous, Florian Huc and Anne-Marie Kermarrec).

6.3. Accidental and Malicious Faults in Distributed Systems

Induced Churn to Face Malicious Behaviors:

In reputation mechanisms, ensuring durable access to feedbacks is a first barrier against simple attacks. To bias the reputation mechanism, an adversary can create and use several distinct identities. In that case, if the reputation mechanism is solely based on statistical measurements, the trustworthiness can be violated. Our contribution is centered around the study of robust mechanisms that can resist such attacks.

Toward this goal, we have first investigating the problem of uniform sampling in large scale open systems in presence of adversarial nodes. Uniform sampling ensures that any individual in a population has the same probability to be selected as sample. Uniform sampling finds its root in many problems such as data collection, dissemination, load balancing, and data-caching.

By relying on the topological properties of structured peer-to-peer systems, it has been shown that it is possible to guarantee with high probability that any node is equally likely to appear in the local view of each other honest node in a number of rounds polynomial in the size of the system. This is achieved by imposing nodes to frequently depart from their position and move to another random position in the system. Indeed, in [15], we have shown that an adversary can very quickly subvert overlays based on distributed hash tables by simply never triggering leave operations. We have also demonstrated that when all nodes (honest and malicious ones) are imposed on a limited lifetime, the system eventually reaches a stationary regime where the ratio of polluted clusters is bounded, independently from the initial amount of corruption in the system.

In unstructured peer-to-peer systems, nodes cannot rely on the topological nature of structured graphs to detect undesirable behaviors. The sampling has to be uniform and ergodic. Informally, this second property guarantees that each received node id infinitely often has a non-null probability to locally appear as a sample. In [21], we determine necessary and sufficient conditions under which uniform and ergodic sampling is achievable in unstructured peer-to-peer systems potentially populated with a large proportion of Byzantine nodes. Strict restrictions are imposed on the number of messages gossiped by malicious nodes during a given period of time and providing each honest node with a very large memory (in the size of the system).

In [38], we consider the problem of targeted attacks in large scale peer-to-peer overlays. These attacks aimed at exhausting key resources of targeted hosts to diminish their capacity to provide or receive services. To defend the system against such attacks, we rely on clustering and implement induced churn to preserve randomness of nodes identifiers so that adversarial predictions are impossible. We propose robust join, leave, merge and split operations to discourage brute force denial of services and pollution attacks.

Sequence of Consensus Instances:

To be able to coordinate efficiently the activities of replicas, a significant body of work on replication techniques, group communication services and agreement problems has been done. The Consensus service has been recognized as a fundamental building block for fault-tolerant distributed systems. Many different protocols to implement such a service have been proposed, however, little effort has been placed in evaluating their performance. During her PhD thesis [14], Izabela Moise has presented a protocol designed to solve several consecutive consensus instances in an asynchronous distributed system prone to crash failures and message omissions. The protocol [31] follows the Paxos approach [49], [47] and integrates two different optimizations to reduce the latency of learning a decision value. As one optimization is risky [48], dynamics triggering criterion are defined to check at runtime if the context seems to be favorable or not. The proposed protocol [34] is adaptive as it tries to obtain the best performance gain depending on the current context. Moreover, it guarantees the persistence of all decision values. Our experimentation results [
[36] focus on the impact of the prediction of collisions (i.e., the cases where the use of the risky optimization is counterproductive).

**Transactional Mobile Agent:**
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. In [33], we propose a solution where 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.
FOCUS Project-Team

6. New Results

6.1. Contracts and sessions

Participants: Ugo Dal Lago, Ornela Dardha, Maurizio Gabbrielli, Elena Giachino, Claudio Guidi, Jacopo Mauro, Fabrizio Montesi.

Contracts are descriptions of the functionalities offered by a component or a service, and of the way these functionalities may be accessed by clients. A contract may include a description of the component capabilities, place constraints on their usage, as well as declare preferences, entitlements and credentials. When a client wants to use one of the functionalities offered, it engages a dialogue (e.g., a sequence of interactions) with the servers; this is usually called a session.

The expected dialogue in a session can be specified by means of types, the session types. We have studied the integration of union and session types in a class-based language for building network applications, which amalgamates sessions and methods in order to flexibly exchange data according to communication protocols. We have established type safety properties guaranteeing that, after a session has started, computation cannot get stuck on a communication deadlock, and studied type inference. On a similar topic is the paper.

We have used types to guarantee bounds on sessions. These are polynomial bounds, on both time and space needed by the interacting processes to carry out the interactions in their sessions. This is the first example of a refinement of session types guaranteeing quantitative properties beyond the usual safety property, and builds on earlier work on soft linear logic.

In service-oriented architectures, the mechanism that allows to manage sessions and, in particular, to assign incoming messages to the correct sessions, is critical for efficiency and performance. A well-known solution to this problem, first introduced by BPEL, makes use of correlation sets. Intuitively these distinguish different sessions by means of the values for some specific variables which are present also in messages, thus allowing for their routing to sessions on the basis of these values. We have studied a typed language for programming services based on correlation sets, that takes into account key aspects of service-oriented systems, such as distribution, loose coupling, open-endedness and integration. We have provided an implementation of the language as an extension of the Jolie language and applied it to a nontrivial real-world example of a fully-functional distributed user authentication system.

In current SOC languages based on correlation sets, a message can be assigned to a unique session. In another line of work on correlation sets, we have studied the possibility – useful in many practical examples – of broadcasting messages to more than one session. We have investigated a data structure, based on radix trees, and an algorithm for managing a correlation mechanism, that support the broadcast primitive without degrading performances.

6.2. Fault Handling, compensations and transactions

Participants: Mila Dalla Preda, Maurizio Gabbrielli, Ivan Lanese, Jacopo Mauro, Gianluigi Zavattaro.

One of the predominant properties of CBUS is the loose coupling among the components. In fact, components can dynamically connect/disconnect and can be modified/updated at run time. It is thus important to support unexpected events, called faults.

In we have studied the problem of fault handling in the kind of object-oriented languages developed in the EU Hats project; notably these languages have asynchronous method calls whose results are returned inside futures. We present an extension for those languages where futures are used to return fault notifications and to coordinate error recovery between the caller and callee. This can be exploited to ensure that invariants involving many objects are restored after faults.
Traditional fault handling mechanisms, including those based on try-catch operators, do not seem sufficient to deal with the non-local errors and failures of distributed systems. At the application level, more advanced transactional models and primitives are needed to guarantee integrity and continuity of the whole system. We study approaches based on long running transactions and compensations. A long running transaction is a computation that either successfully terminates, or it aborts. In case of abort, a compensation is executed to take the system to a consistent state. In [53], extending work started last year, we make a thorough comparison among different approaches to the specification of compensations, in particular static forms of recovery where the compensation is statically defined together with the transaction, and dynamic forms where the compensation is progressively built along with a computation.

We have also continued our study on faults and compensations in Service Oriented Computing. The approach to the interplay between bi-directional request-response interaction and faults, proposed in our past works on the Jolie language, supported the idea that the bi-directional pattern should not be interrupted in case of faults. However, this may cause long delays or even deadlocks if the communicating partner disappears. On the contrary, the approach of WS-BPEL causes no delay, but it does not allow to compensate the remote activity. We have investigated [38] an intermediate approach in which it is not necessary for the fault handler to wait for the response, but it is still possible on response arrival to gracefully close the conversation with the remote service.

A related work, but mainly developed in 2010, is [21].

6.3. Service orchestration and choreography

Participants: Mila Dalla Preda, Maurizio Gabbielli, Claudio Guidi, Ivan Lanese, Jacopo Mauro, Fabrizio Montesi, Marco Pistore, Gianluigi Zavattaro.

Orchestration has to do with the definition of services that should obey given behaviours. The services may be realised by composing services already available. Orchestration is often discussed in relationship with “choreography”, which refers to global descriptions of the intended behaviour of a system of components, stating the role of each participant and the set of coordination requirements.

In [55] we have studied the basic linguistic constructs and a reference implementation for aggregation, a mechanism for composing services that abstracts from the order of their communications. Aggregation is widely used in practice. However, since it is not natively supported by service-oriented languages, it is mostly implemented by means of ad-hoc solutions which typically exploit middleware technology.

A critical aspect for pervasive computing is the possibility to discover and use process knowledge at run time depending on the specific context. This can be achieved by using an underlying service-based application and exploiting its features in terms of dynamic service discovery, selection, and composition. Pervasive process fragments represent a service-based tool that allows to model incomplete and contextual knowledge. In [40] we provide a solution to automatically compose such fragments into complete processes, according to a specific context and specific goals. We compute the solution by encoding process knowledge, domain knowledge and goals into an AI planning problem.

Concerning choreography languages, two main approaches have been followed in their design: the interaction-oriented approach at the basis of WS-CDL and the process-oriented approach of BPEL4Chor. In [52] we investigate the relationships between the two approaches. In particular, we point out several possible interpretations for interaction-oriented choreographies: one synchronous and various asynchronous. Under each of these possible interpretations we characterize the class of interaction-oriented choreographies which have a direct process-oriented counterpart, and we formalize the corresponding notion of equivalence between the initial interaction-oriented choreography and the corresponding process-oriented counterpart.

In [50] we study the issue of checking a multiparty choreography against formal protocol specifications, and then projecting it onto a description of the individual service orchestrators. Contributions are also the definition of a multiparty choreography model, and the correctness proof for the projection.
6.4. Primitives for adaptable and evolvable components

Participants: Mario Bravetti, Ivan Lanese, Michael Lienhardt, Jacopo Mauro, Marco Pistore, Davide Sangiorgi, Gianluigi Zavattaro.

In Focus we study linguistic primitives for components, and models for them following the process calculus approach. A special emphasis is given to the adaptability and evolvability of the components — important issue in complex software systems. Components indeed are often used in contexts that had not been predicted at the time when the components were built. Moreover, the needs and the requirements on a system may change over time: one may find that the original specification was incomplete or ambiguous, or new needs may arise that had not been predicted at design time. As designing and deploying a system is costly, it is important that the system be capable of evolving and adapting itself to changes in the surrounding environment.

Models and linguistic constructs for adaptability and evolvability of components are studied in [34] and [19]. The key features of the component model in [34] are: a hierarchical structure of components; the capacity to move, update, wrap components; method interfaces for components; and capacities to isolate and distribute components. In the model in [19], adaptable processes have a location and can be subject to dynamic update actions at runtime (related to this paper is also [20]).

In [22] we provide an adaptation approach that can automatically adapt business processes to run-time context changes that impede achievement of a business goal. We define a formal framework that adopts planning techniques to automatically derive necessary adaptation activities on demand. The adaptation consists in identifying recovery activities that guarantee that the execution of a business process can be successfully resumed and, as a consequence, the business goals are achieved. The solution proposed is evaluated on a real-world scenario from the logistics domain.

Adaptability and evolvability are major concerns in Software Product Lines. The EU Hats project has developed the idea of delta-oriented programming (DOP) as a technique for implementing Software Product Lines based on modifications (add, remove, modify) to a core object-oriented program. Such modifications can introduce errors into a program, when type signatures of classes are modified. To overcome this problem we have introduced [54] a type system for delta-oriented programs. The system is based on row polymorphism, a well-known method in type systems for records.

6.5. Resource Control

Participants: Ugo Dal Lago, Marco Gaboardi, Daniel Hirschkoff, Simone Martini, Paolo Parisen Toldin, Giulio Pellitta, Davide Sangiorgi.

In Focus, we study both foundations and methodologies for controlling the amount of resources programs and processes make use of. The employed techniques mainly come from the world of type theory and proof theory, and as such have been used extensively in the context of sequential computation. Interesting results have been obtained recently indicating that those techniques can be quite useful in the concurrent context too, thus being potentially interesting for CBUS.

We have continued our work on techniques for ensuring termination of programs. On the one hand we have refined [25] previous techniques, enhancing them by taking into account input/output capabilities of channels and subtyping. On the other hand we have studied [28] how to transport techniques initially devised for processes onto sequential higher-order languages with imperative features (e.g., $\lambda$-calculi with references). The method employed makes it possible to combine term rewriting measure-based techniques for termination of imperative languages with traditional approaches to termination in purely functional languages, such as logical relations.

In [31], a type system of linear and dependent types, called dIPCF, has been proved to be a sound, but also relatively complete, way to prove intensional (but also extensional) properties of PCF programs. In other words, not only all properties of programs proved by typing are operationally valid, but all true properties of programs can be proved so by way of dIPCF. This holds not only for terms of base type, but also for (first-order) functions: this makes the type system more expressive than intersection type disciplines.
A characterization of probabilistic polynomial time classes by way of linear typing systems for a variant of
Gödel’s T called RSLR, has been proposed [51]. Classes like BPP and PP are characterized by RSLR once
appropriate constraints on the probability of error are imposed.

A unifying methodology for the study of resource consumption of processes has been presented in [13]: it is
a refinement of realizability, in which not only termination but also concrete resource bounds can be obtained
by showing a function to be realized by a program.

On the same topic is the paper [26], that polishes previous work.

6.6. Verification of extensional properties

Participants: Elena Giachino, Cosimo Laneve, Tudor Alexandru Lascu, Davide Sangiorgi, Gianluigi Zavattaro.

Extensional refers to properties that have to do with behavioral 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. We mainly employ techniques based on behavioral equivalences (and preorders), and on types
and logics. Type systems offer a good trade-off between expressiveness and efficiency of the techniques.

A substantial amount of the work carried out this year has to do with the transfer of techniques between the
areas of concurrency theory and object-oriented languages.

We have developed [29] a technique for the deadlock analysis of systems of concurrent object groups. The
technique makes use of types in the form of contracts, that is, abstract descriptions of method’s behaviours.
Object groups are collections of objects that perform collective work. Within a group, there can be only one
running thread at a time; the scheduling of threads is cooperative.

We have studied [37] the concept of ownership types, originally introduced for (sequential) object-oriented
languages, in the setting of pure message-passing concurrency. Ownership types have the effect of statically
preventing certain communications, and can block the accidental or malicious leakage of secrets. Intuitively,
a channel defines a boundary and forbids access to its inside from outer channels, thus preserving the secrecy
of the inner names from malicious outsiders.

In a different line of work, we have analyzed ad hoc networks, intended as networks of devices connected
by wireless links and communicating via broadcast. We have considered [27], [18] models in which the
communication topology of a network is represented as a graph. Nodes represent states of individual processes,
and adjacent nodes represent single-hop neighbors. Processes are finite state automata that communicate via
selective broadcast messages. Reception of a broadcast is restricted to single-hop neighbors. In these systems
we have studied various forms of reachability (example: the existence of an initial topology in which the
execution of the protocol can lead to a configuration with at least one node in a certain state).

Induction is a pervasive tool in Computer Science and Mathematics for defining structures and reasoning on
them. Coinduction is the dual of induction, and as such it brings in tools that are quite different from those
provided by induction. The best known instance of coinduction is bisimulation, mainly employed to define
and prove equalities among potentially infinite objects: processes, streams, non-well-founded sets, and so on.
Sangiorgi has completed [47], [49] two comprehensive textbooks on bisimulation and coinduction (in [49],
Sangiorgi is an editor, and author of two chapter contributions [48], [46]). The books explain the fundamental
concepts and techniques, and the duality with induction. A special emphasis is put on bisimulation as a
behavioural equivalence for processes. Thus the books also serve as an introduction to models for expressing
processes, and to the associated techniques of operational and algebraic analysis.

6.7. Tutorial papers on Service-Oriented Computing

Participants: Mario Bravetti, Ivan Lanese, Davide Sangiorgi, Gianluigi Zavattaro.

We have contributed to a few tutorial papers that summarise the work on languages and tools for Service-
Oriented Computing that has been carried out within the EU project Sensoria in Focus and elsewhere. The
papers appear as chapters of a book dedicated to the topic.
The chapters [45] and [43] present and contrast the primitives and the behavioural theories of the main process calculi designed for modeling services.

Languages and models for service-oriented applications usually include primitives and constructs for exception and compensation handling. Exception handling is used to react to unexpected events while compensation handling is used to undo previously completed activities. In [44] we investigate the impact of exception and compensation handling in service-oriented process calculi.

The chapter [42] deals with contracts: descriptions intended to provide support for the automatic on-demand discovery of functionalities offered by a service. The approach followed is to describe such contracts as process calculi expressions. We show how, in certain cases, service contracts can be automatically extracted out of service behaviour, and how they can be used to formally check compliance among the communication protocols followed by the interacting services.

Finally, in [41] we present different tools that have been developed for verifying properties of service implementations with respect to their formal specifications in an automated, or semi-automated, way.

### 6.8. Expressiveness of computational models

**Participants:** Mila Dalla Preda, Ugo Dal Lago, Ivan Lanese, Cosimo Laneve, Davide Sangiorgi.

Expressiveness refers to the study of the expressive power of computational models. In 2011 we have addressed four main aspects.

First, we have continued our investigation of reversible computations. Reversibility is a main ingredient in the study of programming abstractions for reliable systems, e.g. for exception handling. In fact, reversibility can be used for going back to some consistent state after an exception has occurred. In previous years we had defined $ρπ$, a higher-order calculus where processes can both go forward and backward in the computation. This year we have studied fine-grained rollback primitives to control reversibility. The definition of a proper semantics for such a primitive is a surprisingly delicate matter because of the potential interferences between concurrent rollbacks. We have also considered lower-level distributed semantics, which are closer to an actual implementation of the rollback primitives, and their relationship with the high-level semantics.

A thread of research close to that of $ρπ$ is the study of the properties and the expressive power of a simple calculus with reversible transitions, called reversible structures. In [24], we have demonstrated a standardization theorem for these structures. When terms in reversible structures have unique id, the standardization theorem may be strengthened in a form that bears a quadratic algorithm for reachability, a problem that is EXPSPACE-complete for generic structures (as in Petri Nets). The expressive power of reversible structures has been studied in [17], [12] and a compilation of asynchronous Reversible CCS has been provided.

A second aspect has been motivated by the analogy between malicious software and biological infections [39]. In the paper, we have used a formalism originally developed for the analysis of biological systems — the kappa calculus by Danos and Laneve — for the formalization and analysis of malicious software. In particular we have modeled the different actors involved in a malicious code attack in the kappa-calculus. Then, by simulating the behavior, we have shown how to extract relevant information that can drive the choice of the defense technique to apply.

A third aspect has been the refinement [14] of some previous work on the expressiveness and decidability of higher-order concurrent languages. — formalisms for concurrency in which processes can be passed around in communications.

A fourth aspect has been the study of properties of a simple calculus for quantum computation. In [16], we have demonstrated a confluence property both for finite and infinite computations using a novel technique.
6. New Results

6.1. Security

Participants: Ilaria Castellani, Zhengqin Luo, Tamara Rezk [correspondant], José Santos, Manuel Serrano.

6.1.1. Secure session calculi

We have pursued our work on controlling information flow in session calculi, started in previous years in collaboration with colleagues from the university of Torino. We also started investigating a notion of (objective) reputation for principals participating in sessions. The reputation of a principal is based on her previous behaviour as a user of a service. A principal’s reputation can be checked both by the service itself, before admitting again the principal as a user, or by other principals to evaluate the reputation of the current users before they join a service (we consider multi-user services). We plan to apply this idea to refine our previous work on information flow control in multiparty sessions, by considering reputations built on the “security behaviour” of principals.

In the work “Information flow safety in multiparty sessions” [11], we consider a calculus for multiparty sessions enriched with security levels for messages. We propose a monitored semantics for this calculus, which blocks the execution of processes as soon as they attempt to leak information. We illustrate the use of our monitored semantics with various examples, and show that the induced safety property implies the security property studied previously for the same calculus. This work was presented at the 18th International Workshop on Expressiveness in Concurrency (EXPRESS’11).

In the work “A Reputation System for Multirole Sessions” [10], we extend role-based multiparty sessions with reputations and policies associated with principals. The reputation associated with a principal in a service is built by collecting her relevant behaviour as a participant in sessions of the service. The service checks the reputation of principals before allowing them to take part in a session, and decides whether to accept them or not depending on their reputation and on the role they want to play. Furthermore, principals can declare policies that must be fulfilled by the other participants of the same service. These policies are used by principals to check the reputation of the current participants and to decide accordingly whether or not to join the service. Our approach is illustrated by an example describing a real-world protocol. This work was presented at the 6th International Symposium on Trustworthy Global Computing (TGC’11).

Both [11] and [10] were partially funded by the ANR-08-EMER-010 grant PARTOUT.

6.1.2. Automatic Code Injection Prevention for Web Applications

We propose a new technique based on multtier compilation for preventing code injection in web applications. It consists in adding an extra stage to the client code generator which compares the dynamically generated code with the specification obtained from the syntax of the source program. No intervention from the programmer is needed. No plugin or modification of the web browser is required. The soundness and validity of the approach are proved formally by showing that the client compiler can be fully abstract. The practical interest of the approach is proved by showing the actual implementation in the Hop environment.

This work was presented in TOSCA’11 and appeared in the LNCS series [13]. See also software section.

6.1.3. A Certified Lightweight Non-Interference Java Bytecode Verifier

We propose a type system to verify the non-interference property in the Java Virtual Machine. We verify the system in the Coq theorem prover.

This work will appear in the journal of Mathematical Structures in Computer Science [8].
6.1.4. Information-flow types for homomorphic encryptions

We develop a flexible information-flow type system for a range of encryption primitives, precisely reflecting their diverse functional and security features. Our rules enable encryption, blinding, homomorphic computation, and decryption, with selective key re-use for different types of payloads. We show that, under standard cryptographic assumptions, any well-typed probabilistic program using encryptions is secure (that is, computationally non-interferent) against active adversaries, both for confidentiality and integrity. We illustrate our approach using ElGamal and Paillier encryption.

We present two applications of cryptographic verification by typing: (1) private search on data streams; and (2) the bootstrapping part of Gentry’s fully homomorphic encryption.

We provide a prototype typechecker for our system.

This work appeared in CCS’11 [12]. See also software section.

6.1.5. The Mashic compiler

Mashups are a prevailing kind of web applications integrating external gadget APIs often written in the Javascript programming language. Writing secure mashups is a challenging task due to the heterogeneity of existing gadget APIs, the privileges granted to gadgets during mashup executions, and Javascript’s highly dynamic environment.

We propose a new compiler, called Mashic, for the automatic generation of secure Javascript-based mashups from existing mashup code. The Mashic compiler can effortlessly be applied to existing mashups based on a wide-range of gadget APIs. It offers security and correctness guarantees. Security is achieved by using the Same Origin Policy. Correctness is ensured in the presence of benign gadgets, that satisfy confidentiality and integrity constrains with regard to the integrator code. The compiler has been successfully applied to real world mashups based on Google maps, Bing maps, YouTube, and Zwibbler APIs.

See also software section.

6.1.6. Secure Information Flow by Self-Composition

Information flow policies are confidentiality policies that control information leakage through program execution. A common means to enforce secure information flow is through information flow type systems. Although type systems are compositional and usually enjoy decidable type checking or inference, their extensibility is very poor: type systems need to be redefined and proven sound for each new single variation of security policy and programming language for which secure information flow verification is desired. In contrast, program logics offer a general mechanism to enforce a variety of safety policies, and for this reason are favored in Proof Carrying Code, a promising security architecture for mobile code. However, the encoding of information flow policies in program logics is not straightforward, because they refer to a relation between two program executions. The purpose of this work is to investigate logical formulations of secure information flow based on the idea of self-composition, that reduces the problem of secure information flow of a program P to a safety property for program P composed with itself.

This work appeared in the special issue of MSCS of PLID [7].

6.1.7. Secure Information flow enforcement techniques for dynamic security policies

We performed a comprehensive investigation of alternative static mechanisms to enforce information flow policies considering a setting in which programs run under an authority that is only known at runtime and that yields a relaxation of the base security policy. The devised method aims at eliminating the need to reanalyse a program each time its authority changes. The soundness of the proposed approach was established for a concurrent higher-order imperative lambda calculus with reference creation. This work resulted in the report “Typing Illegal Information Flows as Program Effects” available at http://www-sop.inria.fr/members/Jose.Santos/reportInfFlow.pdf.
6.2. Models, semantics, and languages

Participants: Pejman Attar, Gérard Berry [correspondant], Gérard Boudol, Frédéric Boussinot, Johan Grande, Cyprien Nicolas, Manuel Serrano.

6.2.1. HipHop

HipHop is a new Domain Specific Language for Hop dedicated to request and event orchestration. HipHop follows the synchronous reactive model of the Esterel and ReactiveC languages, originally developed for embedded systems programming. It is based on synchronous concurrency and preemption primitives, which are known to be key components for the modular design of complex temporal behaviors. Although the language is concurrent, the generated code is purely sequential and thread-free. HipHop is translated to Hop code to be interpreted by the Hop runtime, either on server or client sides. HipHop has been described in a paper [9] accepted at the new International Workshop on Programming Language And Systems Technologies for Internet Clients (Plastic 2011).

6.2.2. Tesard

Tesard is a programming language of the Caml family, designed to offer simple constructs for shared memory concurrency and a deadlock-free semantics.

It features in particular the 2 following constructs:

- `thread e` which launches a thread that will execute expression `e`, and returns immediately;
- `lock x in e` which takes a lock on mutable value `x` (possibly waiting to be able to do so), executes expression `e`, releases the lock on `x` and returns the result of the execution of `e`.

A type and effect system is used at compile-time to:

- associate a mutex with every mutable value
- make the `lock ... in ...` construct implement deadlock avoidance.

The language is implemented as an interpreter and a bytecode compiler. It is a fork of Llama Light, which is itself derived from Caml. Llama Light is functionnaly roughly equivalent to Caml Light, but a large part of its code comes from OCaml, along with the threads library and runtime that we ported ourselves.

Several key parts of the language have been implemented: the runtime and most of the type and effect system, including inference of the creation of mutexes (`let region r in ...`) but not region polymorphism yet. The project is not in a usable state (no release has been made yet). Development versions are publicly available via GitHub at https://github.com/nahoj/llama.

6.2.3. Synchronous orchestration and beyond

We studied DSL, an orchestration language based on the synchronous/reactive model. In DSL, systems are composed of several sites executed asynchronously. Within each site, scripts are run in a synchronous parallel way. Scripts may call functions that are treated in an abstract way: their effect on the memory is not considered, but only their “orchestration”, i.e., the organisation of their calls in time and in place (the site where they are called). The mapping of sites onto cores allows one to benefit from multicore architectures. Two properties are required by DSL: reactivity of sites and absence of interferences between scripts run by distinct sites. We also introduced DSLM, which adds a memory level to DSL and a way to automatically adapt the execution to get a maximal use of the available cores. This work, presented respectively in [18] and [15], was funded by the ANR-08-EMER-010 grant PARTOUT.

6.3. Web programming

Participants: Zhengqin Luo, Cyprien Nicolas, Tamara Rezk, Bernard Serpette, Manuel Serrano [correspondant].
6.3.1. A new evaluator for Hop

At the time where Hop programs were basic scripts, the performance of the server-side interpreter was not a concern. An inefficient interpreter was acceptable. As Hop expanded, Hop programs got larger and more complex. A more efficient interpreter was necessary. Therefore, this year we have designed and implemented a new interpreter for Hop. It is compact, its whole implementation counting no more than 2.5 KLOC. It is more than twice faster than the old interpreter and consumes less than a third of its memory. The architecture and the performance of the new interpreter are described in [14].

6.3.2. Abstraction of Hop with a bicolored lambda-calculus

We have studied an extension of the $\lambda$-calculus where each expression has a color. These colors designate the sites where expressions are evaluated, i.e., in the server or in the client. Colors are similar to the $\$ and " annotations of Hop. With this, we have defined a transformation, using $\beta$-expansion, which groups together expressions with the same color. Correction, confluence and termination of the transformation was verified using the Coq system and its description was described in paper to appear in 2012 [16].

Following Hop’s syntax, $\$ mentions that the followed expression is evaluated on the server while the " character introduce client code. Inside the client code it is allowed to reintroduce some server expressions by reusing a $\$. We can imagine, for example, that the action associated to a client’s button is dependent on some server’s data (order number, proxies, data bases, ...). This kind of example is depicted with the abstract syntax tree in the upper left part of figure 1.

![Figure 1. Example of Hop colored tree](image)

From the server’s point of view, the client’s code following a " is ignored, but the environment in which this client code is activated must be preserved for all $\$ inside this code. Therefore, there exists a relation between a node " and its including$. These relations are depicted by the dashed arrows in the upper right tree of Figure 1. The proposed transformation highlights this relation and is shown in the bottom of the figure where it can be observed that the transformation groups together expressions of the same color.

6.3.3. HopTeX, an Hop application for authoring documents

HOPTEX is a HOP application for authoring HTML and LaTeX documents. The content of the document is either expressed in HTML or in a blending of HTML and a dedicated wiki syntax, for the sake of conciseness and readability. The rendering of the document is expressed by a set of CSS rules. The main originality of HOPTEX is to consider LaTeX as a new media type for HTML and to express the compilation from HTML to LaTeX by the means of dedicated style sheet rules. HOPTEX can then be used to generate high quality documents for both paper printed version and electronic version.
HOPTEX is implemented in HOP, a multi-tier programming language for the Web 2.0. This implementation extensively relies on two facilities generally only available on the client-side that HOP also supports on the server-side of the application: DOM manipulations and CSS server-side resolutions.

The online version of the paper describing HOPTEX [17] is available at the HOPTEX web page (http://hop.inria.fr/hop/weblets/homepage?weblet=hoptex). Contrary to the PDF version published in the proceeding of the workshop, the online version is convenient for both desktop computers and Smartphones.
6. New Results

6.1. Introduction

Research results are presented according to the research directions of the MYRIADS team.

6.2. Autonomous Management of Virtualized Infrastructures


6.2.1. Cloud Federations


6.2.1.1. Virtual Execution Platforms in Cloud Federations

In the context of the Contrail European project, we have defined the overall architecture of the Contrail software stack for cloud computing on top of cloud federations [51]. We have focused on the design and the implementation of a first basic prototype of the Virtual Execution Platform (VEP) component [52]. It is in charge of provisioning hardware resources from Cloud providers and to deploy and run distributed applications submitted by users under the control of a negotiated Service Level Agreement (SLA) [16]. Within VEP software, REST interface, OVF parsing, SSL security, Authorization modules are under active development and at various levels of integration. A first demo version of VEP running on top of OpenNebula IaaS cloud has been successfully demonstrated at the first annual project review.

6.2.1.2. Efficient virtual cluster migration

We continued our work on Shrinker, a system providing efficient live migration of virtual clusters on wide area networks. The design has been improved to coordinate the deduplication on the source site of the migration. Deduplication is now performed only within an individual virtual cluster, in order to reduce security issues and avoid performance impact on virtual machines of other users. We performed a comprehensive performance evaluation of Shrinker. An article presenting the design, implementation, and performance of Shrinker was published in [28].

6.2.1.3. Elastic Map/Reduce over Cloud Federations

We worked on the development of Resilin, a system to easily create execution platforms over distributed cloud resources for executing MapReduce computations. Resilin implements the Amazon Elastic MapReduce web service API and uses resources from private and community clouds. Resilin takes care of provisioning, configuring and managing cloud-based Hadoop execution platforms, potentially using multiple clouds. An initial implementation of Resilin was presented at the CCA ’11 workshop [36]. Further development was performed in the context of Ancuta Iordache’s master internship. The results of this work were published as a research report [45].

6.2.1.4. Sky Computing Experiments

We continued our collaboration with the University of Florida on sky computing experiments, which led to the publication of a book chapter [38].

6.2.2. Infrastructure as a Service Clouds

Participants: Stefania Costache, Eugen Feller, Yvon Jégou, Christine Morin, Nikos Parlavantzas, Pierre Riteau.
6.2.2.1. Large scale Energy aware self-healing IaaS

The research done in 2011 was two fold. A prototype of the previously proposed scalable, fault-tolerant and energy-aware virtual machine (VM) management framework called Snooze was implemented and evaluated on the Grid5000 testbed [41]. In 2011, we have focused on the implementation of the self-healing mechanisms and protocols, and on integrating in Snooze the system-level mechanisms (e.g. for automatically switching on/off cluster nodes) to support energy aware resource management algorithms. Our experimental results show that the fault-tolerance features of the framework do not impact application performance. Moreover, negligible cost is involved in performing distributed VM management and the system remains highly scalable with increasing amounts of resources. A nature-inspired VM placement algorithm [24] based on the Ant Colony Optimization (ACO) meta-heuristic was developed and evaluated by means of simulations.

6.2.2.2. Resource Management in Private Clouds

We focused on the design of a resource management system for private clouds that provides support for different application SLAs while maximizing the resource utilization of the infrastructure. As we also considered the need of providing users the incentives to truthfully express their valuation for the performance of their application we investigated existing economic models of allocating resources. As a result, we proposed a novel resource management architecture based on a virtual economy. In this system, independent agents monitor the application’s performance and dynamically provision virtual machines from the infrastructure under user’s budget constraints. To provision virtual machines, a proportional share auction is used, allowing a fine-grain resource sharing at a low complexity cost. This work was done as part of a collaboration with EDF R&D and was published at the VHPC 2011 workshop [22]. We have also implemented a first prototype of this proposal. In collaboration with Vydia Rajagopalan (Master student at VU Amsterdam) we have implemented the proportional-share auction scheduler and integrated it with the OpenNebula Virtual Infrastructure Manager. Then, we have extended this work with the design of agents that execute scientific applications (MPI and Bag-of-Task applications) under deadline and budget constraints. Experimental evaluations are currently performed on Grid5000.

6.2.2.3. Resilience

We initiated a collaboration with Box Leangsuksun’s group on high availability of cloud infrastructures. We carried out a preliminary study on how the HA-OSCAR environment developed at the Louisiana Tech University could be used to ensure the high availability of critical services in Nimbus IaaS clouds [30].

6.2.3. XtreamOS Grid Distributed Operating System

Participants: Amine Belhaj, Jérôme Gallard, Rémy Garrigue, Yvon Jégou, Christine Morin, Yann Radenac.

6.2.3.1. Facilitating Experiments with XtreamOS Grid System

XtreamOS Grid system that has been developed in the framework of the XtreamOS European project is now evolving as an open source software in a community driven by INRIA in the framework of the XtreamOS Easy ADT. We have provided first level support to users and maintained XtreamOS website, wiki and mailing-lists. We have updated XtreamOS documentation to reflect the evolution of XtreamOS system. We facilitated the access to XtreamOS in three different ways: maintaining an open public testbed running XtreamOS, providing ready-to-use XtreamOS virtual machines and developing tools to automatically deploy XtreamOS Grid system on the Grid’5000 large-scale experimentation platform. In 2011, we have finalized a new 3.0 version of XtreamOS system and ported it on top of the OpenSuse 11.4 Linux distribution. We performed a number of tests to validate the installation, configuration and execution of the new XtreamOS version based on openSuSE Linux distribution. An incremental integration process has been set up to facilitate the integration of patches and bug fixes. We have run a number of experiments with XtreamOS 3.0 based on Mandriva Linux distribution: MPI programs, Salomé numerical simulation platform, bio-informatics applications. Yann Radenac, in the framework of the COOP project funded by ANR contributed to XtreamOS’s code by fixing bugs, cleaning the source code to improve maintainability, and adding a few minor features.
6.2.3.2. Resource Management for Dynamic Applications

In the framework of the COOP project funded by ANR, we compared the features offered by the CooRM resource manager for dynamic applications developed by Christian Perez and Cristian Klein at ENS Lyon with those provided by the XtreamOS Grid system. A plan has been set to adapt CooRM to XtreamOS system and to extend XtreamOS’s API to include a CooRM-like interface [53]. We worked on the definition of a variant of CooRM that can work in the context of XtreamOS Grid operating system.

6.3. Dynamic Adaptation of Service-based Applications

Participants: Françoise André, Djawida Dib, Erwan Daubert, Guillaume Gauvrit, André Lage, Christine Morin, Nikos Parlavantzas, Jean-Louis Pazat, Chen Wang, Mohamed Zouari.

6.3.1. Dynamic Adaptation in a Distributed Operating System

Participants: Françoise André, Djawida Dib, Christine Morin, Nikos Parlavantzas.

We have studied the feasibility to dynamic adapt the features of a distributed operating system using a framework for self-adaptation of service oriented distributed applications [46]. We have focused on the consistency protocols for replicated data in distributed shared memory systems. We have considered two strict consistency protocols, one based on invalidation and one based on broadcast on write operations. The adaptation framework selects one of these two algorithms based on the inter-node data transfer delay. We have implemented a prototype based on Kerrighed single system image operating system for clusters and the SAFDIS adaptation framework. We have integrated a broadcast based consistency protocol in Kerrighed that already implements a write invalidation consistency protocol. We have implemented the adaptation policy in the SAFDIS framework and the needed adaptation mechanisms in Kerrighed as well as a component for monitoring data transmission delays. An experimental evaluation is being carried out.

6.3.2. Adaptation for Data Management

Participants: Françoise André, Mohamed Zouari.

The usage of context-aware data management in mobile environments has been previously investigated by Françoise André in collaboration with Mayté Segarra and Jean-Marie Gilliot from Telecom Bretagne Brest (previously known as ENST Bretagne). This work focuses on data management in grid and mobile environments; an ambient assisted living application illustrates the approach. This work was realized in the context of the ALORAD project (Architecture LOgicielle pour la Réplication Adaptative de Données), financed by the Brittany council. Mayté Segarra from Telecom Bretagne Brest was co-adviser for the PhD thesis of M. Zouari [12].

6.3.3. Adaptation for Service-Oriented Architectures

Participants: Françoise André, Erwan Daubert, Guillaume Gauvrit, André Lage, Nikos Parlavantzas, Jean-Louis Pazat, Chen Wang.

Service-Oriented Computing is a paradigm that is rapidly spreading in all application domains and all environments - grids, clusters of computers, mobile and pervasive platforms. The following works take place in the context of the S-CUBE European Network of Excellence.

6.3.3.1. Services adaptation in distributed and heterogeneous systems

We are still studying service adaptation in distributed and heterogeneous systems. This work covers different aspects such as structural, behavioral and environmental adaptation, distributed decision and planification of adaptation actions, adaptive allocation of resources for services. A framework called SAFDIS for “Self Adaptation For Distributed Services” has been defined and implemented. It is built as a set of services, providing functionalities useful to build an adaptation system. The analysis phase can take reactive as well as proactive decisions. This gives the ability to either react fast or to take decisions for the long term. This implies the ability to analyze the context with a variable depth of reasoning. Our implementation of the SAFDIS analysis phase also distributes and decentralizes its analysis process to spread the computational
load and make the analysis process scalable. The planning phase seeks the set of actions (the plan) needed to adapt the system according to the strategy chosen by the analysis phase. It also schedules the selected actions to ensure a coherent and efficient execution of the adaptation. The planning topic is a well-known subject in AI research works and many algorithms already exist in that field to produce efficient schedules. With our SAFDIS framework, the planning phase is able to reuse these algorithms. The resulting plan of actions can have actions that can be executed in parallel.

6.3.3.2. Quality Assurance for Distributed Services

In the context of the service-centric paradigm, we have designed and developed the Qu4DS (Quality Assurance for Distributed Services) research prototype. Qu4DS is a cloud PaaS solution which fills the gap between the conception of higher-level SaaS service providers over the resource-level PaaS layer. Qu4DS provides an automatic support for service execution management by aiming at increasing the service provider’s profit. More specific, Qu4DS dynamically acquires resources according to the customer demand, deploys service instances and implements QoS assurance mechanisms in order to prevent SLA violations. Moreover, Qu4DS has been evaluated on Grid5000 and showed to be effective on reducing service provider’s costs [27].

6.3.3.3. Self-configuration for Cloud Platforms

By definition, cloud computing offers an abstraction to manage various needs and concepts such as distributed software design, the deployment of such software on dynamic resources and the management of this kind of resources. Thus it is possible to reconfigure (adapt) according to some needs the software as well as the use of the resources. However these reconfigurations that are used on different layers may also have impacts on the others. Moreover these layers are independent and so are able to adapt themselves independently of the others. In our work, we propose to use some adaptation capabilities offered for example by the infrastructure (IaaS) that manages the resources to adapt the software (SaaS). We also propose to use planning algorithms to coordinate the adaptations between them to avoid inconsistency or inefficiency due to concurrent adaptations.

6.3.3.4. Dynamic Adaptation of Chemical services

We have proposed a QoS-aware middleware for dynamic service execution. In the context of dynamic execution, a workflow is defined by composing a set of abstract activities as place holders. Each activity is bound to a suitable partner service, which is selected at run-time from a set of functional equivalent candidates with different non-functional properties such as quality of service (QoS). The service selection process is modeled as a series of chemical reactions.

6.3.3.5. Prediction of SLA violations and dynamic adaptation in workflows

During execution, run-time QoS is determined by the dynamic execution environment and thus the expected QoS is not always ensured. In addition, infrastructure failures can make a service undeliverable. The adaptive execution reflects the capability to recompose a (part of a) workflow on-the-fly in case that global SLA violation is predicted. Most techniques for predicting global SLA violation require past experiences on executions of a business process. All historical execution instances have the same structure as well as the same bindings. These solutions do not adapt to the case of dynamic execution, where for each execution, partner services are selected and bound at run time.

In order to predict global SLA violation in the context of dynamic service execution, we proposed a 2-phase prediction technique, which is fit for generic workflow composition. The prediction method works with a high accuracy for simple workflows, but when the workflow composes complicated structures (such as loops and exclusive branches), the performance degrades. The reason is that the estimation of global SLA is based on the critical path, which is not definitely executed. To solve this problem, we propose to use data mining technique to predict workflow branches and the number of loop execution. Based on predicted branches, the prediction of global SLA violation is much more accurate. The numerical evaluation will be carried out in the near future.

6.4. A Chemical Approach for Autonomous Service Computing

Participants: Héctor Fernandez, Marko Obrovac, Thierry Priol, Cédric Tedeschi.
6.4.1. Chemistry-Inspired Workflow Management System for e-Science Applications

Participants: Héctor Fernandez, Cédric Tedeschi, Thierry Priol.

In the research track that aims at leveraging the properties of the chemical Programming models for autonomic computing, we have built a software based on the HOCL compiler (part of the HOCL-tools) that was actually deployed and experimented over the Grid’5000 platform. The experiments have shown, that envisioning the execution workflow as an autonomic chemical process is actually viable in practice. Experimented with different well known workflow-based e-Science applications, the software showed a performance level comparable to current top-rated scientific workflow management systems [25].

6.4.2. Solving Workflow Patterns Through Molecular Composition

Participants: Héctor Fernandez, Cédric Tedeschi, Thierry Priol.

In the same area, but on a more conceptual point of view, we have shown how the expressive power of the chemical model can be leveraged to solve complex workflow patterns. This aspect was also integrated into the HOCL-tools and experimented over the Grid’5000 platform, following two architectures with a different level of decentralization, showing the advantages and drawbacks of decentralizing the workflow execution using a chemical workflow engine [26].

6.4.3. Scalable Atomic Capture of Molecules

Participants: Marko Obrovac, Cédric Tedeschi.

Capturing the reactants involved in a reaction constitutes one of the main challenges in the execution of chemical programs. Doing it at large scale is one of the essential barriers hindering the actual execution of chemical programs at large scale. We proposed a protocol solving this issue on top of a distributed hash table (DHT). The DHT secures the scalability of the communications and provides a scalable discovery of reactants. Our protocol is triggered once reactants are found. It is made of two sub-protocols being used at different stages of the computation, according to the density of possible reactions. The protocol is validated through its proof of liveness and simulations showing that it is able to switch from one sub-protocol to the other efficiently, according to the execution’s conditions [18].
6. New Results

6.1. Distributed Programming Models

6.1.1. Multi-active Objects
Participants: L. Henrio, I. Zsolt, F. Huet.

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 extends 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 contribution includes

- publication of the basic principles of the new model [25]
- refinement of the proposal and adding dynamic compatibility
- operational semantic for multi-active objects
- extensive experiments

6.2. Component-oriented Distributed and Large-Scale Programming

6.2.1. Behavioural models for Distributed Components
Participants: E. Madelaine, R. Halalai, A. Savu, M. Alexe, L. Henrio.

In the past, we defined the behavioural semantics of active objects and components, in [3]; we extended last year this work to take group communications. On the practical side, this work contributes to the Vercors platform; the overall picture being to provide tools to the programmer for defining his application, including its behavioural specification. Then some generic properties like absence of deadlocks, but also application specific properties can be validated on the composed model using an existing model-checker. We mainly use CADP model-checker, that also supports distributed generation of state-space. This year our main achievements are the following:

- We provided model for one-to-many component communication
- We provided a model for Byzantine failures, specified a component application supporting some Byzantine faults, and proved its correctness;
- We conducted heavy experiments on distributed model-checking in this context;
- We worked on the formal specification of the behavioural model generation for component systems.

Most of those results were published in [22] and [35].

6.2.2. Enacting large-scale service orchestration using a component-based approach
Participants: F. Baude, V. Legrand.
The distribution of business processes encompasses the inclusion of external service providers in the overall process as well as the usage of external infrastructures like clouds. Both of these approaches lead to decentralization and outsourcing of a part of the global workflow, resulting in a complexified management of the global orchestration. As a matter of fact, the overall data are decentralized among different domains and must, most of the time, be gathered manually. To this extent, we continue our work on agile and distributed orchestration, showing that the framework we develop eases multidomain orchestration management. Our approach extracts, gathers and digests data from the decentralized processes in order to provide an unified and global view of a distributed orchestration. This year we focussed in particular on:

- The specification of the execution framework extending the SCA specification by adding temporal dependencies
- The definition of a use-case allowing the provisioning of a large set of OSGi gateways.

This work resulted in the PhD thesis of Virginie Legrand [12].

6.2.3. Autonomic Monitoring and Management of Components

Participants: F. Baude, C. Ruz, B. Sauvan, R. Dib.

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, and we have tested it in simple applications. Our implementation 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 [16].

- We presented the general description of the framework and its capability to support autonomic behaviour in [30]. This work takes advantage of the componentized membrane of GCM/ProActive, and of the PAGCMScript reconfiguration extensions made in our team.
- We used our proposition to design an integrated framework to cover the life-cycle of a service application from business and design level, to deployment and execution concerns in a Cloud environment, in a work done in conjunction with Adrian Mos and Alain Boulze formerly leading the INRIA ADT Galaxy from INRIA Rhône-Alpes. This work was presented in the SoEA4EE workshop [29].
- We experimented with the use of our autonomic framework to integrate autonomic behaviour into skeletons. This work was taken during the engineering internship of Rima Dib, and included the collaboration with Marco Danelutto from Università di Pisa.

6.3. Middleware for Grid and Cloud computing

6.3.1. RDF Data Storage and Retrieval In P2P Systems


We have proposed in the context of the SOA4ALL FP7-IP project (8.3.1.1) the design and the implementation of a hierarchical Semantic Space infrastructure based on Structured Overlay Networks (SONS) [46], [10]. It aims at the storage and the retrieval of the semantic description of services at the Web scale [47]. This infrastructure combines the strengths of both P2P paradigm at the architectural level and the Resource Description Framework (RDF) data model at the knowledge representation level. As it is designed, the proposed infrastructure enables the processing of simple and complex queries. This year, the following achievements have been realised.

- A thorough survey of the existing works that have adapted the combination of RDF data model and the P2P communication model to build distributed infrastructures for RDF data storage and retrieval has been performed. This effort was published in a journal [34]. A previous but more complete version of this work can be found in a research report [45] and was used extensively in [39], [38], [36].
• We provided and presented in [23] an implementation of a three dimensional CAN overlay network for storing and retrieving RDF triples. At the implementation level, a modular and flexible architecture for the Semantic Space infrastructure has been proposed. The implementation relies on the ProActive Grid middleware and provides a clear separation between its sub-components (overlay, storage, query engine, etc.). The modularity of the architecture is combined with the decentralized aspect of the infrastructure enabling the RDF data storage and retrieval at large scale. The evaluation of the infrastructure through micro-benchmarks experiments on clusters and grids shows the impact of the architecture and data distribution on the performance of the storage and processing mechanisms.

In the context of the FP7 Strep PLAY (8.3.1.2) and French ANR SocEDA (8.2.2) research projects, we have extended the aforementioned work with a content-based Publish/Subscribe abstraction in order to support asynchronous queries for RDF-based events in large scale settings. In order to support these queries efficiently, we worked on an efficient broadcast primitive on top of CAN which we formalized and implemented (see section 6.3.2). We are also working towards a generalization of this broadcast algorithm to a multicast one, and contribute intensively to the general integration effort for offering such innovative semantic described event marketplace platform at cloud scale [41].

6.3.2. An algorithm for efficient broadcast over CAN-like P2P networks

Participants: L. Henrio, F. Bongiovani, A. Craciun.

The nature of some large-scale applications such as content delivery systems or publish/subscribe systems, built on top of structured overlay networks, demands application-level dissemination primitives which do not overwhelm the overlay and which are also reliable. Building such communication primitives in a reliable manner on top of such networks would increase the confidence regarding their behavior prior to deploying them in real settings. In order to come up with real efficient primitives, we take advantage of the underlying geometric topology of the overlay network and we also model the way peers communicate with one another. For this our objective is to design and prove an efficient and reliable broadcast algorithm for CAN-like P2P networks. To this aim, this year we:

• Formalised in Isabelle/HOL a CAN-like P2P system, devised formalised tools to reason on CAN topologies, and on communication protocols on top of CANs. We proved first completeness and correctness properties on some classes of broadcast algorithms.
• Designed an efficient broadcast algorithm on top of CAN and implemented it.

Preliminary results were presented at CFSE and published as a research report [37]; another publication is under way.

6.3.3. Matlab/Scilab parallel programming

Participant: F. Viale.

Matlab & Scilab, with millions of users around the world, are industry standards for numerical computing. They both lack a powerful and modern parallel computing framework to meet the industry’s growing demand in terms of parallel processing. This activity is intended to integrate into both softwares a toolbox for parallel processing, based on ProActive.

• This year, we implemented a ProActive Scilab toolbox with the same functionalities as the ProActive Matlab toolbox we built last year.
• We added in the Matlab toolbox a disconnected mode to allow closing the Matlab session while uncomplete Matlab jobs are still running on the scheduler side, and retrieving the job results at the next connection.
• We reorganized both Matlab and Scilab toolboxes with a cleaner and more intuitive API, a stronger and stabler implementation and an extensive documentation. We designed as well unit-tests to make the toolboxes fully usable for production standards.
• The Scilab toolbox is now deployed on PACAGrid cluster and used extensively by other partners of the OMD2 ANR.
6.3.4. Network Aware Cloud Computing

**Participants:** S. Malik, F. Huet.

We have proposed a cloud scheduler module named Network Awareness Module (NAM), which helps the scheduler to take the more efficient scheduling decisions on the basis of resource features, such as network latency, reliability, environment compatibility and monetary cost issues.

- Currently we are working on Reliability Assessment based Scheduling on Cloud Infrastructure. We are building a model, which enables a scheduler to schedule the tasks on cloud infrastructure, on the basis of adaptive reliability of nodes (virtual machine). The core of this model is a reliability assessment algorithm, which computes the reliability for individual resources and for the group of resources.

- We have proposed, designed and implemented an algorithm for the grouping of nodes on the basis of inter-node latencies. This algorithm can do the dynamic grouping and work with the incomplete latency information available. It groups the nodes on the basis of node latency instead of neighbor count. It produces mutually exclusive groups and can perform group reconfiguration.

- We have designed a model of the Virtual Cloud [27]. The concept of Virtual cloud revolves around the concept, “Rent Out the Rented Resources”. In this model, cloud vendors offer low cost cloud services by acquiring underutilized resources from some big third-party enterprise. The cloud vendor then further rents out those resources/services to the cloud users. The upfront and administrative costs for the Virtual cloud vendor are lower, and the cloud users access services at a cheaper rate.

- We have proposed a fault tolerance model for real time cloud computing [27]. In this model, the system tolerates the faults and makes the decision on the basis of reliability of the virtual machines. The reliability of the virtual machines is adaptive, which changes after every computing cycle. A metric model is given for the reliability assessment. The system provides both the forward and backward recovery mechanisms.
6. New Results

6.1. Leveraging Software Architectures to Guide and Verify the Development of Sense/Compute/Control Applications

A software architecture describes the structure of a computing system by specifying software components and their interactions. Mapping a software architecture to an implementation is a well known challenge. A key element of this mapping is the architecture’s description of the data and control-flow interactions between components. The characterization of these interactions can be rather abstract or very concrete, providing more or less implementation guidance, programming support, and static verification.

In this work, we have introduced a notion of behavioral contract that expresses the set of allowed interactions between components, describing both data and control-flow constraints [15]. This declaration is part of the architecture description, allows generation of extensive programming support, and enables various verifications. We have instantiated our approach in an architecture description language for the domain of Sense/Compute/Control (SCC) applications, and described associated compilation and verification strategies.

The main contributions of this work are the following:

- We have introduced a language for behavioral contracts dedicated to SCC applications.
- We have shown that behavioral contracts can effectively guide the implementation of SCC applications by enabling the generation of highly customized programming frameworks using a dedicated compiler. This approach ensures the conformance between the architecture and the implementation, while facilitating software evolution.
- We have shown that such descriptions are precise enough to verify safety properties such as information flow reachability or behavioral invariants.
- Based on an implementation of behavioral contracts in a design language targeting SCC applications, we have assessed the benefit of behavioral contracts at a conceptual level and in terms of metrics on the resulting code.

6.2. A Step-wise Approach for Integrating QoS throughout Software Development

Non-functional requirements are used to express the quality to be expected from a system. For real-time systems such as avionics, it is critical to guarantee this quality, in particular time-related performance properties. In this domain, deterministic QoS is generally ensured at the execution platform level (e.g., operating systems, distributed systems technologies, hardware specificities), independently of a particular application. When addressing the QoS requirements of a given application, these platform-specific guarantees are not sufficient.

In this work, we have proposed a step-wise QoS approach integrated through all development phases and development artifacts [17]. This approach is dedicated to control-loop systems. Control-loop systems are systems that sense the external environment, compute data, and eventually control the environment accordingly. This kind of systems can be found in a range of domains, including avionics, robotics, and pervasive computing. For example, in the avionics domain, a flight management application is a control-loop system that (1) senses the environment for location and other navigation information, (2) computes the trajectory and (3) modifies the wings configuration accordingly.
The main contributions of this work are the following:

- We have developed a step-wise approach that systematically processes QoS requirements throughout software development. This integrated approach is dedicated to control-loop systems, allowing to rely on a particular architectural pattern and thus enhancing the design and programming support level for non-functional aspects. For now, we focus on time-related performance but the approach could be generalized to other non-functional properties (e.g., CPU or memory consumption).

- Our approach has been integrated into DIASUITE, a tool-based development methodology dedicated to control-loop systems. DIASUITE is based on a dedicated design language that we have enriched with time-related performance properties. This non-functional extension has been used to offer verification and programming support at each development stage.

- Our approach has been applied to the development of avionics applications such as a flight management system and a collision avoidance system. These experiments have demonstrated that our step-wise approach can effectively guide the avionics certification process.

6.3. Architecturing Conflict Handling of Pervasive Computing Resources

The rapid development of new devices (further resources) and development tools being opened to third-parties have paved the way to an increasing number of applications being deployed in pervasive computing environments. These applications anarchically access resources. In this situation, it is very common for a resource to be accessed by multiple applications, potentially leading to conflicts. For example, in a building management system, a security application that grants access inside the building, can conflict with an application dealing with emergency situations like fires, preventing the building to be evacuated.

Managing conflicts consists of three main parts, detection, resolution and prevention. These parts crosscut the development cycle of applications and pervasive computing systems. In this work, we have proposed a conflict management process that cleanly separates conflict management tasks by providing a design method and supporting tools [18]. This facilitates the work of developers, architects and administrators, who can follow clear guidelines to manage conflicts.

The main contributions of this work are the following:

- We have identified requirements at different stages during the development cycle that are necessary to detect, resolve, and prevent conflicts. We have assigned duties and responsibilities to existing roles, that are carried out during the conflict management process without interfering with the standard application development.

- We have extended a domain-specific design language to declare conflict resolution at an architectural level. During the conflict management process conditions are specified and prioritized. Afterwards conflicting applications (inter application) or modules (intra application) are linked to these conditions.

- The declared information is used to generate code dedicated to conflict handling. On the one hand, a compiler generates a dedicated framework that guides the implementation of the conflict handling logic at application and system level. On the other hand, it generates code that orchestrates resource accesses and prevents conflicts.
5. New Results

5.1. RFID and Internet of Things

Participants: Roudy Dagher, Nathalie Mitton, Roberto Quilez, Loic Schmidt, David Simplot-Ryl, Lei Zhang.

5.1.1. Reader anti-collision protocol

In a Radio-Frequency IDentification network, while several readers are placed close together to improve coverage and consequently read rate, reader-reader collision problems happen frequently and inevitably. High probability of collision not only impairs the benefit of multi-reader deployment, but also results in misreadings in moving RFID tags. In order to eliminate or reduce reader collisions, we propose in [28] an Adaptive Color based Reader Anti-collision Scheduling algorithm (ACoRAS) for 13.56 MHz RFID technology where every reader is assigned a set of colors that allows it to read tags during a specific time slot within a time frame. Only the reader holding a color (token) can read at a time. Due to application constraints, the number of available colors should be limited, a perfect coloring scheme is not always feasible. ACoRAS tries to assign colors in such a way that overlapping areas at a given time are reduced. To the best of our knowledge ACoRAS is the first reader anti-collision algorithm which considers, within its design, both application and hardware requirements in reading tags. We show, through extensive simulations, that ACoRAS outperforms several anticollision methods and detects more than 99% of mobile tags while fitting application requirements.

5.1.2. Distributed ALE

Following the Internet of Things concept [14], each object is associated with a unique identifier which will allow to retrieve information about it in large databases. In the process of managing a large amount of objects, and consequently a large amount of events from readers, without overloading the network, these events have to be filtered and aggregated. This is the aim of the Application Level Events (ALE) standard from EPCGlobal, which receives events from readers and sends a useful and well constructed report to the business application. The ALE may be connected to several hundreds of readers. As the number of readers may increase with the increase of the company, a bottleneck may appear with all readers events sent to the ALE. A solution for scalability is to distribute the ALE. In [37], we propose an efficient way to solve this problem based on a Distributed Hash table (DHT). One role of the ALE is to insulate business application from technical concern so in our solution, we present a mechanism to distribute the ALE using Chord, a well-known peer-to-peer lookup system, and being transparent for business applications. This solution is compliant with the EPCglobal existing standard, scalable, robust and transparent for other layers of the middleware. We show that our solution generates only 10% overhead than in a nominal case while offering a better robustness and scalability when numbers of tags and readers increase significantly.

5.1.3. Advance Internet of Things

The Internet of Things (IoT) is a network of Internet-enabled objects, whose original purpose would be to interconnect all things in our daily life to build an always connected world. However, most of studies in the current IoT scientific community only focus on the radio-frequency identification (RFID) and wireless sensor network (WSN) based objects and lose the generality features endowed by the original definition of IoT. Furthermore, the emergence and proliferation of smart objects have been significantly changing our daily lives. It has been becoming evident that the objects should far beyond only "be identified and interconnected", but can also be controlled in an intelligent and transparent way independent of third party object (user) profiles and space and time span. In [39], we proposes a standardization scheme for a new paradigm: Advanced Internet of Things (AIoT), which is based on our proposed Unified Object Description Language (UODL) and allows to identify and interconnect every object and event with a standard format, and makes it easier and flexible for the third party control and management by integrating multiple services issued from cloud computing.
The purpose of our proposed AIoT scheme is to build a smart world of always on, always-awareness, always-connected, always-controllable, and establish an “intelligent networking” based relationship among the objects, service suppliers and the third party users. In the scope of AIoT, all the objects are transparent across the networks and can be identified and controlled (with security guarantees) via a standard prototype anytime and anywhere.

5.2. Topology control and neighbor discovery

Participants: Xu Li, Nathalie Mitton, Jovan Radak, David Simplot-Ryl, Isabelle Simplot-Ryl.

5.2.1. Topology control

Topology control is a tool for self-organizing wireless networks locally. It allows a node to consider only a subset of links/neighbors in order to later reduce computing and memory complexity. Topology control in wireless sensor networks is an important issue for scalability and energy efficiency. It is often based on graph reduction performed through the use of Gabriel Graph or Relative Neighborhood Graph. This graph reduction is usually based on geometric values.

In [35] we tackle the problem of possible connectivity loss in the reduced graph by applying a battery level based reduction graph. Experiments are conducted to evaluate our proposition. Results are compared with RNG [52] reduction which takes into account only the strength of the received signal (RSSI). Results show that our algorithm maintains network connectivity longer than solutions from the literature and balances the energy consumption over nodes.

In [31], we propose a radically new family of geometric graphs, i.e., Hypocomb, Reduced Hypocomb and Local Hypocomb for topology control. The first two are extracted from a complete graph; the last is extracted from a Unit Disk Graph (UDG). We analytically study their properties including connectivity, planarity and degree bound. All these graphs are connected (provided the original graph is connected) planar. Hypocomb has unbounded degree while Reduced Hypocomb and Local Hypocomb have maximum degree 6 and 8, respectively. To our knowledge, Local Hypocomb is the first strictly-localized, degree-bounded planar graph computed using merely 1-hop neighbor position information. We present a construction algorithm for these graphs and analyze its time complexity. Hypocomb family graphs are promising for wireless ad hoc networking. We report our numerical results on their average degree and their impact on FACE [49] routing. We discuss their potential applications and some open problems.

5.2.2. Neighbor discovery

To perform topology control, a node needs to discover its neighbors. Hello protocol is the basic technique for neighborhood discovery in wireless ad hoc networks. It requires nodes to claim their existence/aliveness by periodic "hello" messages. Central to a hello protocol is the determination of hello message transmission rate. No fixed optimal rate exists in the presence of node mobility. The rate should in fact adapt to it, high for high mobility and low for low mobility. In [30], we propose a novel mobility prediction based hello protocol, named ARH (Autoregressive Hello protocol). Each node predicts its own position by an ever-updated autoregression-based mobility model, and neighboring nodes predict its position by the same model. The node transmits "hello" message (for location update) only when the predicted location is too different from the true location (causing topology distortion), triggering mobility model correction on both itself and each of its neighbors. ARH evolves along with network dynamics, and seamlessly tunes itself to the optimal configuration on the fly using local knowledge only. Through simulation, we demonstrate the effectiveness and efficiency of ARH, in comparison with the only competitive protocol TAP (Turnover based Adaptive hello Protocol). With a small model order, ARH achieves the same high neighborhood discovery performance as TAP, with dramatically reduced message overhead (about 50% lower hello rate).
5.2.3. Address allocation

In [9], we propose a localized address autoconfiguration (LaConf) scheme for wireless ad hoc networks. Address allocation information is maintained on the network border nodes, called addressing agents (AAs), which are locally identified by a geographic routing protocol GFG (Greedy-FACE-Greedy). When a node joins the network, it acquires an address from a neighboring AA (if any exists) by local communication or from the head AA (a geographic extreme AA) by GFG-based multi-hop communication. A Geographic Hash Table (GHT) is adopted for duplicate address detection. Each address is hashed to a unique location in the network field, and the associated assignment information is stored along the face perimeter enclosing that location (in the planar graph). When a node receives an address assignment, it consults with the perimeter nodes around the hash location of the assigned address about any conflicts. AAs detects network partitions and mergers locally according to neighborhood change and triggers AA re-selection and network re-configuration (if necessary). We propose to apply a Connected Dominating Set (CDS) to improve the performance. We also evaluate LaConf through simulation using different planar graphs.

5.3. Routing

Participants: Nicolas Gouvy, Xu Li, Nathalie Mitton, David Simplot-Ryl.

In mobile wireless sensor networks, flows sent from data collecting sensors to a sink could traverse inefficient resource expensive paths. Such paths may have several negative effects such as devices battery depletion that may cause the network to be disconnected and packets to experience arbitrary delays. This is particularly problematic in event-based sensor networks (deployed in disaster recovery missions) where flows are of great importance. In [27], we use node mobility to improve energy consumption of computed paths. Mobility is a two-sword edge, however. Moving a node may render the network disconnected and useless. We propose CoMNet (Connectivity preservation Mobile routing protocol for actuator and sensor NETworks), a localized mechanism that modifies the network topology to support resource efficient transmissions. To the best of our knowledge, CoMNet is the first georouting algorithm which considers controlled mobility to improve routing energy consumption while ensuring network connectivity. CoMNet is based on (i) a cost to progress metric which optimizes both sending and moving costs, (ii) the use of a connected dominating set to maintain network connectivity. CoMNet is general enough to be applied to various networks (actuator, sensor). Our simulations show that CoMNet guarantees network connectivity and is effective in achieving high delivery rates and substantial energy savings compared to traditional approaches. CoMNET has then been extended in [26] to multi-hop movement.

In [12] we propose a novel localized Integrated Location Service and Routing (ILSR) scheme, based on the geographic routing protocol GFG, for data communications from sensors to a mobile sink in wireless sensor networks. The objective is to enable each sensor to maintain a slow-varying routing next hop to the sink rather than the precise knowledge of quick-varying sink position. In ILSR, sink updates location to neighboring sensors after or before a link breaks and whenever a link creation is observed. Location update relies on flooding, restricted within necessary area, where sensors experience (next hop) change in GFG routing to the sink. Dedicated location update message is additionally routed to selected nodes for prevention of routing failure. Considering both unpredictable and predictable (controllable) sink mobility, we present two versions. We prove that both of them guarantee delivery in a connected network modeled as unit disk graph. ILSR is the first localized protocol that has this property. We further propose to reduce message cost, without jeopardizing this property, by dynamically controlling the level of location update. A few add-on techniques are as well suggested to enhance the algorithm performance. We compare ILSR with an existing competing algorithm through simulation. It is observed that ILSR generates routes close to shortest paths at dramatically lower (90% lower) message cost.

In [29], we propose a novel trust management scheme for improving routing reliability in wireless ad hoc networks. It is grounded on two classic autoregression models, namely Autoregressive (AR) model and Autoregressive with exogenous inputs (ARX) model. According to this scheme, a node periodically measures the packet forwarding ratio of its every neighbor as the trust observation about that neighbor. These measurements constitute a time series of data. The node has such a time series for each neighbor. By applying
an autoregression model to these time series, it predicts the neighbors future packet forwarding ratios as their trust estimates, which in turn facilitate it to make intelligent routing decisions. With an AR model being applied, the node only uses its own observations for prediction; with an ARX model, it will also take into account recommendations from other neighbors. We evaluate the performance of the scheme when AR, ARX or a previously proposed Bayesian model is used. Simulation results indicate that the ARX model is the best choice in terms of accuracy.

5.4. Self-deployment, localization and area coverage

Participants: Milan Erdelj, Xu Li, Enrico Natalizio, Nathalie Mitton, Tahiry Razafindralambo, David Simplot-Ryl, Isabelle Simplot-Ryl.

5.4.1. Deployment

First steps in order to perform any task, a network needs to be deployed and nodes need to discover each other. To the best of our knowledge, very few scenarios when robots self-deploy to afterwards themselves constitute the network nodes or drop off sensor nodes have been investigated so far and none of them ensure the network connectivity at every step. In [15], we consider the self-deployment of wireless sensor networks. We present a mechanism which allows to preserve network connectivity during the deployment of mobile wireless sensors. Our algorithm is localized and is based on a subset of neighbors for motion decision. Our algorithm maintains a connected topology regardless of the direction chosen by each sensor. To preserve connectivity, the distance covered by the mobile nodes is constrained by the connectivity of the node to its neighbors in a connected subgraph like the relative neighborhood graph. We show the connectivity preservation property of our algorithm through analysis and present some simulation results on different deployment schemes such as full coverage, point of interest coverage or barrier coverage.

Another approach is the one proposed in [34] in which node placement is performed off-line with objective to optimize a criterion. Based on the specific application, different objectives can be taken into account such as energy consumption, throughputs, delay, coverage, etc. Also many schemes have been proposed in order to optimize a specific quality of service (QoS) parameter. Power consumption is an essential issue in wireless multimedia sensor networks (WMSNs) due to the elevated processing capabilities requested by the video acquisition hardware installed on the generic sensor node. Hence, node placement scheme in WMSNs greatly impacts the overall network lifetime. [34] first proposes a suitable hardware architecture to implement a feasible WMS node based on off-the-shelf technology, then it evaluates the energy consumption obtained throughout a wise “energy-spaced” placement of the wireless nodes without affecting the video quality of multimedia traffic. In [4], we propose to use a neural network as a controller for nodes mobility and a genetic algorithm for the training of the neural network through reinforcement learning. This kind of scheme is extremely adaptive, since it can be easily modified in order to consider different objectives and QoS parameters. In fact, it is sufficient to consider a different kind of input for the neural network to aim for a different objective. All things considered, we propose a new method for programming a WSRN and we show practically how the technique works, when the coverage of the network is the QoS parameter to optimize. Simulation results show the flexibility and effectiveness of this approach even when the application scenario changes (e.g., by introducing physical obstacles).

5.4.2. Coverage

The coverage of Points of Interest (PoI) is a classical requirement in mobile wireless sensor applications. Optimizing the sensors self-deployment over a PoI while maintaining the connectivity between the sensors and the sink is thus a fundamental issue. [22] addresses the problem of autonomous deployment of mobile sensors that need to cover a predefined PoI with a connectivity constraints and provides the solution to it using Relative Neighborhood Graphs (RNG) [52]. Our deployment scheme minimizes the number of sensors used for connectivity thus increasing the number of monitoring sensors. Analytical results, simulation results and real implementation are provided to show the efficiency of our algorithm. To the best of our knowledge, only [21] both preserves the network connectivity and validates its proposition through experimentations with real wireless robots. This work has been extended to discovery and coverage of multi-point of interested in [
21]. Indeed, the problems of multiple PoI coverage, environment exploration and data report are still solved separately and there are no works that combine the aforementioned problems into a single deployment scheme. In [21], we present a novel approach for mobile sensor deployment, where we combine multiple PoI coverage with network connectivity preservation and environment exploration in order to capture the dynamics of the monitored area. We examine the performance of our scheme through extensive simulation campaigns.

As sensors are energy constrained devices, one challenge in wireless sensor networks (WSNs) is to guarantee coverage and meanwhile maximize network lifetime. In [7], we leverage prediction to solve this challenging problem, by exploiting temporal-spatial correlations among sensory data. The basic idea lies in that a sensor node can be turned off safely when its sensory information can be inferred through some prediction methods, like Bayesian inference. We adopt the concept of entropy in information theory to evaluate the information uncertainty about the region of interest (RoI). We formulate the problem as a minimum weight submodular set cover problem, which is known to be NP hard. To address this problem, an efficient centralized truncated greedy algorithm (TGA) is proposed. We prove the performance guarantee of TGA in terms of the ratio of aggregate weight obtained by TGA to that by the optimal algorithm. Considering the decentralization nature of WSNs, we further present a distributed version of TGA, denoted as DTGA, which can obtain the same solution as TGA. The implementation issues such as network connectivity and communication cost are extensively discussed. We perform real data experiments as well as simulations to demonstrate the advantage of DTGA over the only existing competing algorithm and the impacts of different parameters associated with data correlations on the network lifetime.

5.4.3. Localization

In mobile-beacon assisted sensor localization, beacon mobility scheduling aims to determine the best beacon trajectory so that each sensor receives sufficient beacon signals with minimum delay. We propose a novel DeteRministic bEAcon Mobility Scheduling (DREAMS) algorithm [32], [10], without requiring any prior knowledge of the sensory field. In this algorithm, beacon trajectory is defined as the track of depth-first traversal (DFT) of the network graph, thus deterministic. The mobile beacon performs DFT under the instruction of nearby sensors on the fly. It moves from sensor to sensor in an intelligent heuristic manner according to RSS (Received Signal Strength)-based distance measurements. We prove that DREAMS guarantees full localization (every sensor is localized) when the measurements are noise-free. Then we suggest to apply node elimination and topology control (Local Minimum Spanning Tree) to shorten beacon tour and reduce delay. Through simulation we show that DREAMS guarantees full localization even with noisy distance measurements. We evaluate its performance on localization delay and communication overhead in comparison with a previously proposed static path based scheduling method.

5.5. Platforms and Substitution Networks


5.5.1. Platforms

In the framework of the ANR SensLAB project, a wireless sensor testbed has been set up in Lille in order to allow the evaluation through experiments of scalable wireless sensor network protocols and applications. All functionalities offered by the platform have then been presented in [17], [16], [42]. SensLAB’s main and most important goal is to offer an accurate open access multisusers scientific tool to support the design, the development tuning, and the experimentation of real large-scale sensor network applications. The SensLAB testbed is composed of 1024 nodes over 4 sites. Each site hosts 256 sensor nodes with specific characteristics in order to offer a wide spectrum of possibilities and heterogeneity. Within a given site, each one of the 256 nodes is able both to communicate via its radio interface to its neighbors and to be configured as a sink node to exchange data with any other “sink node”. The hardware and software architectures that allow to reserve, configure, deploy firmwares and gather experimental data and monitoring information are described. We also present demonstration examples to illustrate the use of the SensLAB testbed and encourage researchers to test
and benchmark their applications/protocols on a large scale WSN testbed. A survey of platforms similar to SensLAB can be found in [6].

5.5.2. Emulation

Although some platforms like SensLAB are very convenient, they do not always fit the application requirements and setting up experimental testbed of large scale wireless sensor networks requires huge cost, space and human resources. A more affordable approach is needed to provide preliminary insights on network protocols performance. To overcome the need for significant number of sensors required to perform a realistic experiment, and/or to experiments with high density networks, we introduce in [43] a novel approach: emulation by using all available sensors as candidate forwarding neighbors of the node S currently holding the packet. Destination position is virtual. After successfully sending message to forwarding node B over realistic wireless channel, the position of virtual destination is adjusted by translating it for vector BS and possibly rotating it to change the neighborhood configuration. The same node S then again selects new forwarding neighbor. Such selection of best forwarding neighbor continues until virtual destination appears close to a real node, and the later then becomes final destination node. Compared to real testbeds, our emulation has advantages of testing networks with very large densities (which may not be possible in small scale implementations), and in unlimited scalability of our physical implementations (e.g. we can emulate network with a million nodes).

5.5.3. Substitution network

A substitution network is a rapidly deployable backup wireless solution to quickly react to network topology changes due to failures or to flash crowd effects on the base network. Unlike other ad hoc and mesh solutions, a substitution network does not attempt to provide new services to customers but rather to restore and maintain at least some of the services available before the failure. Furthermore, a substitution network is not deployed directly for customers but to help the base network to provide services to customers. Therefore, a substitution network is not, by definition, a stand-alone network. [36] describes the quality of service architecture for substitution networks and discuss provisioning, maintenance, and adaptation of QoS inside and between the base network and the substitution network. In the same context, [33] shows the impact of the router mobility on the QoS of such networks.

5.6. Data collection and management

Participants: Thierry Delot, Geoffroy Cogniaux, Arnaud Fontaine, Alia Ghaddar, Michael Hauspie, Samuel Hym, Xu Li, Nathalie Mitton, Tahiry Razafindralambo, David Simplot-Ryl, Isabelle Simplot-Ryl.

5.6.1. Data collection

Wireless sensors networks (WSNs) are deployed to collect huge amounts of data from the environment. This produced data has to be delivered through sensor’s wireless interface using multi-hop communications toward a sink. The position of the sink impacts the performance of the wireless sensor network regarding delay and energy consumption especially for relaying sensors. Optimizing the data gathering process in multi-hop wireless sensor networks is, therefore, a key issue. [19] and [18] address the problem of data collection using mobile sinks in a WSN. We provide a framework that studies the trade-off between energy consumption and delay of data collection. This framework provides solutions that allow decision makers to optimally design the data collection plan in wireless sensor networks with mobile sinks.

In [20], [5], we focus on information gathering in vehicular ad hoc networks. Until now, only a few research works have addressed this problem. They have lead to solutions relying on push models, where potentially useful data are pushed towards vehicles. To the best of our knowledge, no work has tackled the use of pull models in VANETs. Such models would allow users to send queries to a set of cars in order to find the desired information. In order to propose such a query processing scheme, the main challenge to address is to route the different results towards their recipient in a highly dynamic network where the nodes move very quickly.

To solve this issue, we propose GeoVanet, a DHT-based geographic routing protocol which ensures that the sender of a query can get a consistent answer. Our goal is not to compute the query result “instantaneously” but to ensure that the user will be able to retrieve it within a bounded time. To prove the effectiveness of GeoVanet,
an experimental evaluation is provided in the paper. It shows that up to 80% of the available query results are delivered to the user.

Another way to optimize data collection is to send data only when necessary. Knowledge discovery and data analysis in resource constrained wireless sensor networks faces different challenges. One of the main challenges is to identify misbehaviors or anomalies with high accuracy while minimizing energy consumption in the network. In [25], we extend a previous work of us and we present an algorithm for temporal anomalies detection in wireless sensor networks. Our experiments results show that our algorithm can efficiently and accurately detect anomalies in sensor measurements. It also produces low false alarm rate for slow variation time series measurements without harvesting the source of energy.

In data aggregation, sensor measurements from the whole sensory field or a sub-field are collected as a single report at an actor using aggregate functions such as sum, average, maximum, minimum, count, deviation, etc. We propose a localized Delay-bounded and Energy-efficient Data Aggregation (DEDA) protocol [11], [38] for request-driven wireless sensor networks with IEEE 802.11 CSMA/CA MAC layer. This protocol uses a novel two-stage delay model, which measures end-to-end delay using either hop count or degree sum along a routing path depending on traffic intensity. It models the network as a unit disk graph (UDG) and constructs a localized minimal spanning tree (LMST) sub-graph. Using only edges from LMST, it builds a shortest path (thus energy-efficient) tree rooted at the actor for data aggregation. The tree is used without modification if it generates acceptable delay, compared with a given delay bound. Otherwise, it is adjusted by replacing LMST sub-paths with UDG edges. The adjustment is done locally on the fly, according to the DEsired Progress (DEP) value computed at each node. We further propose to integrate DEDA with a localized sensor activity scheduling algorithm and a localized connected dominating set algorithm, yielding two DEDA variants, to improve its energy efficiency and delay reliability. Through an extensive set of simulation, we evaluate the performance of DEDA with various network parameters. Our simulation results indicate that DEDA far outperforms the only existing competing protocol.

5.6.2. Data management

The use of reliable high-level languages based on virtual machines, such as java, is now possible on systems as small as smart cards or sensors. However, the potential of these languages is widely limited by hardware constraints as memory storage capacity etc. We claim that it may be leveraged by coupling cache mechanisms with external memory storages. [40] is a preliminary study of the set up of such an approach. Thanks to simulation based results, we identify three main factors which tend to decrease the performances of cache setting code in Java.

5.6.3. Data security

[41], [24] presents the enforcement of control flow policies for Java bytecode devoted to open and constrained devices. On-device enforcement of security policies mostly relies on run-time monitoring or inline checking code, which is not appropriate for strongly constrained devices such as mobile phones and smart-cards. We present a proof-carrying code approach with on-device lightweight verification of control flow policies statically at loading time. Policies are expressed by finite automata, the technique is in-between security automata and control flow policies of Jensen et al. Our approach is suitable for evolving, open and constrained Java-based systems as it is compositional, to avoid re-verification of already verified bytecode upon loading of new bytecode, and it is regressive, to cleanly support bytecode unloading.

While mobile devices have become ubiquitous and generally multi-application capable, their operating systems provide few high level mechanisms to protect services offered by application vendors against potentially hostile applications coexisting on the device. In [23], we tackle the issue of controlling application interactions including collusion in Java-based systems running on open, constrained devices such as smart cards or mobile phones. We present a model specially designed to be embedded in constrained devices to verify at install-time that interactions between applications abide by the security policies of each involved application without resulting in run-time computation overheads; this model deals with application (un)installations and policy changes in an incremental fashion. We show the feasibility of our approach and its security.
enhancements on a multi-application use case for GlobalPlatform/Java Card smart cards. This approach is developed in EVe - TCF.

Telecommunication software systems, containing security vulnerabilities, continue to be created and released to consumers. We need to adopt improved software engineering practices to reduce the security vulnerabilities in modern systems. Contracts can provide a useful mechanism for the identification, tracking, and validation of security vulnerabilities. In [8], we propose a new contract-based security assertion monitoring framework (CB SAMF) that is intended to reduce the number of security vulnerabilities that are exploitable across multiple software layers, and to be used in an enhanced systems development life cycle (SDLC). We show how contract-based security assertion monitoring can be achieved in a live environment on Linux. Through security activities integrated into the SDLC we can identify potential security vulnerabilities in telecommunication systems, which in turn are used for the creation of contracts defining security assertions. Our contract model is then exercised, as runtime probes, against two common security related vulnerabilities in the form of a buffer overflow and a denial of service.
6. New Results

6.1. Introduction

In 2011, we focused our research on the following areas:

- distributed algorithms for large and dynamic networks,
- Complex queries over peer-to-peer networks
- Trust and reputation management on P2P networks
- dynamic adaptation of virtual machines,
- services management in large scale environments,
- Formal and practical study of optimistic replication, incorporating application semantics.
- Decentralized commitment protocols for semantic optimistic replication.

6.2. Distributed algorithms

Participants: Luciana Arantes [correspondent], Franck Petit, Maria Potop-Butucaru [correspondent], Swan Dubois, Pierre Sens, Julien Sopena.

Our current research in the context of distributed algorithms focuses on two main axes. We are interested in providing fault-tolerant and self-* (self-organizing, self-healing and self-stabilizing) solutions for fundamental problems in distributed computing. More precisely, we target the following basic blocks: mutual exclusion, resources allocation, agreement and communication primitives.

In dynamic systems we are interested in designing building blocks for distributed applications such as: failure detectors, adequate communication primitives (publish/subscribe) and overlays. Moreover, we are interested in solving fundamental problems such as synchronization, leader election, membership and naming, and diffusion of information.

6.2.1. Failure Detectors for Dynamic Systems

Since 2009, we explore a distributed computing model of dynamic networks such as (MANET or Wireless sensor networks). The temporal variations in the network topology implies that these networks can not be viewed as a static connected graph over which paths between nodes are established beforehand. Path between two nodes is in fact built over the time. Furthermore, lack of connectivity between nodes (temporal or not) makes of dynamic networks a partitionable system, i.e., a system in which nodes that do not crash or leave the system might be not capable to communicate between themselves. In 2011 we propose a new failure detector protocol which implements an eventually strong failure detectors ($\diamond S$) on a dynamic network with an unknown membership. Failure detector is a fundamental service, able to help in the development of fault-tolerant distributed systems. Our failure detector has the interesting feature to be time-free, so that it does not rely on timers to detect failures; moreover, it tolerates mobility of nodes and message losses [41].

6.2.2. Self-* Distributed Algorithms

The main challenges of our research activity over 2011 year were to develop self-* (self-stabilizing, self-organizing and self-healing) distributed algorithms for various type of networks. Self-stabilization is a general technique to design distributed systems that can tolerate arbitrary transient faults. Since topology changes can be considered as a transient failures, self-stabilization turns out to be a good approach to deal with dynamic networks. This is particularly relevant when the distributed (self-stabilizing) protocol does not require any global parameters, like the number of nodes or the diameter of the network. With such a self-stabilizing protocol, it is not required to change global parameters in the program when nodes join or leave the system. Therefore, self-stabilization is very desirable to achieve scalability and dynamicity.

- **Snap-stabilizing Committee Coordination.** The classic committee coordination problem characterizes a general type of synchronization called $n$-ary rendezvous as follows:\[3\]:
Professors in a certain university have organized themselves into committees. Each committee has an unchanging membership roster of one or more professors. From time to time a professor may decide to attend a committee meeting; it starts waiting and remains waiting until a meeting of a committee of which it is a member is started. All meetings terminate in finite time. The restrictions on convening a meeting are as follows: (1) meeting of a committee may be started only if all members of that committee are waiting, and (2) no two committees may convene simultaneously, if they have a common member. The problem is to ensure that (3) if all members of a committee are waiting, then a meeting involving some member of this committee is convened."

In [31], we propose two snap-stabilizing distributed algorithms for the committee coordination problem. Snap-stabilization is a versatile technique allowing to design algorithms that efficiently tolerate transient faults. Indeed, after a finite number of such faults (e.g. memory corruptions, message losses, etc), a snap-stabilizing algorithm immediately operates correctly, without any external intervention. The first algorithm maximizes the concurrency, whereas the latter maximizes the fairness.

- **Snap-stabilizing Message Forwarding.** We focus on end-to-end request and response delivery of messages that are carried over the network. This problem is also known as the message forwarding problem. It consists in the management of network resources in order to forward messages, i.e., protocols allowing messages to move from a sender to receiver over the network. Combined with an self-stabilizing routing protocol, achieving snap-stabilization for the message forwarding problem is a very desirable property because every message sent by the sender is delivered in finite time to the receiver. In other words, no message that was actually sent after the system started is lost. In [46], we present a snap-stabilizing algorithm for the message forwarding that works on a tree topology. It uses a constant number of buffers per link, mainly \(2\delta + 1\) buffers by node, where \(\delta\) is the degree of the node. Therefore, it is particularly well suited for large-scale and dynamic systems, e.g, overlays used in peer-to-peer systems.

- **Self-Organizing Swarms of Robots.**

Consider a distributed system where the computing units are weak mobile robots, i.e., devices equipped with sensors and are designed to move. By weak, we mean that the robots are anonymous, autonomous, disoriented, and oblivious, i.e., devoid of (1) any local parameter (such that an identity) allowing to differentiate any of them, (2) any central coordination mechanism or scheduler, (3) any common coordinate mechanism or common sense of direction, and (4) any way to remember any previous observation nor computation performed in any previous step. Furthermore, all the robots follow the same program (uniform or homogeneous), and there is no kind of explicit communication medium. The robots implicitly “communicate” by observing the position of the others robots. Two different environments are considered: (i) the continuous two-dimensional Euclidian space wherein robot can observe, compute and move with infinite decimal precision, and (ii) the discrete model in which the space is partitioned into a finite number of locations, conveniently represented by a graph where nodes represent locations that can be sensed, and where edges represent the possibility for a robot to move from one location to the other.

During 2011, we mainly investigated the following problems: the gathering onto the plane, and the exploration of a finite discrete environment.

1. **Gathering.** This problem can be stated as follows: Robots, initially located at various positions, gather at the same position in finite time and remain at this position thereafter. In [20], we investigate the self-stabilizing gathering problem in the plane, that is gathering the robots deterministically with no kind of restriction on the initial configuration. In particular, robots are allowed to share same positions in the initial configuration. Strong multiplicity detection is the ability for the robots to count the exact number of robots located at a given position. We show that assuming strong multiplicity detection, it is

---

possible to solve the self-stabilizing gathering problem with $n$ weak robots in the semi-synchronous model if, and only if, $n$ is odd. By contrast, we show that with an odd number of robots, the problem becomes solvable. Our proof is constructive, as we present and prove a deterministic self-stabilizing algorithm for the gathering problem.

In [19], we address the gathering in the discrete environment. We prove some asymptotical time and space complexity lower bounds to solve the problem. We propose an algorithm that is asymptotically optimal in both space and round complexities. Finally, we show that most of the assumptions we made are necessary to deterministically solve the rendezvous considering our initial scenario.

2. Graph Exploration.

The exploration problem is to visit a discrete and finite environment by a swarm of robots. We consider two types of explorations: the finite exploration and the perpetual exploration. The former requires that $k$ robots, initially placed at different nodes, collectively explore a graph before stopping moving forever. By “collectively” explore we mean that every node is eventually visited by at least one robot. In [73], we propose optimal (w.r.t, the number of robots) solutions for the deterministic exploration of a grid shaped network by a team of $k$ asynchronous oblivious robots in the asynchronous and non-atomic asynchronous model. In more details, we first show that no exploration protocol exists with less than three robots for any grid with more than three nodes, less than four robots for the $(2,2)$-Grid, and less than five robots for the $(3,3)$-Grid. Next, we show that the problem is solvable using only 3 robots for any $(i,j)$-Grid, provided that $j > 3$. Our result is constructive as we present a deterministic algorithm that performs in the non-atomic asynchronous model. We also present specific deterministic protocols for the $(3,3)$-Grid using five robots.

The second type of exploration is the perpetual exploration. It requires every possible location to be visited by each robot infinitely often. In [32], we investigate the exclusive perpetual exploration of grid shaped networks. We focus on the minimal number of robots that are necessary and sufficient to solve the problem in general grids. In more details, we prove that three deterministic robots are necessary and sufficient, provided that the size of the grid is $n \times m$ with $3 \leq n \leq m$ or $n = 2$ and $m \geq 4$. Perhaps surprisingly, and unlike results for the exploration with stop problem (where grids are "easier" to explore and stop than rings with respect to the number of robots), exclusive perpetual exploration requires as many robots in the ring as in the grid. Furthermore, we propose a classification of configurations such that the space of configurations to be checked is drastically reduced. This pre-processing lays the bases for the automated verification of our algorithm for general grids as it permits to avoid combinatorial explosion.

6.2.3. Combining Fault-Tolerance and Self-stabilization in Dynamic Systems

Recently, we started to investigate complex faults scenarios. Distributed fault-tolerance can mask the effect of a limited number of permanent faults, while self-stabilization provides forward recovery after an arbitrary number of transient faults hit the system. FTSS (Fault-Tolerant Self-Stabilizing) protocols combine the best of both worlds since they tolerate simultaneously transient and (permanent) crash faults. To date, deterministic FTSS solutions either consider static (i.e., fixed point) tasks, or assume synchronous scheduling of the system components. We proposed in [30] a fault-tolerant and stabilizing simulation of an atomic register. The simulation works in asynchronous message-passing systems, and allows a minority of processes to crash. The simulation stabilizes in a pragmatic manner, by reaching a long execution in which it runs correctly. A key element in the simulation is a new combinatorial construction of a bounded labeling scheme accommodating arbitrary labels, including those not generated by the scheme itself. Our simulation uses a self-stabilizing implementation of a data-link over non-FIFO channels [61]. In [23] we present the first study of deterministic FTSS solutions for dynamic tasks in asynchronous systems, considering the unison problem as a benchmark.
Unison can be seen as a local clock synchronization problem as neighbors must maintain digital clocks at most one time unit away from each other, and increment their own clock value infinitely often. We present several impossibility results for this difficult problem and propose a FTSS solution (when the problem is solvable) for the state model that exhibits optimal fault containment.

6.3. Peer-to-peer systems

Participants: Pierre Sens [correspondent], Nicolas Hidalgo, Sergey Legtchenko, Olivier Marin, Sébastien Monnet, Gilles Muller, Maria Potop-Butucaru, Mathieu Valero.

6.3.1. Peer-to-peer storage

Distributed Hash Table (DHTs) provide a means to build a completely decentralized, large-scale persistent storage service from the individual storage capacities contributed by each node of the peer-to-peer overlay. However, persistence can only be achieved if nodes are highly available, that is, if they stay most of the time connected to the overlay. Churn (i.e., nodes connecting and disconnecting from the overlay) in peer-to-peer networks is mainly due to the fact that users have total control on their computers, and thus may not see any benefit in keeping their peer-to-peer client running all the time.

When connection/disconnection frequency is too high in the system, data-blocks may be lost. This is true for most current DHT-based system’s implementations. To avoid this problem, it is necessary to build really efficient replication and maintenance mechanisms. Since 2008 we study the effect of churn on an existing DHT-based P2P system namely PAST/Pastry. We have proposed RelaxDHT [25], a churn-resilient peer-to-peer DHT. RelaxDHT proposes an enhanced replication strategy with relaxed placement constraints, avoiding useless data transfers and improving transfer parallelization. This new replication strategy is able to cut down by 2 the number of data-block losses compared to PAST DHT. We are now starting to study the use of erasure coding mechanisms along with replication within DHTs. Our goal is to propose hybrid mechanisms to find a good tradeoff among 1) churn-resilience, 2) maintenance cost, and 3) storage space.

6.3.2. Overlays

Large-scale distributed systems gather thousands of peers spread all over the world. Such systems need to offer good routing performances regardless of their size and despite high churn rates. To achieve that requirement, the system must add appropriate shortcuts to its logical graph (overlay). However, to choose efficient shortcuts, peers need to obtain information about the overlay topology. In case of heterogeneous peer distributions, retrieving such information is not straightforward. Moreover, due to churn, the topology rapidly evolves, making gathered information obsolete. State of-the-art systems either avoid the problem by enforcing peers to adopt a uniform distribution or only partially fulfill these requirements. To cope with this problem, we propose DONUT [47], a mechanism to build a local map that approximates the peer distribution, allowing the peer to accurately estimate graph distance to other peers with a local algorithm. The evaluation performed with real latency and churn traces shows that our map increases the routing process efficiency by at least 20% compared to the state-of-the-art techniques. It points out that each map is lightweight and can be efficiently propagated through the network by consuming less than 10 bps on each peer.

6.3.3. Distributed trees

Publish/Subscribe implemented on top of distributed R-trees (DR-trees) overlays offer efficient DHT-free communication primitives. We have then extend the distributed R-trees (DR-trees) in order to reduce event delivery latency in order to meet the requirements of massively distributed video games such that pertinent information is quickly distributed to all the interested parties without degrading the load of nodes neither increasing the number of noisy events. In 2011, we explore how to improve robustness of distributed trees. Since each single crash can potentially break the tree structure connectivity, DR-trees are crash-sensitive. We this have proposed a fault tolerant approach which exploits replication of non leaf nodes in order to ensure the tree connectivity in presence of crashes. This work will be published in [56].
In [85], we consider a complete binary tree and construct a random pairing between leaf nodes and internal nodes. We prove that the graph obtained by contracting all pairs (leaf-internal nodes) achieves a constant node expansion with high probability. In the context of P2P overlays our result can be interpreted as follows: if each physical node participating to the tree overlay manages a random pair that couples one virtual internal node and one virtual leaf node then the physical-node layer exhibits a constant expansion with high probability which improves the robustness of the overlay.

6.3.4. Complex query over peer-to-peer networks

A major limitation of DHTs is that they only support exact-match queries. In order to offer range queries over a DHT it is necessary to build additional indexing structures. Prefix-based indexes, such as Prefix Hash Tree (PHT), are interesting approaches for building distributed indexes on top of DHTs. Nevertheless, the lookup operation of these indexes usually generates a high amount of unnecessary traffic overhead which degrades system performance by increasing response time. In [42], we propose a novel distributed cache system called Tabu Prefix Table Cache (TPT-C), aiming at improving the performance of the Prefix-trees. We have implemented our solution over PHT, and the results confirm that our searching approach reduces up to a 70% the search latency and traffic overhead. In [44], we propose DRing an efficient layered solution that directly supports range queries over a ring-like DHT structure. We improve load balancing by using only the nodes that store data, and by updating neighbour information through an optimistic approach.

6.3.5. Trust management in peer-to-peer networks

An ongoing research work focuses on trust assessment in dynamic systems. Even if it is near impossible to fully trust a node in a P2P system, managing a set of the most trusted nodes in the system can help to implement more trusted and reliable services. Using these nodes can reduce the probability of introducing malicious nodes in distributed computations. Our work aims at the following objectives: 1. To design a distributed membership algorithm for structured Peer to Peer networks in order to build a group of trusted nodes. 2. To design a maintenance algorithm to periodically clean the trusted group so as to avoid nodes whose reputation has decreased under the minimum value. 3. To provide a way for a given node X to find at least one trusted node. 4. To design a prototype of an information system, such as a news dissemination system, that relies on the trusted group. In 2011, we propose the CORPS system for building a community of reputable peers in Distributed Hash Tables [26].

6.4. Virtual machine (VM)

Participants: Harris Bakiras, Bertil Folliot, Gaël Thomas [correspondent], Gilles Muller [correspondent], Julia Lawall, Arie Middlekoop, Thomas Preud’homme, Suman Saha.

Our research interests are in improving the way systems software is developed. One theme of our research is the development of virtual machines with a specific focus on resource management, isolation and concurrency management. Another theme of our research is related to bug finding in systems software.

6.4.1. Virtual machines

Isolation in OSGi: The OSGi framework is a Java-based, centralized, component oriented platform. It is being widely adopted as an execution environment for the development of extensible applications. However, current Java Virtual Machines are unable to isolate components from each other’s. By modifying shared variables or allocating too much memory, a malicious component can freeze the complete platform. We work on I-JVM, a Java Virtual Machines that provides a lightweight approach to isolation while preserving the compatibility with legacy OSGi applications. Our evaluation of I-JVM shows that it solves the 15 known OSGi vulnerabilities due to the Java Virtual Machine with an overhead below 20%. I-JVM has been presented in DSN 2009.
VMKit: Managed Runtime Environments (MREs), such as the JVM and the CLI, form an attractive environment for program execution, by providing portability and safety, via the use of a bytecode language and automatic memory management, as well as good performance, via just-in-time (JIT) compilation. Nevertheless, developing such a fully featured MRE, including features such as a garbage collector and JIT compiler, is a herculean task. As a result, new languages cannot easily take advantage of the benefits of MREs, and it is difficult to experiment with extensions of existing MRE based languages. VMKit is a first attempt to build a common substrate that eases the development of high-level MREs. We have successfully used VMKit to build two MREs: a Java Virtual Machine (J3) and a Common Language Runtime (N3). VMKit has performance comparable to the well established open source MREs Cacao, Apache Harmony and Mono. VMKit is freely distributed under the LLVM licence with the LLVM framework developed by the University of Illinois at Urbana-Champaign and now maintained by Apple.

A third MRE is being built in cooperation with the "Algorithms, Programmes and Resolution" team in LIP6. This integrates a functional machine (the Zinc Abstract Machine) in VMKit and shows that the adaption at the language level of our virtual machine. This project has been funded by the LIP6 in 2009-10 and 2010-11.

6.4.2. Semantic patches

Open source infrastructure software, such as the Linux operating system, Web browsers and n-tier servers, has become a well-recognized solution for implementing critical functions of modern life. Furthermore, companies and local governments are finding that the use of open source software reduces costs and allows them to pool their resources to build and maintain infrastructure software in critical niche areas. Nevertheless, the increasing reliance on open source infrastructure software introduces new demands in terms of security and safety. In principle, infrastructure software contains security features that protect against data loss, data corruption, and inadvertent transmission of data to third parties. In practice, however, these security features are compromised by a simple fact: software contains bugs.

We are developing a comprehensive solution to the problem of finding bugs in API usage in open source infrastructure software based on our experience in using the Coccinelle code matching and transformation tool, and our interactions with the Linux community.

Coccinelle has been successfully used for finding and fixing bugs in systems code. One of our main recent results is an extensive study of bugs in Linux 2.6 [51] that has permitted us to demonstrate that the quality of code has been improving over the last six years, even though the code size has more than doubled.

We have used Coccinelle to generate traditional patches for improving the safety of Linux. Some Linux developers have also begun to use the tool. Over 800 patches developed using Coccinelle have been integrated into the mainline Linux kernel. As part of the ABL ANR project, we are building on the results of Coccinelle by designing semantic patches to identify API protocols and detect violations in their usage [24].

Another work done as part of the ANR ABL project, and as the topic of Suman Saha’s PhD thesis, is the improvement of error handling code in Linux. We developed a program analysis for identifying the code structures used to represent error handling code and a transformation to convert existing error handling code to use gotos to shared cleanup code, which is the style preferred by the Linux community [53]. We subsequently worked on finding bugs in error handling code, following an approach that focuses on local patterns, i.e., within the current function, rather than patterns occurring across the entire code base. This approach has a low rate of false positives and can find bugs in the use of rarely called functions [39].

6.5. Hosted database replication service

Participant: Mesaac Makpangou [correspondent].

Today, the vast majority of content distributed on the web are produced by web 2.0 applications. Examples of such applications include social networks, virtual universities, multi-players games, e-commerce web sites, and search engines. These applications rely on databases to serve end-users’ requests. Hence, the success of these applications/services depends mainly on the scalability and the performance of the database backend.
The objective of our research is to provide a hosted database replication service. With respect to end-users applications, this service offers an interface to create, to register, and to access databases. Internally, each hosted database is fragmented and its fragments are replicated towards a peer-to-peer network. We anticipate that such a service may improve the performance and the availability of popular web applications, thanks to partial replications of backend databases. Partial database replication on top of a peer-to-peer network raises a number of difficult issues: (i) enforcing replica consistency in presence of update transactions, without jeopardizing the scalability and the performance of the system? (ii) accommodating the dynamic and the heterogeneity of a peer-to-peer network with the database requirements?

We designed a database access protocol, capable to spread out a transaction’s accesses over multiple database fragments replicas while guaranteeing that each transaction observes a consistent distributed snapshot of a partially replicated database. We have also proposed a replica control substrate that permits to enforce 1-Copy SI for database fragments replicated over a wide area network. For that, unlike most database replication, we separate the synchronisation from the certification concerns.

A small-scale group of schedulers that do not hold database replicas, cooperate with one another to certify update transactions. Only certified transactions are notified to replicas. Furthermore, each replica will be notified only the transactions that impact the that it stores. Thanks to this separation, we avoid waste of computation resource at replicas that will be used to decide whether to abord or commit an update transaction; Our design choices also permit to reduce bandwidth consumption.

In 2010, we focus on the development of a prototype implementation of the complete system. The current prototype includes: a tool that helps fragment a database into fragments; the support to deploy dynamically, for each fragment, the suitable number of replicas towards the network of hosting peers; the implementation of our proposal (i.e. our database access protocol and our replica control substrate); and the JDBC API extension for accessing replicated databases.

6.6. Commitment protocols for WAN replication

Participants: Marc Shapiro [correspondent], Pierre Sutra, Masoud Saeida Ardekani.

In a large-scale distributed system, replication is an essential technique for improving availability and read performance. However, writes raise the issue of consistency, especially in the presence of concurrent updates, network failures, and hardware or software crashes. So-called consensus constitutes a major primitive to solving these issues. The performance of large-scale systems depends crucially on the latency of consensus, especially in wide-area networks; to decrease it, we focus on generalised consensus algorithms, i.e., ones that leverage the commutativity of operations and/or the spontaneous ordering of messages by the network. One such algorithms is Generalized Paxos, which does not order concurrent commutative operations. However, when a collision occurs (i.e., two replicas receive non-commuting operations in a different order) Generalized Paxos requires a very high latency to recover, completely negating the gain. We designed FGGC, a new generalised consensus algorithm that minimises the cost of recovering from a collision, without decreasing resilience to faults. FGGC achieves the optimal latency (two communication steps when processes receive non-commutative operations in the same order, and three otherwise) when there are no faults. FGGC remains optimally fault-tolerant, as it tolerates \( f < n/2 \) crash faults and requires only \( f + 1 \) processes to make progress. Our experimental evaluation of FGGC shows that it is more efficient than the competing protocols. Another topic of relevance in WANs is partial replication, i.e., where any given server holds only a fraction of all shared objects. This decreases the workload per server and improves access times. However, this makes transactional concurrency control more difficult; indeed most existing algorithms assume full replication. We designed and implemented two genuine consensus protocols for partial replication, i.e., ones in which only relevant replicas need participate in the commit of a transaction. They were evaluated experimentally above the BerkeleyDB database engine. This work is the topic of Pierre Sutra’s PhD thesis.

6.7. Optimistic approaches in collaborative editing

Participants: Marc Shapiro [correspondent], Marek Zawirski, Pierpaolo Cincilla.
In recent years, the Web has seen an explosive growth of massive collaboration tools, such as wiki and weblog systems. By the billions, users may share knowledge and collectively advance innovation, in various fields of science and art. Existing tools, such as the MediaWiki system for wikis, are popular in part because they do not require any specific skills. However, they are based on a centralised architecture and hence do not scale well. Moreover, they provide limited functionality for collaborative authoring of shared documents.

A natural research direction is to use P2P techniques to distribute collaborative documents. This raises the issue of supporting collaborative edits, and of maintaining consistency, over a massive population of users, shared documents, and sites.

In order to avoid complex and unnatural concurrency control and synchronisation, and to enable different styles of collaboration (from online “what you see is what I see” to fully asynchronous disconnected work) we invented the concept of a Commutative Replicated Data Type (CRDT). A CRDT is one where all concurrent operations commute. The replicas of a CRDT converge automatically, without complex concurrency control.

In the context of collaborative editing, we propose, a novel CRDT design called Treedoc. An essential property is that the identifiers of Treedoc atoms are selected from a dense space. We study practical alternatives for implementing the identifier space based on an extended binary tree. We also focus storage alternatives for data and meta-data, and mechanisms for compacting the tree. In the best case, Treedoc incurs no overhead with respect to a linear text buffer. We validate the results with traces from existing edit histories.

Treedoc will be used in ANR projects PROSE (Section 7.1.5) and STREAMS, and will be further studied and developed in ANR project ConcoRDanT (Section 7.1.3) and under a Google European Doctoral Fellowship.

6.8. CRDTs, a principled approach to eventual consistency

Participants: Marc Shapiro [correspondent], Marek Zawirski.

Most well-studied approaches to replica consistency maintain a global total order of operations. This serialisation constitutes a performance and scalability bottleneck, while the CAP theorem imposes a trade-off between consistency and partition-tolerance. An alternative approach, eventual consistency or optimistic replication, is attractive. A replica may execute an operation without synchronising a priori with other replicas. The operation is sent asynchronously to other replicas; every replica eventually applies all updates, but possibly in different orders. This approach ensures that data remains available despite network partitions, and is perceived to scale well and to provide acceptable quality of service. The consensus bottleneck remains but is off the critical path. However, reconciliation is generally complex. There is little theoretical guidance on how to design a correct optimistic system, and ad-hoc approaches have proven brittle and error-prone. We propose a simple, theoretically sound approach to eventual consistency, the concept of a convergent or commutative replicated data type (CRDT), for which some simple mathematical properties ensure eventual consistency. Provably, any CRDT converges to a common state that is equivalent to some sequential execution. A CRDT requires no synchronisation, thus every update can execute immediately, unaffected by network latency, faults, or disconnection. It is extremely scalable and is fault-tolerant, and does not require much mechanism. Previously, only a handful of CRDTs were known. Our current research aims to push the CRDT envelope, to study the principles of CRDTs, and to design a library of useful CRDTs. So far we have designed variations on registers, counters, sets, maps (key-value stores), graphs, and sequences. Potential application areas include computation in delay-tolerant networks, latency tolerance in wide-area networks, disconnected operation, churn-tolerant peer-to-peer computing, and partition-tolerant cloud computing. CRDTs are the main topic of ANR project ConcoRDanT (Section 7.1.3). This research is also funded in part by a Google European Doctoral Fellowship.
5. New Results

5.1. Package understanding and Assessing

Participants: Stéphane Ducasse, Nicolas Anquetil, Usman Bhatti, Jannik Laval.

To support the understanding of large systems is to offer ways to understand and fix dependencies between software elements. We worked on how to semi-automatically reorganize packages to minimize coupling.

Efficient Retrieval and Ranking of Undesired Package Cycles in Large Software Systems. Many design guidelines state that a software system architecture should avoid cycles between its packages. Yet such cycles appear again and again in many programs. We believe that the existing approaches for cycle detection are too coarse to assist the developers to remove cycles from their programs. We describe an efficient algorithm that performs a fine-grained analysis of the cycles among the packages of an application. In addition, we define a metric to rank cycles by their level of undesirability, prioritizing the cycles that seem the least desirable to the developers. Our approach is validated on two large and mature software systems in Java and Smalltalk. [19]

Legacy Software Restructuring: Analyzing a Concrete Case. Software re-modularization is an old pre-occupation of reverse engineering research. The advantages of a well structured or modularized system are well known. Yet after so much time and efforts, the field seems unable to come up with solutions that make a clear difference in practice. Recently, some researchers started to question whether some basic assumptions of the field were not overrated. The main one consists in evaluating the high-cohesion/low-coupling dogma with metrics of unknown relevance. In this paper, we study a real structuring case (on the Eclipse platform) to try to better understand if (some) existing metrics would have helped the software engineers in the task. Results show that the cohesion and coupling metrics used in the experiment did not behave as expected and would probably not have helped the maintainers reach there goal. We also measured another possible restructuring which is to decrease the number of cyclic dependencies between modules. Again, the results did not meet expectations. [14]

An empirical model for continuous and weighted metric aggregation. It is now understood that software metrics alone are not enough to characterize software quality. To cope with this problem, most of advanced and/or industrially validated quality models aggregate software metrics: for example, cyclomatic complexity is combined with test coverage to stress the fact that it is more important to cover complex methods than accessors. Yet, aggregating and weighting metrics to produce quality indexes is a difficult task. Indeed certain weighting approaches may lead to abnormal situations where a developer increasing the quality of a software component sees the overall quality degrade. Finally, mapping combinations of metric values to quality indexes may be a problem when using thresholds. In this paper [20], we present the problems we faced when designing the Squale quality model, then we present an empirical solution based on weighted aggregations and on continuous functions. The solution has been termed the Squale quality model and validated over 4 years with two large multinational companies: Air France-KLM and PSA Peugeot-Citroen.

Modularization Metrics: Assessing Package Organization in Legacy Large Object-Oriented Software. In systems consisting of several thousands of classes, classes cannot be considered as units for software modularization. In such context, packages are not simply classes containers, but they also play the role of modules: a package should focus on providing well identified services to the rest of the software system. Therefore, understanding and assessing package organization is primordial for software maintenance tasks. Although there exist a lot of works proposing metrics for the quality of a single class and/or the quality of inter-class relationships, there exist few works dealing with some aspects for the quality of package organization and relationship. We believe that additional investigations are required for assessing package modularity aspects. The goal of these papers [13], [28] is to provide a complementary set of metrics that assess some modularity principles for packages in large legacy object-oriented software: Information-Hiding, Changeability and Reusability principles. Our metrics are defined with respect to object-oriented dependencies
that are caused by inheritance and method call. We validate our metrics theoretically through a careful study of the mathematical properties of each metric.

5.2. Tools and Tool Infrastructure

Participants: Stéphane Ducasse, Veronica Uquillas-Gomez, Jannik Laval.

Reengineering large applications implies an underlying tool infrastructure that can scale and also be extended.

Ring: a Unifying Meta-Model and Infrastructure for Smalltalk Source Code Analysis Tools. Source code management systems record different versions of code. Tool support can then compute deltas between versions. To ease version history analysis we need adequate models to represent source code entities. As a first step to provide an infrastructure to support history analysis, this article [12] presents Ring, a unifying source code meta-model that can be used to support several activities and proposes a unified and layered approach to be the foundation for building an infrastructure for version and stream of change analyses. We re-implemented three tools based on Ring to show that it can be used as the underlying meta-model for remote and off-image browsing, scoping refactoring, and visualizing and analyzing changes. As a future work and based on Ring we will build a new generation of history analysis tools.

AspectMaps: A Scalable Visualization of Join Point Shadows. When using Aspect-Oriented Programming, it is sometimes difficult to determine at which join point an aspect executes. Similarly, when considering one join point, knowing which aspects will execute there and in what order is non-trivial. This makes it difficult to understand how the application will behave. A number of visualizations have been proposed that attempt to provide support for such program understanding. However, they neither scale up to large code bases nor scale down to understanding what happens at a single join point. In this paper [18], we present AspectMaps - a visualization that scales in both directions, thanks to a multi-level selective structural zoom. We show how the use of AspectMaps allows for program understanding of code with aspects, revealing both a wealth of information of what can happen at one particular join point as well as allowing to see the “big picture” on a larger code base. We demonstrate the usefulness of AspectMaps on an example and present the results of a small user study that shows that AspectMaps outperforms other aspect visualization tools.

Challenges to support automated random testing for dynamically typed languages. Automated random testing is a proved way to identify bugs and precondition violations, and this even in well tested libraries. In the context of statically typed languages, current automated random testing tools heavily take advantage of static method declaration (argument types, thrown exceptions) to constrain input domains while testing and to identify errors. For such reason, automated random testing has not been investigated in the context of dynamically typed languages. We present the key challenges that have to be addressed to support automated testing in dynamic languages. [17]

SmartGroups, Focusing on Task-Relevant Source Artifacts in IDEs. Navigating large software systems, even when using a modern IDE (Integrated Development Environment) is difficult, since conceptually related software artifacts are distributed in a huge software space. For most software maintenance tasks, only a small fraction of the entire software space is actually relevant. The IDE, however, does not reveal the task relevancy of source artifacts, thus developers cannot easily focus on the artifacts required to accomplish their tasks. Smart Groups help developers to perform software maintenance tasks by representing groups of source artifacts that are relevant for the current task. Relevancy is determined by analyzing historical navigation and modification activities, evolutionary information, and runtime information. The prediction quality of Smart Groups is validated with a benchmark evaluation using recorded development activities and evolutionary information from versioning systems. [24]

5.3. Constructs for Dynamic Languages

Participants: Stéphane Ducasse, Marcus Denker, Veronica Uquillas-Gomez, Gwenael Casaccio, Camillo Bruni, Jean-Baptiste Arnaud, Damien Pollet.

To support our research on secure dynamic languages, we focused on improving language infrastructure.
Efficient Proxies in Smalltalk. A proxy object is a surrogate or placeholder that controls access to another target object. Proxy objects are a widely used solution for different scenarios such as remote method invocation, future objects, behavioral reflection, object databases, inter-languages communications and bindings, access control, lazy or parallel evaluation, security, among others. Most proxy implementations support proxies for regular objects but they are unable to create proxies for classes or methods. Proxies can be complex to install, have a significant overhead, be limited to certain type of classes, etc. Moreover, most proxy implementations are not stratified at all and there is no separation between proxies and handlers. We present Ghost, a uniform, light-weight and stratified general purpose proxy model and its Smalltalk implementation. Ghost supports proxies for classes or methods. When a proxy takes the place of a class it intercepts both, messages received by the class and lookup of methods for messages received by instances. Similarly, if a proxy takes the place of a method, then the method execution is intercepted too. [22]

Bootstrapping a Smalltalk. Smalltalk is a reflective system. It means that it is defined in itself in a causally connected way. Traditionally, Smalltalk systems evolved by modifying and cloning what is called an image (a chunk of memory containing all the objects at a given point in time). During the evolution of the system, objects representing it are modified. However, such an image modification and cloning poses several problems: (1) There is no operational machine-executable algorithm that allows one to build a system from scratch. A system object may be modified but it may be difficult to reproduce its exact state before the changes. Therefore it is difficult to get a reproducible process. (2) As a consequence, certain classes may not have been initialized since years. (3) Finally, since the system acts as a living system, it is not simple to evolve the kernel for introducing new abstractions without performing some kind of brain surgery on oneself. There is a need to have a step by step process to build Smalltalk kernels from scratch. After an analysis of past and current practices to mutate or generate kernels, we describe a kernel bootstrap process step-by-step. First the illusion of the existence of a kernel is created via stubs objects. Second the classes and meta-classes hierarchy are generated. Code is compiled and finally information needed by the virtual machine and execution are generated and installed. [15]

Flexible Object Layouts: Enabling Lightweight Language Extensions by Intercepting Slot Access. Programming idioms, design patterns and application libraries often introduce cumbersome and repetitive boilerplate code to a software system. Language extensions and external DSLs (domain specific languages) are sometimes introduced to reduce the need for boilerplate code, but they also complicate the system by introducing the need for language dialects and inter-language mediation. To address this, we propose to extend the structural reflective model of the language with object layouts, layout scopes and slots. Based on the new reflective language model we can 1) provide behavioral hooks to object layouts that are triggered when the fields of an object are accessed and 2) simplify the implementation of state-related language extensions such as stateful traits. By doing this we show how many idiomatic use cases that normally require boilerplate code can be more effectively supported. We present an implementation in Smalltalk, and illustrate its usage through a series of extended examples. [25]

5.4. Resources

Participants: Stéphane Ducasse, Marcus Denker, Mariano Martinez-Peck, Nick Papoylias.

Resource management is important in the context of resource constrained devices as well as situations were a large amount of data is modeled but not accessed often. One example for resource constrained devices are autonomous robots. An example for large models that are accessed infrequently are typical models of systems for software re-engineering.

With Ecole des Mines de Douai we explore how to analyze and improve memory in the case of unused data.

Problems and Challenges when Building a Manager for Unused Objects. Large object-oriented applications may occupy hundreds of megabytes or even gigabytes of memory. During program execution, a large graph of objects is created and constantly changed. Most object runtimes support some kind of automatic memory management based on garbage collectors (GC) whose idea is the automatic destruction of unreferenced objects. However, there are referenced objects which are not used for a long period of time or that are
used just once. These are not garbage-collected because they are still reachable and might be used in the future. Due to these unused objects, applications use much more resources than they actually need. We present the challenges and possible approaches towards an unused object manager for Pharo. The goal is to use less memory by swapping out the unused objects to secondary memory and leaving in primary memory only those objects that are needed and used. When one of the unused objects is needed, it is brought back into primary memory. [23]

**Clustered Serialization with Fuel.** Serializing object graphs is an important activity since objects should be stored and reloaded on different environments. There is a plethora of frameworks to serialize objects based on recursive parsing of the object graphs. However such approaches are often too slow. Most approaches are limited in their provided features. For example, several serializers do not support class shape changes, global references, transient references or hooks to execute something before or after being stored or loaded. Moreover, to be faster, some serializers are not written taking into account the object-oriented paradigm and they are sometimes even implemented in the Virtual Machine hampering code portability. VM-based serializers such as ImageSegment are difficult to understand, maintain, and fix. For the final user, it means a serializer which is difficult to customize, adapt or extend to his own needs. We present a general purpose object graph serialization based on a pickling format and algorithm. We implement and validate this approach in the Pharo Smalltalk environment. We demonstrate that we can build a really fast serializer without specific VM support, with a clean object-oriented design, and providing most possible required features for a serializer. We show that our approach is faster that traditional serializers and compare favorably with ImageSegment as soon as serialized objects are not in isolation. [16]

**Towards Structural Decomposition of Reflection with Mirrors** Mirrors are meta-level entities introduced to decouple reflection from the base-level system. Current mirror-based systems focus on functional decomposition of reflection. We advocate that mirrors should also address structural decomposition. Mirrors should not only be the entry points of reflective behavior but also be the storage entities of meta-information. This decomposition can help resolve issues in terms of resource constraints (e.g., embedded systems and robotics) or security. Indeed, structural decomposition enables discarding meta-information. [21]

### 5.5. Empirical Studies in Software Product Line Engineering

**Participant:** Nicolas Anquetil.

Software Product Line Engineering (SPLE) is a new development paradigm that promises to offer faster development, with better quality. The idea is to develop a generic application (the Software Product Line) with pre-defined variation points. From this, new applications are derived from the generic application and the options are chosen for the possible variation points. Because it is still a new paradigm, Software Product Line development is still an active research domain where empirical research is useful to check the validity of the results.

These publications are the results of an earlier research project to which N. Anquetil participated.

**Managing information flow in the SPL development processes.** Traceability is a quality attribute in software engineering that establishes the ability to describe and follow the life of a requirement in both the forward and backward directions (i.e., from its origins through its specification, implementation, deployment, use and maintenance, and vice-versa). The IEEE Standard Glossary of Software Engineering Terminology defines traceability as “the degree to which a relationship can be established between two or more products of the development process, especially products having a predecessor-successor or master-subordinate relationship to one another”. According to (Palmer, 1997) “traceability gives essential assistance in understanding the relationships that exist within and across software requirements, design, and implementation”. Thus, trace relationships help in identifying the origin and rationale for artefacts generated during development lifecycle and the links between these artefacts. Identification of sources helps understanding requirements evolution and validating implementation of stakeholders’ requirements. The main advantages of traceability are: (i) to relate software artefacts and design decisions taken during the software development cycle; (ii) to give feedback to architects and designers about the current state of the development, allowing them to reconsider alternative
design decisions, and to track and understand bugs; and (iii) to ease communication between stakeholders. [26]

**Empirical research in software product line engineering.** Empirical evaluation has for many years been utilized to validate theories in other science disciplines. One of the first well-known reported examples of empirical evaluation occurred when Galileo wanted to prove that the rate of descent of objects was independent of their mass. This would disprove a theory put forward by Aristotle that the rate of descent is directly proportional to their weight. To prove his theory Galileo dropped two balls made from the same material but different masses from the top of the Tower of Pisa. When the experiment was performed Galileo’s theory was proved correct through the empirical evidence collected. What this story demonstrates is the importance of empirical validation to verify or disprove theories and hypotheses. The purpose of this publication [27] is to emphasize the importance and difficulties of empirical evaluation in the domain of SPLE.
5. New Results

5.1. Languages and Foundations: Process algebra

Participants: Damien Pous, Alan Schmitt, Jean-Bernard Stefani, Claudio Mezzina, Cinzia di Giusto.

The goal of this work is to study process algebraic foundations for component-based distributed programming. Most of this work takes place in the context of the ANR PiCoq project.

To develop composable abstractions for programming dependable systems, we investigate concurrent reversible models of computation, where arbitrary executions can be reversed, step by step, in a causally consistent way. This year we have continued the study of the reversible higher-order pi-calculus and obtained a new encoding of it in the higher-order pi-calculus which improves on the result we published in Concur 2010 by proving the faithfulness of the encoding with a much finer equivalence relation. We also developed a reversible variant of the higher-order pi-calculus where reversibility can be controlled by means of an explicit rollback primitive [37]. We have proved that this rollback primitive is sound and complete in that it provides a causally consistent and complete reversal of concurrent computations, and we have developed a low-level semantics for this primitive, closer to an actual distributed implementation, which we have proved equivalent to the high-level one. All these results are presented in detail in Claudio Mezzina’s forthcoming PhD thesis, and have been developed in cooperation with the INRIA Focus team at the University of Bologna.

An interesting and expressive component model for embedded systems is the BIP component model [58], developed by J. Sifakis’ team at the Verimag Laboratory, which features hierarchical software architectures, explicit constructs for specifying component compositions (glues), and multipoint synchronization under priority constraints. We have begun a process calculus analysis of BIP, with a view to combine the reactive features of BIP with the dynamic reconfiguration features of Fractal. Our first result takes the form of new process calculus, called CAB, which we have proved to be a conservative extension of BIP. CAB also enabled us to study the intrinsic expressivity of the BIP model and to prove that priority constraints are essential to BIP expressivity [34].

We have made significant progress on the formalization in the Coq proof assistant of a core higher-order π-calculus, called HOcore [20]. We have in particular adapted a canonical locally nameless representation of binding to handle alpha-conversion in our formalization. Several major theorems of HOcore, in particular the fact that IO-bisimulation is correct in relation to barbed congruence and is decidable. This work has been submitted for publication.

A longer version of our work on untyping theorems and cyclic linear logic has been accepted for publication in LMCS [24], and a book chapter on up-to techniques for bisimulations, written with Davide Sangiorgi from the INRIA Focus team in Bologna, has been published by Cambridge University Press [46].

Together with Filippo Bonchi (LIP, ENS Lyon), we have worked on a new algorithm for checking the language equivalence of non-deterministic finite automata (NFA). This algorithm improves on the standard Hopcroft and Karp’ algorithm, by using up-to techniques. The first empirical tests look really promising [47].

Together with Tom Hirschowitz (LAMA, U. de Chambéry), we have worked on a categorical model of CCS, where innocent strategies are pre-sheaves. This work has been presented at the ICE workshop [36], and a long version has been submitted to SACS.

5.2. Languages and Foundations: Proof tactics

Participants: Damien Pous, Thomas Braibant.
The goal of this work is to develop proof-assistant-based tools for verifying distributed systems and distributed abstract machines. In particular, we aim to support the derivation of fully formal proofs of correctness for abstract machines supporting the component-based languages and programming models we develop.

We have presented our work about tools for rewriting modulo AC in Coq at CPP’11 [32]. An extended version of our work on Kleene algebra (ATBR, first published at ITP’10), was accepted for publication in LMCS [19]. Also on the Coq side, we have developed a library for verifying hardware circuits, which was also presented at CPP’11 [31].

### 5.3. Control for adaptive systems: Discrete control for adaptive and reconfigurable systems

**Participants:** Eric Rutten, Noël de Palma, Olivier Gruber, Fabienne Boyer, Tayeb Bouhadiba, Xin An.

The goal of this work is to apply control techniques based on the behavioral model of reactive automata and the algorithmic techniques of discrete controller synthesis. We adopt the synchronous approach to reactive systems, and use an associated effective controller synthesis tool, Sigali, developed at INRIA Rennes. Both are integrated into a programming language, called BZR, and its compiler, as an extension of the Heptagon language. We thus have a complete tool-supported method from control modeling down to concrete execution, considering different execution models, and targetting either software or hardware. We explore control theory for computer science, as an original alternative to computer science for control (as more usually in embedded systems), and to classical discrete control systems (as more usually applied to manufacturing). We are exploring several target application domains, where we expect to find commonalities in the control problems, and variations in the definitions of configurations, and in the criteria motivating adaptation.

We have obtained this year the following results:

- At the programming language level, we are continuing the development of a modelling and controller generation language called BZR, which involves DCS in its compilation. BZR is designed and developed in cooperation with the Pop Art and VerTecs (INRIA Rennes) teams [40].
- We have developed a technique for designing reconfiguration controllers in the Fractal component-based framework, where discrete control loops automatically enforce safety properties on the interactions between components, concerning, e.g., mutual exclusions, forbidden or imposed sequences [29] [48].
- We have integrated BZR with Orccad, a programming environment for real-time control systems, in cooperation with the NeCS and SED teams at INRIA Grenoble [28].
- We are investigating administration loops in virtual machine-based distributed systems [44], and the coordination of such loops, especially in relation with green computing problems. We are starting a new ANR project called Ctrl-Green on this topic in 2012.
- We work on the formal modelling and control of dynamic reconfiguration in FPGA circuits, in cooperation with the DaRT team (INRIA Lille) [43] and the Lab-STIC laboratory in Lorient [42], building upon earlier work related to the MARTE framework.
- In cooperation with GIPSA-Lab and ENSI Tunis, we have adapted the discrete controller synthesis technique to the control of decentralized systems that are composed of several subsystems spread across remote sites [17].
- In cooperation with Orange labs and GIPSA-Lab we are beginning to explore the application of discrete event systems and supervisory control to the domain of Machine to Machine and Internet of Things, with the objective to manage energy aspects; this will start with the CIFRE PhD (U. Grenoble) of Mengxuan Zhao (co-advised with H. Alla, G. Privat).

### 5.4. System configuration and deployment

**Participants:** Loris Bouzonnet, Fabienne Boyer, Willy Malvault, Noël de Palma, Vivien Quéma, Jean-Bernard Stefani.
The goal of this work is to study system configuration and software deployment issues in large distributed systems.

System configuration and software deployment in a distributed environment can be greatly aided by the use of a uniform component model to support software assembly, software configuration and deployment, as well as runtime system configuration. We have developed a specialization of the Fractal component model that provides a reference model for heterogeneous software assembly and configuration. In particular, we have shown how this reference model can be used to assemble and configure software architectures built from heterogeneous software packages (e.g. OSGI bundles for Java packages, Debian or RPM packages for Linux modules and applications). The definition of this model, a description of its implementation and its evaluation are documented in Loris Bouzonnet’s PhD thesis [13].

As an alternative to current public cloud infrastructures, which rely on large data centers, we have started the study of a cloud infrastructure based on a peer-to-peer (P2P) overlay network built on gossip-based protocols. More precisely, we have studied how to implement a distributed resource allocation service in a P2P environment maintained by a gossip-based peer-sampling protocol [81]. The resulting system, called Salute, provides for the allocation of application-specific overlays out of an underlying P2P network. By combining several P2P services (including peer-sampling, topology maintenance, and node synchronization), and by partitioning available nodes into free nodes (available for the allocation of new application overlays) and reserve nodes (nodes dedicated to the maintenance of allocated overlays), Salute provides a churn-resistant, completely decentralized cloud infrastructure. In addition, we have shown that Salute can provide its allocation service while maintaining fairness and avoiding starvation. The Salute architecture has been validated through simulations using network traces from different real-world P2P environments. The Salute architecture, algorithms and their validation are documented in Willy Malvault’s PhD thesis [14].

In a cloud computing context the complexity of deploying and configuring non-trivial software architectures is exacerbated. In line with our previous work on architecture-based distributed system management, we have proposed a novel algorithm for configuring component-based distributed applications deployed within several virtual machines in an IaaS environment. The algorithm is completely decentralized, relies on a message queuing middleware and exploits the software architecture descriptions of the applications to deploy and configure, written in an extension of the Fractal Architecture Description Language. A first version of this algorithm, that does not take into account potential failures during the configuration process, has been formally specified in collaboration with Gwen Salaün from the INRIA Vasy team in Grenoble, and presented at IEEE Cloud 2011 [35].

5.5. System support: System support for multicore machines

Participants: Vivien Quema, Renaud Lachaize, Fabien Gaud, Baptiste Lepers, Sylvain Genevès, Fabien Mottet.

Multicore machines with Non-Uniform Memory Accesses (NUMA) are becoming commodity platforms. Efficiently exploiting their resources remains an open research problem. Most of the body of existing work focuses on increasing locality between computations and memory or I/O resources. This is achieved by allocating data items preferably in local memory nodes, by moving computations close to I/O devices or by moving already allocated memory pages close to the applications which use them most. In all these works, researchers always assume that all processors have equal memory performance. Nevertheless, this assumption is not always valid. In 2011, we have studied the performance achieved by a 16-core NUMA exhibiting an irregular connectivity between processors. Some processors are directly connected to all other processors and access memory nodes with a low latency. Other processors have a lower degree of connectivity and need more hops to access some memory nodes and access memory with a higher latency.

Current operating systems are not aware of such performance characteristics. We have shown that the completion time of applications taken from the PARSEC benchmark suite can vary by up to 15% depending on the processor they are scheduled on. We have thus proposed a new OS scheduler that takes this asymmetry into account in order to make efficient decisions. This scheduler relies on a new metric, called MAPI (number
of main Memory Accesses Per retired Instruction), to predict the impact of processor interconnect asymmetry on the performance of applications. We have empirically evaluated the relevance of this metric on applications taken from the PARSEC benchmark suite. We have shown that this metric helps estimating the performance gap between running an application on a "well-interconnected" processor and on a "weakly-interconnected" one. Using this metric, the scheduler we proposed makes efficient decisions. More precisely, we have observed that the scheduler always performed within 3% of the best possible scheduling decision. This work is currently under submission.

5.6. System support: Protocols for resilient systems

Participants: Vivien Quéma, Alessio Pace.

We have worked on replication protocols for P2P systems. In particular, we have worked on replication in so-called Distributed Hash-Tables (DHTs). DHTs provide a simple high-level put/get abstraction that can be used to build efficient distributed storage systems. DHTs gained wide popularity in the last decade, fostering a large amount of interest in the academia, and inspiring the design of key/value distributed storage systems deployed in production.

DHTs provide a way to deterministically map objects to nodes and allow efficiently retrieving objects in a distributed fashion. Nodes and objects are logically arranged in a large numeric key-space, according to a given variant of consistent hashing. Typically, the node in charge of an object is the one whose position immediately follows the object in the key-space.

To guarantee that objects are reliably stored, DHTs rely on replication. A replication protocol is in charge of ensuring that, at any time, each object is replicated on a sufficiently large number of replicas. Several replication strategies have been proposed in the last years. The most efficient ones use predictions about the availability of nodes to reduce the number of object migrations that need to be performed: objects are preferably stored on highly available nodes.

We have proposed an alternative replication strategy. Rather than exploiting highly available nodes, we have designed a protocol that leverages nodes that exhibit regularity in their connection pattern. Roughly speaking, the strategy consists in replicating each object on a set of nodes that is built in such a way that, with high probability, at any time, there are always at least k nodes in the set that are available. We have evaluated this new replication strategy using traces of two real-world systems: eDonkey and Skype. Our evaluation showed that our regularity-based replication strategy induces a systematically lower network usage than existing state-of-the-art replication strategies. This work has been published at the International Symposium on Reliable Distributed Systems, in October 2011.

5.7. System support: End-to-end caching

Participants: Sara Bouchenak, Dàmian Serrano.

Cloud Computing is a paradigm for enabling remote, on-demand access to a virtually infinite set of configurable computing resources. This model aims to provide hardware and software services to customers, while minimizing human efforts in terms of service installation, configuration and maintenance, for both cloud provider and cloud customer. A cloud may have the form of an Infrastructure as a Service (SaaS), a Platform as a Service (PaaS) or a Software as a Service (SaaS). Clouds pose significant challenges to the full elasticity of clouds, their scalability and their dependability in large scale data management and large scale computing resources. Caching is a means for high performance and scalability of distributed systems. Although caching solutions have been successfully studied for individual systems such as database systems or web servers, if collectively applied, these solutions violate the coherence of cached data. We precisely studied this issue in e-Caching, a novel end-to-end caching system.

The contribution of this work is twofold: guaranteeing the coherence of cached data at multiple locations of a distributed system, while improving the overall performance of the system. In collaboration with Marta Patino and Ricardo Jimenez from Universidad Politecnica de Madrid, we proposed a novel distributed caching protocol, implemented it and evaluated it with real online services. The experiments showed that e-Caching was successfully able to improve service performance by two orders of magnitude.
5.8. System support: Performance and dependability benchmarking

Participants: Amit Sangroya, Dàmian Serrano, Sara Bouchenak [correspondant].

MapReduce has become a popular programming model and runtime environment for developing and executing distributed data-intensive and compute-intensive applications. It offers developers a means to transparently handle data partitioning, replication, task scheduling and fault tolerance on a cluster of commodity computers. MapReduce allows a wide range of applications such as log analysis, data mining, Web search engines, scientific computing, bioinformatics, decision support and business intelligence.

There has been a large amount of work on MapReduce towards improving its performance and reliability. Several efforts have explored task scheduling policies in MapReduce, cost-based optimization techniques, replication and partitioning policies. There has also been a considerable interest in extending MapReduce with other fault tolerance models, or with techniques from database systems. However, there has been very little in the way of empiric evaluation for the comparison of the different systems. Most evaluations of these systems have relied on microbenchmarks based on simple MapReduce programs. While microbenchmarks may be useful in targetting specific system features, they are not representative of full distributed applications, and they do not provide multi-user realistic workloads. Furthermore, as far as we know, no studies have investigated dependability benchmarking of MapReduce.

Thus, we provide MapReduce Benchmarking (MRB), a novel MapReduce benchmark suite to enable a thorough analysis of a wide range of features of MapReduce systems. MRB has the following features. First, it enables the empirical evaluation of the performance and dependability of MapReduce systems. This provides a means to analyze the effectiveness of scalability and fault tolerance, two key features of MapReduce. Second, it covers a variety of application domains, workload and faultload characteristics, ranging from compute-oriented to data-oriented applications, batch applications to online real-time applications. While MapReduce frameworks were originally limited to offline batch processing, recent works are exploring the extension of MapReduce beyond batch processing. Moreover, in order to stress MapReduce dependability and performance, the benchmark suite enables different fault injection rates, workloads and concurrency levels. Finally, the benchmark suite is portable and easy to use on a wide range of platforms, covering different MapReduce frameworks and cloud infrastructures. This work has been submitted for publication.

5.9. System support: Self-adaptive Internet services

Participant: Sara Bouchenak.

Although distributed services provide a means for supporting scalable Internet applications, their ad-hoc provisioning and configuration pose a difficult tradeoff between service performance and availability. This is made harder as Internet service workloads tend to be heterogeneous, and vary over time in amount of concurrent clients and in mixture of client interactions. This work proposes an approach for building self-adaptive Internet services through utility-aware capacity planning and provisioning. First, an analytic model is presented to predict Internet service performance, availability and cost. Second, a utility function is defined and a utility-aware capacity planning method is proposed to calculate the optimal service configuration which guarantees SLA performance and availability objectives while minimizing functioning costs. Third, an adaptive control method is proposed to automatically apply the optimal configuration to the Internet service. Finally, the proposed model, capacity planning and control methods are implemented and applied to an online bookstore. Experimental evaluations show that the service successfully self-adapts to both workload mix and workload amount variations, and present significant benefits in terms of performance and availability, with a saving of resources underlying the Internet service.

This work is part of the MyCloud ANR project. It has been described in a chapter of the book titled Performance and Dependability in Service Computing, 2011. There has been an industrial transfer of the MoKA software prototype.
5.10. Self-Configuration of distributed system in the Cloud

Participants: Fabienne Boyer, Noël de Palma.

Cloud computing environments fall under three main kinds of offers according to the resources they provide. The Infrastructure as a Service (IaaS) level enables the access to virtualized hardware resources (processing, storage and network). The Software as a Service (SaaS) layer aims at providing the end-users with software applications. The intermediary layer, called Platform as a Service (PaaS), offers a set of tools and runtime environments that allow managing the applications life-cycle. This life-cycle includes the phases related to the design, the development, the deployment of applications, and generally speaking all their management stages (workload, fault tolerance, security). This article focuses on the deployment of distributed applications in virtualized environments such as cloud computing. Such deployments require to generate the virtual images that will be instantiated as virtual machines, thus ensuring the execution of the application on an IaaS platform. Each image embeds technical elements (operating system, middleware pieces) and functional ones (data and applicative software entities). Once it has been instantiated, each virtual machine is subjected to a stage of dynamic settings, which finalizes the global configuration of the distributed application.

On the whole, the deployment solutions currently available do not take into account these different configuration parameters, which are mostly managed by dedicated scripts. Moreover these solutions are not able to automate the images generation, their instantiation as virtual machines and their configuration independently from the kind of distributed application to be deployed. For instance, Google App Engine solution only deals with Web services organized into precisely defined tiers. In our opinion, the absence of general solutions results essentially from a lack of formalism for describing the distributed application architecture with its configuration constraints in a virtualized infrastructure such as cloud computing. Our work focused on a general solution, for Virtual Applications Management Platform, that automates the deployment of any distributed applications in the cloud. The suggested approach is architectural, meaning that it is based upon an explicit representation of the applications’ distributed architecture. We offer, on the one hand, a formalism for describing an application as a set of interconnected virtual machines and, on the other hand, an engine for interpreting this formalism and automating the application deployment on an IaaS platform. Specifically, we study three contributions:

- A formalism that offers a global view of the application to be deployed in terms of components with the associated configuration- and interconnection constraints and with their distribution within virtual machines. This formalism extends OVF language, dedicated to virtual machines description, with an architecture description language (ADL) that allows describing a distributed application software architecture;
- A deployment engine, i.e. a runtime support able to deploy automatically an application described with this formalism. This engine is based on a decentralized protocol for self-configuring the instantiated virtual machines. In our opinion it can ease the scalability of the dynamic configuration stage;
- A performance evaluation of the proposed solution on an industrial IaaS platform.

We published in this context two journal articles (TPDS [26] and TAAS [18]) and three conference papers (Cloud11 [35], UCC11 [39] and SAC12).

5.11. Virtual Machine

Participants: Olivier Gruber, Fabienne Boyer, Damien Pous, Ludovic Demontes, Clément Deschamps.

A core aspect of the Synergy virtual machine is its ability to reconfigure component-based applications at execution time. We have focused on the reconfiguration protocol with the intent of verifying and proving its robustness.
In a first step, we have formalized and verified that any correct and complex reconfiguration through our reconfiguration protocol can be processed as a sequence of elementary reconfiguration operations and always results in a component assembly that is architecturally consistent. This aspect has been verified using model-checking techniques. This work has been done in collaboration with Gwen Salaün from the VASY team (Inria Rhône-Alpes). It lead to a publication in the Formal Method (FM’11) conference [30].

In a second step, we have considered software failures that may occur during a reconfiguration. Although the protocol is trusted code, it invokes components to reconfigure them, thereby executing unsafe code that may fail. This work with Damien Pous produced a high-level formalisation of our reconfiguration protocol and a completely certified modelisation of these algorithms in Coq [50]. This work resulted in a submitted publication.

Finally, we have also investigated the control of complex reconfiguration through using discrete synchronous control techniques with Eric Rutten and Gwenael Delaval [44].
6. New Results


Processes have received a lot of attention in the last decade and proposed workflow solutions for office automation. The topic is subject today to a lot of interests carried by the expansion of business on the Web. However it is required need to satisfy new application requirements and execution contexts. We are interested in different aspects of process engineering: the management of change in business process; modeling and implementing Quality of Services properties (time, security, constraints...); composing existing process fragments of different nature and models; decentralizing a global process for a distributed execution with organizational constraints; process governance. Most of these aspects are considered within the frame of Web services and/or peer to peer architectures, and we are also interested in proposing new models of process for new applications domains.

6.1.1. Optimized decentralization and synchronization of cross-organizational business processes

Participants: Claude Godart, Walid Fdhila.

Globalization and the increase of competitive pressures created the need for agility in business processes, including the ability to outsource, offshore, 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. An organization that aims for such fragmentation of its business processes needs to be able to separate the process into different parts. Therefore, 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. Additionally, this needs to be done in a networked environment, which is where Service Oriented Architecture plays a vital role.

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. It also proposes methods to optimize the fragmentation process in terms of QoS properties and communication overhead [21], [10]. Moreover, it presents a framework to model web service choreographies in Event Calculus formal language.

Walid Fdhila has successfully defended his thesis on October, 7th [1].

6.1.2. A Declarative Approach to Web Services Computing

Participants: Olivier Perrin, Ehtesham Zahoor, Claude Godart.

Web services composition and monitoring are still highly active and widely studied research directions. Little work however has been done in integrating these two dimensions using an unified framework and formalism. Classical approaches introduce an additional layer for handling the composition monitoring and thus do not provide the important execution time violations feedback to the composition process. This year, we proposed the DISC framework which aims to provide a highly declarative event-oriented model to accommodate various aspects such as composition design and exceptions, data relationships and constraints, business calculations and decisions, compliance regulations, security or temporal requirements. Then, the same model is used for combining the control of the composition definition, its execution and the composition monitoring. We proposed a service oriented architecture with a flexible logic, including complex event patterns and choreographies, business oriented rules, and dynamic control of compositions. Advantages of this unified framework are the higher level of abstraction to design, execute, and reason upon a composition, the flexibility
of the approach, and the ability to easily include non-functional requirements such as temporal or security issues and we implement the DISC framework using the Discrete Event Calculus reasoner. Ehtesham Zahoor defended his thesis in November [ 3 ], and had presented the DISC framework at ICWS 2011 [ 20 ].

We also continued the previous work initiated within the Associate Team INRIA VanaWeb about the provisioning of Web services composition using constraints solvers. The approach consists in instantiating this abstract representation of a composite Web service by selecting the most appropriate concrete Web services. This instantiation is based on constraint programming techniques which allow matching Web services according to a given request. The proposal performs this instantiation in a distributed manner, i.e., the solvers for each service type are solving some constraints at one level, and they are forwarding the rest of the request (modified by the local solution) to the next services. When a service cannot provision part of the composition, a distributed backtrack mechanism enables to change previous solutions (i.e., provisions). A major interest of this approach is to preserve privacy: solutions are not sent to the whole composition, services know only the services to which they are connected, and parts of the request that are already solved are removed from the next requests.

6.1.3. Alignment between Business Process and Service Architecture

Participants: François Charoy, Karim Dahman, Claude Godart.

In the continuation of work done previously on change management during process execution, we are conducting work on the governance of change at the business level and on its implications at the architecture and infrastructure level of an information system. Last year was devoted to the definition of the transformation rules that allowed to go from a business model to an IT model, i.e. a transformation between model based on different paradigms. During this year, a great deal of effort has been done in order to extend our work on Business to IT alignment management. Our goal is still to maintain this alignment at the lowest possible cost when the business process are changing [ 9 ]. Further than that we are trying to describe and validated an engineering method to help designer to maintain this alignment [ 8 ].

6.1.4. Distributed Processes with Security Constraints

Participants: Olivier Perrin, Aymen Baouab, Claude Godart.

Distributed processes governance is a very important challenge. In the past, we proposed various approaches for dealing with distribution, particularly for computing a set of sub-processes that can be distributed and that are equivalent to a given process. However, we did not deal with non-functional requirements as the focus was only on control and data flows. In this work, we try to deal not only with functional requirements, but also with non-functional requirements, in particular the security aspects. With Aymen Baouab, we already proposed an event-based approach that is able to verify that choreographies are valid with respect to given constraints (security constraints for instance) [ 7 ].

6.1.5. A Crisis Management Process Model

Participants: François Charoy, Joern Franke.

As said before, crisis management has been a very fruitful domain to investigate new approaches in the domain of high value, human driven activity coordination in a multi organisational setting. Our work benefits from a large amount of use cases and detailed accounts of previous dramatic events to analyze requirements and confront our proposals. 2011 has been devoted to terminate the evaluation and the validation of the model that we have defined during the previous years. It has also been devoted to complete the work done in the previous years on the inter organisational dimension of the coordination management [ 11 ]. The entire contribution on crisis management, i.e. the model, the system and the evaluation are described in Joern Franke PhD document [ 2 ].

In order to try to leverage this work in a more information system oriented way, we have started some collaboration to confront our view on coordination with existing reference model for humanitarian operations[ 12 ]. We are currently looking for alternative financing vehicle in order to continue this work.
This work was conducted as a cooperation with SAP Research Sophia Antipolis and partially funded by a CIFRE Grant.

6.2. Distributed Collaborative Systems

Starting with Web2.0 era, the web became easily writable and changeable, and nowadays, it is getting more real-time. Rather than requiring that users or their software check a source periodically for updates, real-time web is a paradigm based on the principle of pushing information to users as soon as it is available. We are faced with an explosion of real-time social software (Twitter, Facebook, etc.). Even if many social software are currently available, most of them rely on collaborative systems with a centralized architecture or authority and consequently suffer of intrinsic problems of centralization: lack of fault tolerance, poor scalability, costly infrastructure, problems of privacy.

Distributed collaborative systems (DCS) ease and coordinate collaboration among multiple users who jointly fulfill common tasks over computer networks without the need of a central architecture or authority.

We continued our work on migrating DCS to pure peer-to-peer architectures. This year we focussed on the real-time aspect of the collaboration. We evaluated replication mechanisms suitable for real-time collaboration over peer-to-peer architectures.

Moreover, peer-to-peer collaborative systems need revisiting traditional security models that prevent users from accessing to data and granted rights are checked before access is allowed. These access control mechanisms are too strict and they do not scale well in a peer-to-peer architecture. We make the assumption that an effective collaboration should rely on a flexible optimistic security model based on trust. This year we proposed a new collaboration model based on contracts where we rely on an optimistic security model. Rather than adopting an a priori strict enforcement of security rules, in this optimistic solution, access is given first to data without control but with restrictions that are verified a posteriori.

6.2.1. Evaluation of algorithms for Peer-to-Peer Real-time collaboration

Participants: Mehdi Ahmed-Nacer, Claudia Ignat, Gérald Oster, Hyun-Gul Roh, Pascal Urso.

Nowadays, real-time collaborative editing systems such as Etherpad or Google Docs became very popular. The operational transformation (OT) approach is a traditional optimistic replication mechanism that was used for real-time collaboration. Recently, Commutative Replicated Data Types (CRDTs) were introduced as a new class of replication mechanisms whose concurrent operations are designed to be natively commutative. CRDTs, such as WOOT, Logoot, Treedoc, and RGAs, are expected to be substitutes of replication mechanisms in collaborative editing systems.

We investigated the suitability of CRDTs for realtime collaborative editing [6]. To reflect the tendency of decentralized collaboration, which can resist censorship, tolerate failures, and let users have control over documents, we collect editing logs from real-time peer-to-peer collaborations. We provided a theoretical evaluation as well as an experimental one by replaying the editing logs on various CRDTs and OT algorithms implemented in the same environment. We found out that CRDT algorithms initially designed for peer-to-peer asynchronous collaboration are suitable for real-time collaboration. Moreover, they outperform some representative operational transformation approaches that were well established for real-time collaboration in terms of generation time, remote integration time and space complexity.

6.2.2. Contract-based collaboration

Participants: Claudia Ignat, Hien Thi Thu Truong.

In the push-pull-clone collaborative editing model widely used in distributed version control systems users replicate shared data, modify it and redistribute modified versions of this data without the need of a central authority. However, in this model no usage restriction mechanism is proposed to control what users can do with the data after it has been released to them. We extended the push-pull-clone model with contracts that express usage restrictions and that are checked a posteriori by users when they receive the modified data [18], [25]. We proposed a merging algorithm that deals not only with modifications on data but also with contracts. A
log-auditing protocol [19] was used to detect users who do not respect contracts after they received data and to adjust user trust levels. The associated trust values can be computed by using any existing decentralised trust model. Our proposed contract-based model has been implemented and evaluated by using PeerSim simulator.

6.3. Interoperability and Enterprise Modeling

**Participants:** Nacer Boudjlida [contact], Khalid Benali.

In the continuation of our previous work on semantic-based and model-based solutions for interoperability, we applied and experienced a variety of semantic annotation types (structural, terminological and behavioral) in the frame of dynamic web services discovery and for competence management systems. In addition, we explored semantic graphs as a formal framework for competence description and management. Further, in order to ease semantic interoperability of heterogeneous competence management systems, we are defining a generic representation model that could serve as a shared ontology for these types of systems.
6. New Results

6.1. Model Driven and Aspect Oriented Design

6.1.1. Requirements Engineering

Participants: Olivier Barais, Benoit Baudry, Benoit Combemale, Maha Driss, Jean-Marc Jézéquel, Emmanuelle Rouillé, Nicolas Sannier, Didier Voitisék.

Model-driven engineering can have a huge impact on the early design and analysis of complex systems. We have investigated modeling for requirements engineering in three areas:

- We use executable metamodelling techniques developed in the team to capture formal relationships between regulatory requirements and accepted practices in systems engineering [47], [42].
- We propose an approach for facilitating Web service selection according to user requirements. These requirements specify the needed functionality and expected QoS, as well as the composability between each pair of services. The originality of our approach is embodied in the use of Formal Concept Analysis (FCA) and its extension Relational Concept Analysis (RCA) [33] [25].
- We have analyzed a real industrial software process to illustrate the need for bridging the gap between software processes and software development tools to automate the development tools configuration, deployment, integration and adaptation [46].

6.1.2. Dynamically adaptive interactive systems

Participants: Arnaud Blouin, Jean-Marc Jézéquel, Grégory Nain.

Combining Aspect-Oriented Modeling with Property-Based Reasoning to Improve User Interface Adaptation: in this work we combined aspect-oriented modeling with an interactive system architecture to support dynamic adaptation of interactions and user interfaces [28].

6.1.3. Dynamically adaptive component-based systems

Participants: François Fouquet, Olivier Barais, Viet-Hoa Nguyen, Noël Plouzeau.

Continuous Design to Achieve Intelligent Reflection in Distributed Systems: we defined an intelligent reflection model to support fast adaptation of distributed systems by architecture modification without stopping the system. This adaptation mechanism is well suited to rapidly changing needs (continuous design of eternal systems) or fast paced modifications of the context of the running system (for instance for Internet of Things distributed systems) [34].

6.1.4. Architecture for Services-based applications

Participants: Olivier Barais, Johann Bourcier, Erwan Daubert, Jean-Marc Jézéquel.

The architecture of service-based applications can have a huge impact on their dynamic adaptability. We have investigated various framework for architecting service-based applications:

- Designing SAFDIS: a self adaptive framework for distributed applications based on services. SAFDIS includes facilities to support the coordination of distributed reconfigurations [24]. SAFDIS also takes benefit of the Infrastructure As A Service to dynamically reconfigure Software As A Service [44].
- Analyzing and improving consistency between functional and business view of telecom services architecture. This work is based on the definition of a strategic alignment of the target functional view with the target business view. Alignment is validated with a real case study implemented and deployed at Orange–France Telecom on their messaging service [22].
• Designing AutoHome: a service oriented framework to simplify the development and runtime adaptive support of autonomic pervasive applications. This includes the amalgamation of the two computing areas of Autonomics and Service Orientation, to produce a Component-based platform providing facilities. This infrastructure uniquely blends the advantages of distributed autonomic control with global conflict management in a management hierarchy [17].

6.2. Model V&V and Testing

6.2.1. Formal MDE Foundations

Participants: Benoit Baudry, Benoit Combemale.

• Formally Tracing Executions From an Analysis Tool Back to a Domain Specific Modeling Language’s Operational Semantics: in this work, we propose a formal and operational framework for tracing results back (e.g., a program crash log, or a counterexample returned by a model checker) from execution and verification tools to an original DSML’s syntax and operational semantics [31].

• A Proof Assistant Based Formalization of components in MDE: using the Coq proof assistant we propose a formalization of some operators for model fragment extraction and composition, as defined in the ReuseWare toolset [39].

• We have developed a methodology to explicitly model the context in which a temporal property must be verified. This contextual information is expressed in the requirements, and an explicit model allows to reduce the complexity of automated verification [41].

6.2.2. Pairwise testing for highly variable systems

Participants: Benoit Baudry, Aymeric Hervieu.

Variability management is a major concern for the development of software intensive systems. In particular, the explosion of variants is an issue for testing and analysis. Feature models allow to explicitly capture the variability in a formal model and get a complete view on all possible variants of the system. We have investigated pair-wise generation from feature models in order to test software product lines [36], and to evaluate QoS contracts in variable web service compositions [38].

6.2.3. Testing aspect-oriented programs

Participant: Benoit Baudry.

Aspect-oriented mechanisms introduce new risks for reliability that must be tackled by specific testing techniques in order to fully benefit from the use of this paradigm. We have investigated a monitoring mechanism of advices in an aspect-oriented program and use this information to build test cases that target faults in pointcut descriptors [18].

6.2.4. Modeling model quality metrics

Participants: Benoit Baudry, Jean-Marc Jézéquel.

We have developed a model-driven measurement approach to measure models of a domain specific modeling language. The approach uses models as unique and consistent metric specifications for the automated generation of a metric tool. The benefit from applying the approach is evaluated by four case studies [20]. In particular, we have evaluated the ability of the approach to build a tool for the measurement of requirements documents [21].
6.3. Meta-Modeling

6.3.1. Model Driven Language Engineering

Participants: Benoit Baudry, Arnaud Blouin, Juan-Jose Cadavid Gomez, Benoit Combemale, Clément Guy, Jean-Marc Jézéquel, Didier Vojtisek.

- Model-Driven Engineering and Optimizing Compilers: A bridge too far? In this work, we report and analyze an experience about the use of MDE technologies to build and evolve compiler infrastructures in the optimizing compiler domain. From this study, we highlight challenges and propose a roadmap for the cross-fertilization of the MDE and compiler domains [35], [45].
- Modeling Model Slicers: model slicing is a model operation that consists in extracting a subset of a model. Because the creation of a new DSL implies the creation from scratch of a new model slicer, we proposed the Kompren language that models and generates model slicers for any DSL [27].
- Empirical Evaluation of the Conjunct Use of MOF and OCL: we evaluate in this work the conjunct usage of MOF (Meta-Object Facility) and OCL (Object Constraint Language) in the development of Domain-Specific Modeling Languages. We observe the state of practice to understand how experts use them and find patterns on its usage, in order to provide techniques to improve the experience [29].

6.3.2. Model Transformation and Composition

Participants: Olivier Barais, Benoit Baudry, Arnaud Blouin, Mickaël Clavreul, Benoit Combemale, Xavier Dolques, Jean-Marc Jézéquel.

- Model operations such as transformation and composition declare source metamodels that are usually larger than the set of concepts and relations actually used by the operation. We have proposed and validated a static operation analyzer to retrieve the metamodel footprint of the operation [37].
- Service-Oriented Architecture Modeling: Bridging the Gap between Structure and Behavior: In this approach, we propose to detect divergences among structural and behavioral models to support a semi-automatic process of synchronization between class diagrams and workflow models [30].
- The paper propose a technique for discovering matchings between two model elements modeling the same system, but being instances of different metamodels. This is achieved by using property names and models structure thanks to the adaptation of a schema matching technique named Anchor-PROMPT [32].
- Specifying and implementing UI Data Bindings with Active Operations: based on the concept of active operations, this work proposes a framework to bind models at runtime and more precisely to bind data and their possible representations [26].
- We propose a requirement-centric approach for Web service composition which allows: (i) modeling users’ requirements with the MAP formalism and specifying required services using an Intentional Service Model (ISM); (ii) discovering and selecting relevant Web services and high QoS services; and (iv) generating automatically BPEL coordination processes by applying the model transformation technique [19].
6. New Results

6.1. Structuring of Applications for Scalability

Participants: Sylvain Contassot-Vivier, Thomas Jost, Jens Gustedt, Soumeya Leila Hernane, Constantinos Makassikis, Stéphane Vialle.

6.1.1. Large Scale and Interactive Fine Grained Simulations

Our library parXXL allows the validation of a wide range of fine grained applications and problems. We were able to test the interactive simulation of PDEs in physics, see [5], on a large scale. Also, biologically inspired neural networks have been investigated using parXXL and the InterCell software suite. The InterCell suite and these applicative results have been presented in [29].

6.1.2. Large Scale Models and Algorithms for Random Structures

A realistic generation of graphs is crucial as an input for testing large scale algorithms, theoretical graph algorithms as well as network algorithms, e.g. platform generators.

Commonly used techniques for the random generation of graphs have two disadvantages, namely their lack of bias with respect to history of the evolution of the graph, and their incapability to produce families of graphs with non-vanishing prescribed clustering coefficient. In this work we propose a model for the genesis of graphs that tackles these two issues. When translated into random generation procedures it generalizes well-known procedures such as those of Erdős & Rényi and Barabási & Albert. When just seen as composition schemes for graphs they generalize the perfect elimination schemes of chordal graphs. The model iteratively adds so-called contexts that introduce an explicit dependency to the previous evolution of the graph. Thereby they reflect a historical bias during this evolution that goes beyond the simple degree constraint of preference edge attachment. Fixing certain simple statical quantities during the genesis leads to families of random graphs with a clustering coefficient that can be bounded away from zero.

A journal article describing intensive simulations of these models that confirm the theoretical results and that show the ability of that approach to model the properties of graphs from application domains has been published as [13].

6.1.3. Development environment for co-processing units

In the framework of the PhD thesis of Wilfried Kirschenmann, co-supervised by Stéphane Vialle (SUPELEC & AlGorille team) and Laurent Plagne (EDF SINETICS team), we have designed and implemented a unified framework based on generic programming to achieve a development environment adapted both to multi-core CPUs, multi-core CPUs with SSE units, and GPUs, for linear algebra applied to neutronic computations. Our framework is composed of two layers: (1) MTPS is a low-level layer hiding the real parallel architecture used, and (2) Legolas++ is a high-level layer allowing the application developer to rapidly implement linear algebra operations. The Legolas++ layer aims at decreasing the development time, while the MTPS layer aims at automatically generating very optimized code for the target architecture, thus leading to decreased execution times. Experimental performances of the MTPS layer appeared very good, the same source code achieved performances close to 100% of the theoretical ones, on any of the supported target architectures. Our strategy is to generate optimized data storage and data access code for each target architecture, not just different computing codes.

A new version of Legolas++ is under development, and a minimal version has been implemented in 2011. It is optimized to use the MTPS layer: source code is generic while an optimized code is automatically generated to efficiently use all SSE/AVX vector units of a multicore CPU. An article on that work is accepted in the post-proceedings of PARA 2010 and will be published at the end of 2011; the thesis of Wilfried Kirschenmann will be defended in early 2012.
6.1.4. Structuring algorithms for co-processing units

Since 2009, we have designed and experimented several algorithms and applications, in the fields of option pricing for financial computations, generic relaxation methods, and PDE solving applied to a 3D transport model simulating chemical species in shallow waters. We aim at designing a large range of algorithms for GPU cluster architectures, to develop a real knowledge about mixed coarse and fine grained parallel algorithms, and to accumulate practical experience about heterogeneous cluster programming.

Our PDE solver on GPU cluster has been designed in the context of a larger project on the study of asynchronism (see 3.1 and 6.1.5). The iterations of the asynchronous parallel algorithm runs faster, but it requires more iterations and a more complex detection of convergence, see Section 6.1.5 below. We measured both computing and energy performances of our PDE solver in order to track the best solution, in function of the problem size, the cluster size and the features of the cluster nodes. We are tracking the most efficient solution for each configuration. It can be based on a CPU or a GPU computing kernel, and on a synchronous or asynchronous parallel algorithm. Moreover, the fastest solution is not always the less energy consuming. Our recent results are introduced in [26], and in an article accepted in the post-proceedings of PARA 2010. In 2011 we improved our asynchronous implementation. However, the most asynchronous version has led to significantly more complex code (with an increased probability of remaining bugs) but to similar performances. At the opposite, we designed and implemented different convergence detection mechanisms in our asynchronous version, and some versions seem to achieve really better performances. Execution time and energy consumption performances have now to be measured again for many configurations. We aim to get new complete performance evaluation at the beginning of 2012. Then we will design an automatic selection of the right kernel and the right algorithm, and we will implement an auto-setting application function of a global instruction of the user (to achieve a fast run, or a low consumption run, or a compromise...).

At last, we have continued to design option pricers on clusters of GPUs, with Lokman Abbas-Turki (PhD student at University of Marne-la-Valée) and some colleagues from financial computing. In the past we developed some European option pricers, distributing independent Monte-Carlo computations on the nodes of a GPU cluster. In 2010 we succeeded to develop an American Option pricer on our GPU clusters, distributing strongly coupled Monte-Carlo computations. The Monte-Carlo trajectories depend on each others, and lead to many data transfers between CPUs and GPUs, and to many communications between cluster nodes. First results were encouraging, we achieve speedup and size up. In 2011 we optimized a major step of our algorithm, consisting in a 4D to 2D reduction on GPU. Performances have increased, and are significantly easier to achieve. The configuration tuning of the application, function of the problem size and the number of computing nodes, has been simplified. Again, we investigate both computing and energy performances of our developments, in order to compare interests of CPU clusters and GPU clusters considering execution speed and the exploitation cost of our solution.

6.1.5. Asynchronism

In the previous paragraph is mentioned a project including the study of sparse linear solvers on GPU. That project deals with the study of asynchronism in hierarchical and hybrid clusters mentioned in 3.1. In that context, we study the adaptation of asynchronous iterative algorithms on a cluster of GPUs for solving PDE problems. In our solver, the space is discretized by finite differences and all the derivatives are approximated by Euler equations. The inner computations of our PDE solver consist in solving linear equations (generally sparse). Thus, a linear solver is included in our solver. As this part is the most time consuming, to decrease the overall computation time it is essential to get a version that is as fast as possible. This is why we have decided to implement it on GPU, as discussed in the previous paragraph. Our parallel scheme uses the Multisplitting-Newton which is a more flexible kind of block decomposition. In particular, it allows for asynchronous iterations.

Our first experiments, conducted on an advection-diffusion problem, have shown very interesting results in terms of performances [8]. However, we investigate the possibility to insert periodic synchronous iterations inside the asynchronous scheme in order to improve the convergence detection delay. This is especially interesting on small/middle clusters with efficient networks.
Moreover, another aspect which is worth being studied is the full use of all the computational power present on each node, in particular the multiple cores, in conjunction with the GPU. This is still a work in progress.

### 6.1.6. New Control and Data Structures for Efficiently Overlapping Computations, Communications and I/O

With the thesis of Pierre-Nicolas Clauss we introduced the framework of ordered read-write locks, ORWL, see [3]. These are characterized by two main features: a strict FIFO policy for access and the attribution of access to lock-handles instead of processes or threads. These two properties allow applications to have a controlled pro-active access to resources and thereby to achieve a high degree of asynchronism between different tasks of the same application. For the case of iterative computations with many parallel tasks which access their resources in a cyclic pattern we provide a generic technique to implement them by means of ORWL. It was shown that the possible execution patterns for such a system correspond to a combinatorial lattice structure and that this lattice is finite iff the configuration contains a potential deadlock. In addition, we provide efficient algorithms: one that allows for a deadlock-free initialization of such a system and another one for the detection of deadlocks in an already initialized system.

We have developed a standalone distributed implementation of the API that is uniquely based on C and POSIX socket communications. Our goal is to simplify the usage of ORWL and to allow portability to a large variety of platforms. This implementation runs on different flavors of Linux and BSD, on different processor types Intel and ARM, and different compilers, gcc, clang, opencc and icc. An experimental evaluation of the performance is on its way. An engineering support from the local INRIA center has allowed to advance this implementation and to perform intensive benchmarks. The results have been presented in [28].

**Data Handover**, DHO, is a general purpose API that combines locking and mapping of data in a single interface. The access strategies are similar to ORWL, but locks and maps can also be hold only partially for a consecutive range of the data object. It is designed to ease the access to data for client code, by ensuring data consistency and efficiency at the same time.

In the thesis of Soumeya Hernane, we use the Grid Reality And Simulation (GRAS) environment of SimGrid, see 5.4, as a support for an implementation of DHO. GRAS has the advantage of allowing the execution in either the simulator or on a real platform. A first series of tests and benchmarks of that implementation demonstrates the ability of DHO to provide a robust and scalable framework, [18]. A step forward towards a distributed algorithm that allows distributed read-write locks with dynamic participation of process has been achieved in [30].

### 6.1.7. Energy performance measurement and optimization

Several experiments have been done on the GPU clusters of SUPÉLEC with different kinds of problems ranging from an embarrassingly parallel one to a strongly coupled one, via some intermediate levels. Our first results tend to confirm our first intuition that the GPUs are a good alternative to CPUs for problems which can be formulated in a SIMD or massively multi-threading way. However, when considering not embarrassingly parallel applications the supremacy of a GPU cluster tends to decrease when the number of nodes increases. This observation was the starting point of our participation to the COST-IC0804 about energy efficiency in large scale distributed systems, and an article accepted in the post-proceedings of PARA 2010 introduces our results achieved with our PDE solver distributed on our GPU clusters.

In 2011 we conducted new experiments and optimizations of our PDE solver and our American option pricer on the GPU clusters of SUPELEC. These experiments are still ongoing, and the optimization of this software should be achieved at the beginning of 2012. Simultaneously, we designed the foundations of a complete software architecture of self-configuring applications, choosing the right compute kernel and the right parallel algorithm to use, automatically. The global objectives to respect would either the overall speed, low energy consumption, or a speed-energy compromise. This global objective can be set by the user, or by an intelligent scheduler that aims to optimize a set of runs on a large cluster. This software architecture foundation has been introduced to a COST-IC0804 meeting in Budapest in June 2011.
In order to achieve this goal we need to establish some models of energy consumption of our applications on our CPU+GPU clusters, to be able to implement some heuristic of auto-setting. In 2011 we published a book chapter [26] introducing our first modeling strategies. Next step will be to implement a first auto-setting application, and to experiment its performances.

6.1.8. Load balancing

A load-balancing algorithm based on asynchronous diffusion with bounded delays has been designed to work on dynamical networks [34]. It is by nature iterative and we have provided a proof of its convergence in the context of load conservation. Also, we have given some constraints on the load migration ratios on the nodes in order to ensure the convergence. This work has been extended, especially with a detailed study of the imbalance of the system during the execution of a parallel algorithm simulated in the SimGrid platform.

The perspectives of that work are double. The first one concerns the internal functioning of our algorithm. There is an intrinsic parameter which tunes the load migration ratios and we would like to determine the optimal value of that ratio. The other aspect is on the application side in a real parallel environment. Indeed, we are currently applying this algorithm to a parallel version of the AdaBoost learning algorithm. This will allow us to study the best parameter to choose and to compare our load-balancing scheme to other existing ones.

Concerning the Neurad project, our parallel learning proceeds by decomposing the data-set to be learned. However, using a simple regular decomposition is not sufficient as the obtained sub-domains may have very different learning times. Thus, we have designed a first domain decomposition of the data set yielding sub-sets of similar learning times [40]. One of the main issue in this work has been the determination of the best estimator of the learning time of a sub-domain. As the learning time of a data set is directly linked to the complexity of the signal, several estimators taking into account that complexity have been tested, among which the entropy. However, the entropy is not the best estimator in that context, and we had to design a specific estimator. Also, we have optimized the decomposition process and added a selection phase that produces learning subsets of the same size [20]. Finally, we have also developed a parallel multi-threaded version of that decomposition/selection process.

6.1.9. Fault Tolerance

6.1.9.1. Application-level fault tolerance

Concerning fault tolerance, we have worked with Marc Sauget, from the University of Franche-Comté, on a parallel and robust algorithm for neural network learning in the context of the Neurad project [35]. A short description of that project is given in Section 4.1.5.

Our fault-tolerance strategy has shown to be rather efficient and robust in our different experiments performed with real data on a local cluster where faults were generated. Although those results are rather satisfying, we would like to investigate yet more reactive mechanisms as well as the insertion of robustness at the server level.

6.1.9.2. Programming model and frameworks for fault-tolerant applications

During the PhD thesis of Constantinos Makassikis [11], supervised by Stéphane Vialle, we have designed a new fault tolerance programming model (MoLoToF) to ease the development of fault-tolerant distributed applications. Main features of MoLoToF include so-called “fault-tolerant skeletons” to embed checkpoint-based fault tolerance within applications, and enable various collaborations, such as application-semantic knowledge supplied by users to the underlying system (e.g.: middleware), in order to fine tune fault tolerance.

Two development frameworks have been designed according to two different parallel programming paradigms: ToMaWork for Master-Workers applications [17] and FT-GReLoSS (FTG) for some kind of SPMD applications including inter-node communications [10]. The programmer’s task is limited. He only needs to supply some computing routines (functions of the application), has to add some extra code to specify a fault-tolerant parallel programming skeletons and then to tune the checkpointing frequency.
Our experiments have exhibited limited runtime overheads when no failure occurs and acceptable runtime overheads in the worst case failures. Observed runtime overheads are less than the ones obtained with all other system-level fault tolerance solutions we have experimented, while maintaining very limited development time overhead. Moreover, detailed experiments up to 256 nodes of our cluster have shown that it is possible to finely tune the checkpointing policies of the frameworks in order to implement different fault tolerance strategies, for example, according to cluster reliability.

In 2011, we have used the FTG framework to make fault-tolerant an existing parallel financial application [31] from EDF R&D, where it is used for gas storage valuation. The resulting application kept its initial runtime performance despite some source code modifications which are required in order to use FTG. As it was the case in earlier experiments with other applications, these modifications accounted for a limited development time overhead and fault tolerance remained more efficient than system-level fault tolerance solutions.

6.2. Experimentation Methodology

Participants: Tomasz Buchert, Sébastien Badia, Pierre-Nicolas Clauss, El Mehdi Fekari, Jens Gustedt, Lucas Nussbaum, Martin Quinson, Cristian Rosa, Luc Sarzyniec, Sylvain Contassot-Vivier.

6.2.1. Overall Improvements of the SimGrid Framework

See 4.2.1 for the scientific context of this result.

This year was the third year of the USS-SimGrid project on the simulation of distributed applications. We are principal investigator (see 8.2.7) of this project, funded by the ANR. It was prolonged until October 2012, giving us the ability to finish properly what was started. Several improvements have therefore been added to the framework, with numerous contributions from the ANR participants. This served as a flagship for the whole SimGrid project and hosted several of our research efforts, detailed in the subsequent sections (up to 6.2.5). Also this year, the SONGS project got accepted by the ANR, paving the road for our research in this context for the next four years. Our team also coordinates this project, devoted to the “Simulation Of Next Generation Systems” (see 8.2.7).

In addition, the software quality efforts were pursued further through the second year of the INRIA ADT project (see 8.2.1) to maximize the impact of our research on our user community. First, we improved further our automated regression tests (by increasing the test coverage from below 60% to almost 80%) and by fixing the bugs found through the automated builds conducted on the INRIA pipol infrastructure. We also reduced the amount of possible configurations to reduce the test and maintenance burden. As usual, performance tuning deserved a lot of our attention this year. The bindings were solidified and improved. They are very well received by the user community. Finally, the port to the Mac architecture was improved while the experimental port to Windows was revived.

Finally, several operations were conducted to increase our user community. A publication summarizing all improvements made in the recent years were written and submitted [27]. The SimGrid team was represented at SuperComputing’11 (through our partners of Lyon) to meet potential users and distribute informative leaflets designed and printed to that extend.

6.2.2. Formal Verification of Distributed Applications

The context of this work is presented in 4.2.2.

In 2011, we started using the model-checker integrated last year into the SimGrid framework with the goal to evaluate its limitations. Due to its generic design, it is able to verify protocols written using several APIs of SimGrid. We tested it on both a MPI toy program written to that extend and on an implementation of the Chord P2P protocol. In this later case, the tested program was not written for the purpose of being model-checked but to assess the scalability limits of the simulator. The model-checker was used to track down a bug that was near to impossible to find with the simulator alone. This experiment and the formalism underlying our model-checker were described in the publication [19]. It is also described in further detail in Cristian Rosa’s PhD, defended this year [12].
A second axis of our work this year consisted in extending the semantic power of the verified properties. In the work presented above, only local assertions and invariants can be verified. We started to investigate how to improve this during the internship of Marion Guthmuller. The major difficulty is that the reduction techniques based on the transition independences that we used so far are not sufficient for vivacity properties and must be extended to deal with the visibility of atomic properties [37]. One of the specificity of our work is the use of actual implementations were most of the literature uses handmade abstract models. This work continued in a PhD program, but didn’t led to any publication, yet.

6.2.3. Parallel Simulation within SimGrid

In addition to the software tuning and improvement described in 6.2.1, we tackled the issue of running SimGrid simulators in parallel. Our work differs from the state of the art, because we do not aim to parallelize the simulation kernel itself but the execution of the user code processes running on top of the simulated system. Interestingly enough, this benefits greatly from the work on formal verification introduced in the previous section, and particularly of the new network abstraction layer that was added. It greatly reduced the code locations where the global state is modified, making the parallel execution possible.

This allowed for example the simulation of up to 2 million Chord hosts on a single computer. This work was described in [33] and a publication in a major conference is under preparation. Since the available memory constitutes the main scalability limit now, we will work on distributing the simulation to leverage the memory of several computers at the same time.

6.2.4. Simulating MPI Applications

The final goal of SMPI is to simulate a C/C++/FORTRAN MPI program designed for a multi-processor system on a single computer without any source code modification. This addresses one of the main limitation of SimGrid, which requires the application to be written using one of the specific interfaces atop the simulator. New efforts have been put since July 2009 in this project, hereby continuing the work initiated by Henri Casanova and Mark Stilwell at University of Hawai’i at Manoa.

Previous work included a prototype implementation of various MPI primitives such as send, recv, isend, irecv and wait. Since the project’s revival, many of the collective operations (such as bcast, alltoall, reduce) have been implemented. The standard network model used in SimGrid has also been reworked to reach a higher precision in communication timings. Indeed, MPI programs are traditionally run on high performance computers such as clusters, and this requires to capture fine network details to correctly model the program behavior. Starting from the existing, validated network model of SimGrid, we have derived for SMPI a specific piece-wise linear model which closely fits real measurements. In particular, it enables to correctly model small messages and messages above the eager/rendezvous protocol limit. This work has been published at the IPDPS conference this year [15].

Ongoing work is now targeting a panel of MPI applications to have a better understanding of the applicability of our proposition. Pierre-Nicolas Clauss, who has been working full-time on the project between mid-2010 and mid-2011 has left, and we plan to put new workforce on SMPI with the support of the SONGS ANR project in 2012.

6.2.5. Simulating Real Applications

This work aims at providing a solution to simulate arbitrary applications on top of SimGrid. The approach consists in intercepting the application actions at system level while they are executed on a test platform, and then replay these actions on top of the simulator.

Concerning trace capture, we continued our work on the Simterpose software, which intercepts the actions of the application and save them to file for further use by the simulator. This work, presented in a national conference [22], will be continued during the PhD work of Marion Guthmuller.
Concerning trace replay, we proposed a replay mechanism specific to MPI applications in collaboration with F. Suter from the Computing Center at IN2P3 together with F. Desprez and G. Markomanolis from the Graal team at INRIA Rhônes-Alpes. The originality is to rely on time-independent execution traces. This allows to completely decouple the acquisition process from the actual replay of the traces in a simulation context. We are able to acquire traces for large application instances without being limited to an execution on a single cluster. Finally, our replay framework is built directly on top of the SIMGrid simulation kernel. This work was published in [16].

6.2.6. Emulation & Distem

During the internship of Luc Sarzyniec, we re-implemented an emulator from scratch with the goal of having a more reliable basis for further developments. This new development, Distem (see 5.2), already includes support for CPU performance emulation (internship of Tomasz Buchert in 2010) and network emulation. We are currently preparing a first release of Distem, and are working on its validation.

6.2.7. Grid’5000 and ADT Aladdin-G5K

Grid’5000 is an experimental platform for research on distributed systems. Two new sites were added to Grid’5000 in 2011: Reims and Luxembourg. This should reinforce the impact of Grid’5000 in the east of France. It is worth noting that the system administrator of the Luxembourg Grid’5000 site was formerly a student in Nancy, and did a student project using Grid’5000 managed by Lucas Nussbaum. Also, more collaboration on technical aspects is expected thanks to this geographical proximity.

On the local level, power consumption sensors are being added to the graphene cluster, which will allow an accurate monitoring of energy consumption during experiments.

On the national level, Lucas Nussbaum is now mandated by the Grid’5000 executive committee to follow the work of the technical team. He contributed to two publications [23], [24] at Journées Réseaux 2011 that describe the Grid’5000 software stack. He also gave invited talks during a Grid’5000 day at RenPar, and during the Support for experimental computer science workshop as SuperComputing’11.

Local scientific contributions include the automation of the deployment of the gLite middleware on Grid’5000. That work [21] was presented at Rencontres France Grilles and received the Best Poster award. We hope that this work will serve as a basis for further collaborations with the production grids community.

We also started the ADT Kadeploy project that will continue the development of the Kadeploy software, which already plays a key role on Grid’5000.

6.2.8. Experimental cluster of GPUs

The experimental platform of SUPÉLEC for "GPGPU", see Section 4.2.6, is composed of two GPU clusters, and its electrical line has been improved in 2011.

The first cluster is currently composed of 16 PCs, each one hosting a dual-core CPU and a GPU card: a nVIDIA GeForce GT285, with 1GB of RAM (on the GPU card). The 16 nodes are interconnected across a devoted Gigabit Ethernet switch. The second cluster has 16 more recent nodes, composed of an Intel Nehalem CPU with 4 hyper-threaded cores at 2.67GHz, and a nVIDIA GTX480 ("Fermi") GPU card with 1.5GB of memory. This cluster has a Gigabit Ethernet interconnection network too. These 2 clusters can be accessed and used like one 32-nodes heterogeneous cluster of hybrid nodes. This platform has allowed us to experiment different algorithms on an heterogeneous cluster of GPUs.

The energy consumption of each node of the cluster hosting the GTX285 GPUs is monitored by a Raritan DPXS20A-16 device that continuously measures the electric power consumption (in Watts). The nodes of the cluster hosting the GTX480 are monitored by two Raritan devices, because the energy consumed by this cluster exceeds the maximum energy supported by a Raritan DPXS20A-16 device.

A set of Perl and shell scripts, developed by our team, sample the electrical power (Watt) measured by the Raritan devices and compute the energy (Joule or Watt Hour) consumed by the computation on each node and on the complete cluster (including the interconnection switch).
In 2011 we have increased the amount of electrical energy supplied to these cluster, in order to support the experiments of our new distributed American option pricer. This application achieves high performances but consumes more energy on our GPU clusters than our previous codes, and exceeded the limit of our previous electrical line.

This platform has been intensively used to get experimental performance measures introduced in 2011 meetings of the COST IC0804 about *Energy efficiency in large scale distributed systems*, and published in a book chapter [26].
6. New Results

6.1. Models

6.1.1. Using the Last-mile Model as a Distributed Scheme for Available Bandwidth Prediction

Participants: Olivier Beaumont, Lionel Eyraud-Dubois, Young Won.

Several Network Coordinate Systems have been proposed to predict unknown network distances between a large number of Internet nodes by using only a small number of measurements. These systems focus on predicting latency, and they are not adapted to the prediction of available bandwidth. But end-to-end path available bandwidth is an important metric for the performance optimisation in many high throughput distributed applications, such as video streaming and file sharing networks. In [34], we propose to perform available bandwidth prediction with the last-mile model, in which each node is characterised by its incoming and outgoing capacities. This model has been used in several theoretical works for distributed applications. We design decentralised heuristics to compute the capacities of each node so as to minimise the prediction error. We show that our algorithms can achieve a competitive accuracy even with asymmetric and erroneous end-to-end measurement datasets. A comparison with existing models (Vivaldi, Sequoia, PathGuru, DMF) is provided. Simulation results also show that our heuristics can provide good quality predictions even when using a very small number of measurements.

6.1.2. Divisible Load Scheduling

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois.

Malleable tasks are jobs that can be scheduled with preemptions on a varying number of resources. In [31] we focus on the special case of work-preserving malleable tasks, for which the area of the allocated resources does not depend on the allocation and is equal to the sequential processing time. Moreover, we assume that the number of resources allocated to each task at each time instant is bounded. We consider both the clairvoyant and non-clairvoyant cases, and we focus on minimizing the weighted sum of completion times. In the weighted non-clairvoyant case, we propose an approximation algorithm whose ratio (2) is the same as in the unweighted non-clairvoyant case. In the clairvoyant case, we provide a normal form for the schedule of such malleable tasks, and prove that any valid schedule can be turned into this normal form, based only on the completion times of the tasks. We show that in these normal form schedules, the number of preemptions per task is bounded by 3 on average. At last, we analyze the performance of greedy schedules, and prove that optimal schedules are greedy for a special case of homogeneous instances. We conjecture that there exists an optimal greedy schedule for all instances, which would greatly simplify the study of this problem. Finally, we explore the complexity of the problem restricted to homogeneous instances, which is still open despite its very simple expression. (Join work with Loris Marchal from ENS Lyon)

6.1.3. Modeling and Practical Evaluation of a Service Location Problem in Large Scale Networks

Participants: Olivier Beaumont, Nicolas Bonichon, Hubert Larchevêque.
In [33], we consider a generalization of a classical optimization problem related to server and replica location problems in networks. More precisely, we suppose that a set of users distributed over a network wish to have access to a particular service proposed by a set of providers. The aim is then to distinguish a set of service providers able to offer a sufficient amount of resources in order to satisfy the requests of the clients. Moreover, a quality of service following some requirements in terms of latencies is desirable. A smart repartition of the servers in the network may also ensure good fault tolerance properties. We model this problem as a variant of Bin Packing, namely Bin Packing under Distance Constraint (BPDC) where the goal is to build a minimal number of bins (i.e. to choose a minimal number of servers) so that (i) each client is associated to exactly one server, (ii) the capacity of the server is large enough to satisfy the requests of its clients and (iii) the distance between two clients associated to the same server is minimized. We prove that this problem is hard to approximate even when using resource augmentation techniques; we compare the number of obtained bins when using polynomial time algorithms allowed to build bins of diameter at most $\beta d_{\text{max}}$, for $\beta > 1$, to the optimal number of bins of diameter at most $d_{\text{max}}$. On the one hand, we prove that (i) if $\beta = (2 - \epsilon)$, BPDC is hard to approximate within any constant approximation ratio, for any $\epsilon > 0$; and that (ii) BPDC is hard to approximate at a ratio lower than $\frac{3}{2}$ even if resource augmentation is used. On the other hand, if $\beta = 2$, we propose a polynomial time approximation algorithm for BPDC with approximation ratio $\frac{7}{3}$ in the general case. We show how to turn an approximation algorithm for BPDC into an approximation algorithm for the non-uniform capacitated $K$-center problem and vice-versa. Then, we present a comparison of the quality of results for BPDC in the context of several Internet latency embedding tools such as Sequoia and Vivaldi, using datasets based on PlanetLab latency measurements.

6.1.4. Use of Internet Embedding Tools for Heterogeneous Resources Aggregation

Participants: Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Hubert Larchevêque.

In [28], we are interested in large scale distributed platforms like BOINC, consisting of heterogeneous resources and using the Internet as underlying communication network. In this context, we study a resource clustering problem, where the goal is to build clusters having at least a given capacity and such that any two participants to the same cluster are not too far from each other. In this context, the distance between two participants corresponds to the latency of a communication between them. Our goal is to provide algorithms with provable approximation ratios. In such large scale networks, it is not realistic to assume that the whole latency matrix (that gives the latency between any two participants) is known, and we need to rely on embedding tools such as Vivaldi or Sequoia. These tools enable to work on compact descriptions and well described metric spaces in which the distance between two points can be obtained directly from a small amount of information available at each node. We present the Bin Covering under Distance Constraint problem (BCDC for short), and propose dedicated algorithms for this problem for each metric space induced by each of the embedding tools. Then, we propose a comparison of these algorithms based on actual latency measures, that enables to decide which algorithm/embedding tool pair offers in practice for realistic datasets the best balancing between distance prediction and approximation ratios for the resource clustering problem.

6.1.5. Broadcasting on Large Scale Heterogeneous Platforms with Connectivity Artifacts under the Bounded Multi-Port Model

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois, Przemyslaw Uznanski.

In [32], we consider the classical problem of broadcasting a large message at an optimal rate in a large scale distributed network. The main novelty of our approach is that we consider that the set of participating nodes can be split into two parts: "green" nodes that stay in the open-Internet and "red" nodes that lie behind firewalls or NATs. Two red nodes cannot communicate directly, but rather need to use a green node as a gateway for transmitting a message. In this context, we are interested in both maximizing the throughput (i.e. the rate at which nodes receive the message) and minimizing the degree at the participating nodes, i.e. the number of TCP connections they must handle simultaneously. We consider both cyclic and acyclic solutions for the flow graph. In the cyclic case, our main contributions are a closed form formula for the optimal cyclic throughput and the proof that the optimal solution may require arbitrarily large degrees. In the acyclic case, we propose an algorithm to achieve the optimal throughput with low degree. Then, we prove a worst case ratio between...
the optimal acyclic and cyclic throughput and show through simulations that this ratio is on average very close to 1, which makes acyclic solutions efficient both in terms of throughput and of number of connections.

6.2. Overlays and distributed algorithms

6.2.1. Locally Fair Graph Exploration Strategies

Participants: David Ilcinkas, Ralf Klasing, Adrian Kosowski.

In [16], we considered the problem of exploring an anonymous undirected graph using an oblivious robot. The studied exploration strategies are designed so that the next edge in the robot’s walk is chosen using only local information, and so that some local equity (fairness) criterion is satisfied for the adjacent undirected edges. Such strategies can be seen as an attempt to derandomize random walks, and are natural counterparts for undirected graphs of the rotor-router model for symmetric directed graphs. The first of the studied strategies, known as Oldest-First (OF), always chooses the neighboring edge for which the most time has elapsed since its last traversal. Unlike in the case of symmetric directed graphs, we show that such a strategy in some cases leads to exponential cover time. We then consider another strategy called Least-Used-First (LUF) which always uses adjacent edges which have been traversed the smallest number of times. We show that any Least-Used-First exploration covers a graph $G = (V,E)$ of diameter $D$ within time $O(D|E|)$, and in the long run traverses all edges of $G$ with the same frequency.

6.2.2. Black Hole Search in Directed Graphs

Participant: Adrian Kosowski.

In [21] we considered a team of agents which has to explore a graph $G$ where some nodes can be harmful. Robots are initially located at the so called home base node. The dangerous nodes are the so called black hole nodes, and once a robot enters in one of them, it is destroyed. The goal is to find a strategy in order to explore $G$ in such a way that the minimum number of robots is wasted. The exploration ends if there is at least one surviving robot which knows all the edges leading to the black holes. As many variations of the problem have been considered so far, the solution and its measure heavily depend on the initial knowledge and the capabilities of the robots. We assume that $G$ is a directed graph, the robots are associated with unique identifiers, they know the number of nodes $n$ of $G$ (or at least an upper bound on $n$), and they know the number of edges $\Delta$ leading to the black holes. Each node is associated with a white board where robots can read and write information in a mutual exclusive way. A recently posed question (Czyzowicz et al. 2009) is whether some number of robots, expressed as a function of parameter Delta only, is sufficient to detect black holes in directed graphs of arbitrarily large order $n$. We give a positive answer to this question for the synchronous case, i.e., when the robots share a common clock, showing that $O(\Delta^2 \Delta)$ robots are sufficient to solve the problem. This bound is nearly tight, since it is known that at least $2^\Delta$ robots are required for some instances. Quite surprisingly, we also show that unlike in the case of undirected graphs, for the directed version of the problem, synchronization can sometimes make a difference: for $\Delta = 2$, in the synchronous case 4 robots are always sufficient, whereas in the asynchronous case at least 5 robots are sometimes required.

6.2.3. Rendezvous for Location-Aware Agents

Participant: Adrian Kosowski.

In [35] we studied rendezvous of two anonymous agents, where each agent knows its own initial position in the environment. Their task is to meet each other as quickly as possible. The time of the rendezvous is measured by the number of synchronous rounds that agents need to use in the worst case in order to meet. In each round, an agent may make a simple move or it may stay motionless. We consider two types of environments, finite or infinite graphs and Euclidean spaces. A simple move traverses a single edge (in a graph) or at most a unit distance (in Euclidean space). The rendezvous consists in visiting by both agents the same point of the environment simultaneously (in the same round). In [35], we propose several asymptotically optimal rendezvous algorithms. In particular, we show that in the line and trees as well as in multi-dimensional Euclidean spaces and grids the agents can rendezvous in time $O(d)$, where $d$ is the distance between the initial
positions of the agents. The problem of location-aware rendezvous was studied before in the asynchronous model for Euclidean spaces and multi-dimensional grids, where the emphasis was on the length of the adopted rendezvous trajectory. We point out that, contrary to the asynchronous case, where the cost of rendezvous is dominated by the size of potentially large neighborhoods, the agents are able to meet in all graphs of at most \( n \) nodes in time almost linear in \( d \), namely, \( O(d \log^2 n) \). We also determine an infinite family of graphs in which synchronized rendezvous takes time \( \Omega(d) \).

6.2.4. Boundary Patrolling by Mobile Agents

**Participant:** Adrian Kosowski.

In the boundary patrolling problem, a set of \( k \) mobile agents are placed on the boundary of a simply connected planar object represented by a cycle of unit length. Each agent has its own predefined maximal speed, and is capable of moving around this boundary without exceeding its maximal speed. The agents are required to protect the boundary from an intruder which attempts to penetrate to the interior of the object through a point of the boundary, unknown to the agents. The intruder needs some time interval of length \( \tau \) to accomplish the intrusion. Will the intruder be able to penetrate into the object, or is there an algorithm allowing the agents to move perpetually along the boundary, so that no point of the boundary remains unprotected for a time period \( \tau \)? Such a problem may be solved by designing an algorithm which defines the motion of agents so as to minimize the idle time \( I \), i.e., the longest time interval during which any fixed boundary point remains unvisited by some agent, with the obvious goal of achieving \( I < \tau \). Depending on the type of the environment, this problem is known as either boundary patrolling or fence patrolling in the robotics literature. The most common heuristics adopted in the past include the cyclic strategy, where agents move in one direction around the cycle covering the environment, and the partition strategy, in which the environment is partitioned into sections patrolled separately by individual agents. We have obtained, to our knowledge, the first study of the fundamental problem of boundary patrolling by agents with distinct maximal speeds. In this scenario, we give special attention to the performance of the cyclic strategy and the partition strategy. In [ 36 ], we propose general bounds and methods for analyzing these strategies, obtaining exact results for cases with 2, 3, and 4 agents. We show that there are cases when the cyclic strategy is optimal, cases when the partition strategy is optimal and, perhaps more surprisingly, novel, alternative methods have to be used to achieve optimality.

6.2.5. Rendezvous in trees

**Participant:** Adrian Kosowski.

In the rendezvous problem in trees, 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. In [ 51 ] we have presented 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 unknown size \( n \) 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 \( \log \log n \) memory is necessary for rendezvous, even for agents starting simultaneously in a \( n \)-node line.

6.2.6. How many oblivious robots can explore a line

**Participant:** David Ilcinkas.

In [ 20 ] we consider the problem of exploring an anonymous line by a team of \( k \) identical, oblivious, asynchronous deterministic mobile robots that can view the environment but cannot communicate. We completely characterize sizes of teams of robots capable of exploring a \( n \)-node line. For \( k < n \), exploration by \( k \) robots turns out to be possible, if and only if either \( k = 3 \), or \( k \geq 5 \), or \( k = 4 \) and \( n \) is odd. For all values of \( k \) for which exploration is possible, we give an exploration algorithm. For all others, we prove an impossibility result.
6.2.7. Asynchronous deterministic rendezvous in bounded terrains

**Participant:** David Ilcinkas.

Two mobile agents (robots) have to meet in an a priori unknown bounded terrain modeled as a polygon, possibly with polygonal obstacles. Robots are modeled as points, and each of them is equipped with a compass. Compasses of robots may be incoherent. Robots construct their routes, but the actual walk of each robot is decided by the adversary that may, e.g., speed up or slow down the robot. In [18], we consider several scenarios, depending on three factors: (1) obstacles in the terrain are present, or not, (2) compasses of both robots agree, or not, (3) robots have or do not have a map of the terrain with their positions marked. The cost of a rendezvous algorithm is the worst-case sum of lengths of the robots’ trajectories until they meet. For each scenario we design a deterministic rendezvous algorithm and analyze its cost. We also prove lower bounds on the cost of any deterministic rendezvous algorithm in each case. For all scenarios these bounds are tight.

6.2.8. On the Power of Waiting when Exploring Public Transportation Systems

**Participants:** David Ilcinkas, Ahmed Mouhamadou Wade.

We study the problem of exploration by a mobile entity (agent) of a class of dynamic networks, namely the periodically-varying graphs (the PV-graphs, modeling public transportation systems, among others). These are defined by a set of carriers following infinitely their prescribed route along the stations of the network. Flocchini, Mans, and Santoro [58] (ISAAC 2009) studied this problem in the case when the agent must always travel on the carriers and thus cannot wait on a station. They described the necessary and sufficient conditions for the problem to be solvable and proved that the optimal number of steps (and thus of moves) to explore a \( n \)-node PV-graph of \( k \) carriers and maximal period \( p \) is in \( \Theta(k \cdot p^2) \) in the general case.

In [46], we study the impact of the ability to wait at the stations. We exhibit the necessary and sufficient conditions for the problem to be solvable in this context, and we prove that waiting at the stations allows the agent to reduce the worst-case optimal number of moves by a multiplicative factor of at least \( \Theta(p) \), while the time complexity is reduced to \( \Theta(n \cdot p) \). (In any connected PV-graph, we have \( n \leq k \cdot p \).) We also show some complementary optimal results in specific cases (same period for all carriers, highly connected PV-graphs). Finally this new ability allows the agent to completely map the PV-graph, in addition to just explore it.

6.2.9. The impact of edge deletions on the number of errors in networks

**Participants:** Christian Glacet, Nicolas Hanusse, David Ilcinkas.

In [41], we deal with an error model in distributed networks. For a target \( t \), every node is assumed to give an advice, i.e. to point to a neighbor that take closer to the destination. Any node giving a bad advice is called a liar. Starting from a situation without any liar, we study the impact of topology changes on the number of liars.

More precisely, we establish a relationship between the number of liars and the number of distance changes after one edge deletion. Whenever \( \ell \) deleted edges are chosen uniformly at random, for any graph with \( n \) nodes, \( m \) edges and diameter \( D \), we prove that the expected number of liars and distance changes is \( O\left(\frac{\ell^2 Dm}{n}\right) \) in the resulting graph. The result is tight for \( \ell = 1 \). For some specific topologies, we give more precise bounds.

6.2.10. Computations in interconnection networks with a shared whiteboard

**Participant:** Adrian Kosowski.

In [52], we study the computational power of graph-based models of distributed computing in which each node additionally has access to a global whiteboard. A node can read the contents of the whiteboard and, when activated, can write one message of \( O(\log n) \) bits on it. A message is only based on the local knowledge of the node and the current content of the whiteboard. When the protocol terminates, each node computes the output based on the final contents of the whiteboard in order to answer some question on the network’s topology. We propose a framework to formally define several scenarios modelling how nodes access the whiteboard, in a synchronous way or not. This extends the work of Becker et al. where nodes were imposed to create their messages only based on their local knowledge (i.e., with the whiteboard empty). We prove that the four
models studied have increasing power of computation: any problem that can be solved in the weakest one can be solved in the second, and so on. Moreover, we exhibit problems that separate models, i.e., that can be solved in one model but not in a weaker one. These problems are related to Maximal Independent Set and detection of cycles. Finally we investigate problems related to connectivity as the construction of spanning- or BFS-tree in our different models.

6.2.11. Network Verification

Participant: Ralf Klasing.

In [27], we address the problem of verifying the accuracy of a map of a network by making as few measurements as possible on the nodes of the network. In the past, this task has been formalized as an optimization problem that, given a graph $G = (V, E)$, asks for finding a minimum-size subset $Q$ of vertices of $G$ such that the information returned by the queries on $Q$ uniquely identifies $G$. Previously, two global query models have been studied. In [27], we propose a query model that uses only local knowledge about the network. Quite naturally, we assume that a query at a given node $q$ returns the associated routing table, namely a set of entries which provides, for each destination node, a corresponding (set of) first-hop node(s) along an underlying shortest path. First, we show that any network of $n$ nodes needs $\Omega(\log \log n)$ queries to be verified. Then, we prove that there is no $o(\log n)$ approximation algorithm for the problem, unless $P = NP$, even for graphs with diameter 2. On the positive side, we provide an $O(\log n)$-approximation algorithm to verify a network of diameter 2, and we give exact polynomial-time algorithms for paths, trees and cycles of even length.

6.3. Compact and distributed data structures

6.3.1. Query optimization in databases

Participants: Nicolas Hanusse, Sofian Maabout.

Datacubes are data structures designed to query optimization in databases. In [42] we provide some algorithmic solutions in a user centric setting: the request time is guaranteed and the amount of memory space is minimized.

6.3.2. Parallel computations of Borders

Participants: Nicolas Hanusse, Sofian Maabout.

Borders are fundamental building blocks in data mining. They are used to find frequent patterns, dependencies between attributes, ... In [43] we provide an algorithm that computes borders with a speedup of $p$ (under reasonable hypothesis) for $p$ cores.

6.3.3. Node-disjoint multipath spanners and their relationship with fault-tolerant spanners

Participants: Cyril Gavoille, Quentin Godfroy.

Motivated by multipath routing, we introduce a multi-connected variant of spanners. For that purpose we introduce in [40] the $p$-multipath cost between two nodes $u$ and $v$ as the minimum weight of a collection of $p$ internally vertex-disjoint paths between $u$ and $v$. Given a weighted graph $G$, a subgraph $H$ is a $p$-multipath $s$-spanner if for all $u, v$, the $p$-multipath cost between $u$ and $v$ in $H$ is at most $s$ times the $p$-multipath cost in $G$. The $s$ factor is called the stretch.

Building upon recent results on fault-tolerant spanners, we show how to build $p$-multipath spanners of constant stretch and of $O(n^{1+1/k})$ edges, for fixed parameters $p$ and $k$, $n$ being the number of nodes of the graph. Such spanners can be constructed by a distributed algorithm running in $O(k)$ rounds.

Additionally, we give an improved construction for the case $p = k = 2$. Our spanner $H$ has $O(n^{3/2})$ edges and the $p$-multipath cost in $H$ between any two node is at most twice the corresponding one in $G$ plus $O(W')$, $W$ being the maximum edge weight.
6.3.4. On approximate distance labels and routing schemes with affine stretch

Participant: Cyril Gavoille.

For every integral parameter $k > 1$, given an unweighted graph $G$, we construct in polynomial time in $[25]$, for each vertex $u$, a distance label $L(u)$ of size $O(n^{2/(2k - 1)})$. For any $u, v \in G$, given $L(u), L(v)$ we can return in time $O(k)$ an affine approximation $\tilde{d}(u, v)$ on the distance $d(u, v)$ between $u$ and $v$ in $G$ such that $d(u, v) \leq \tilde{d}(u, v) \leq (2k - 2)d(u, v) + 1$. Hence we say that our distance label scheme has affine stretch of $(2k - 2)d + 1$. For $k = 2$ our construction is comparable to the $O(n^{5/3})$ size, $2d + 1$ affine stretch of the distance oracle of Pătraşcu and Roditty (FOCS ’10), it incurs a $o(\log n)$ storage overhead while providing the benefits of a distance label.

For any $k > 1$, given a restriction of $o(n^{1+1/(k-1)})$ on the total size of the data structure, our construction provides distance labels with affine stretch of $(2k - 2)d + 1$ which is better than the stretch $(2k - 1)d$ scheme of Thorup and Zwick (J. ACM ’05).

Our second contribution is a compact routing scheme with poly-logarithmic addresses that provides affine stretch guarantees. With $O(n^{3/(3k-2)})$-bit routing tables we obtain affine stretch of $(4k - 6)d + 1$, for any $k > 1$.

Given a restriction of $o(n^{1/(k-1)})$ on the table size, our routing scheme provides affine stretch which is better than the stretch $(4k - 5)d$ routing scheme of Thorup and Zwick (SPAA ’01).

6.3.5. Sparse spanners vs. compact routing

Participant: Cyril Gavoille.

Routing with multiplicative stretch 3 (which means that the path used by the routing scheme can be up to three times longer than a shortest path) can be done with routing tables of $\Theta(\sqrt{n})$ bits$^3$ per node. The space lower bound is due to the existence of dense graphs with large girth. Dense graphs can be sparsified to subgraphs, called spanners, with various stretch guarantees. There are spanners with additive stretch guarantees (some even have constant additive stretch) but only very few additive routing schemes are known.

In $[39]$, we give reasons why routing in unweighted graphs with additive stretch is difficult in the form of space lower bounds for general graphs and for planar graphs. We prove that any routing scheme using routing tables of size $M$ bits per node and addresses of poly-logarithmic length has additive stretch $\Omega(\sqrt{n/M})$ for general graphs, and $\Omega(\sqrt{n}/M)$ for planar graphs. Routing with tables of size $O(n^{1/3})$ thus requires a polynomial additive stretch of $\Omega(n^{1/3})$, whereas spanners with average degree $O(n^{1/3})$ and constant additive stretch exist for all graphs. Spanners, however sparse they are, do not tell us how to route. These bounds provide the first separation of sparse spanner problems and compact routing problems.

On the positive side, we give an almost tight upper bound: we present the first non-trivial compact routing scheme with $o(\log^2 n)$-bit addresses, additive stretch $O(n^{1/3})$, and table size $O(n^{1/3})$ bits for planar graphs. Actually, our scheme applies to general graphs, and the above bound of $n^{1/3}$ on the stretch and the routing table size holds for all graphs of linear local tree-width, a class of graphs including bounded-genus graphs and apex-minor-free graphs.

$^3$Tilde-big-$O$ notation is similar to big-$O$ notation up to factors poly-logarithmic in $n$. 


6. New Results

6.1. Scheduling Strategies and Algorithm Design for Heterogeneous Platforms


6.1.1. Virtual Machine Resource Allocation for Service Hosting on Heterogeneous Distributed Platforms

We proposed algorithms for allocating multiple resources to competing services running in virtual machines on heterogeneous distributed platforms. We developed a theoretical problem formulation, designed algorithms, and compared these algorithms via simulation experiments based in part on workload data supplied by Google. Our main finding is that vector packing approaches proposed in the homogeneous case can be extended to provide high-quality solutions in the heterogeneous case, and combined to provide a single efficient algorithm. We also considered the case when there may be errors in estimates of performance-related resource needs. We provided a resource sharing algorithm and proved that for the single-resource, single-node case, when there is no bound on the error, its performance ratio relative to an omniscient optimal algorithm is \( \frac{2J}{J+1} \), where \( J \) is the number of services. We also provided a heuristic approach for compensating for bounded errors in resource need estimates that performs well in simulation.

6.1.2. Dynamic Fractional Resource Scheduling vs. Batch Scheduling

We finalized this work in which we proposed a novel job scheduling approach for homogeneous cluster computing platforms. Its key feature is the use of virtual machine technology to share fractional node resources in a precise and controlled manner. Other VM-based scheduling approaches have focused primarily on technical issues or extensions to existing batch scheduling systems, while we take a more aggressive approach and seek to find heuristics that maximize an objective metric correlated with job performance. We derived absolute performance bounds and developed algorithms for the online, non-clairvoyant version of our scheduling problem. We further evaluated these algorithms in simulation against both synthetic and real-world HPC workloads and compared our algorithms to standard batch scheduling approaches. We found that our approach improves over batch scheduling by orders of magnitude in terms of job stretch, while leading to comparable or better resource utilization. Our results demonstrated that virtualization technology coupled with lightweight online scheduling strategies can afford dramatic improvements in performance for executing HPC workloads.

6.1.3. Greedy algorithms for energy minimization

This year, we have revisited the well-known greedy algorithm for scheduling independent jobs on parallel processors, with the objective of energy minimization. We have assessed the performance of the online version, as well as the performance of the offline version, which sorts the jobs by non-increasing size before execution. We have derived new approximation factors, as well as examples that show that these factors cannot be improved, thereby completely characterizing the performance of the algorithms.

6.1.4. Energy-aware mappings on chip multiprocessors

This year, in collaboration with Rami Melhem at Pittsburgh University (USA), we have studied the problem of mapping streaming applications that can be modeled by a series-parallel graph, onto a 2-dimensional tiled CMP architecture. The objective of the mapping is to minimize the energy consumption, using dynamic and voltage scaling techniques, while maintaining a given level of performance, reflected by the rate of processing the data streams. This mapping problem turned out to be NP-hard, but we identified simpler instances, whose
optimal solution can be computed by a dynamic programming algorithm in polynomial time. Several heuristics were proposed to tackle the general problem, building upon the theoretical results. Finally, we assessed the performance of the heuristics through comprehensive simulations using the StreamIt workflow suite and various CMP grid sizes.

We are pursuing this work by investigating the routing of communications in chip multiprocessors (CMPs). The goal is to find a valid routing in the sense that the amount of data routed between two neighboring cores does not exceed the maximum link bandwidth while the power dissipated by communications is minimized. Our position is at the system level: we assume that several applications, described as task graphs, are executed on a CMP, and each task is already mapped to a core. Therefore, we consider a set of communications that have to be routed between the cores of the CMP. We consider a classical model, where the power consumed by a communication link is the sum of a static part and a dynamic part, with the dynamic part depending on the frequency of the link. This frequency is scalable and it is proportional to the throughput of the link. The most natural and widely used algorithm to handle all these communications is XY routing: for each communication, data is first forwarded horizontally, and then vertically, from source to destination. However, if it is allowed to use all Manhattan paths between the source and the destination, the consumed power can be reduced dramatically. Moreover, some solutions may be found while none existed with the XY routing. We have compared XY routing and Manhattan routing, both from a theoretical and from a practical point of view. We considered two variants of Manhattan routing: in single-path routing, only one path can be used for each communication, while multi-paths routing allows to split a communication between different routes. We established the NP-completeness of the problem of finding a Manhattan routing that minimizes the dissipated power, we exhibited the minimum upper bound of the ratio power consumed by an XY routing over power consumed by a Manhattan routing, and finally we performed simulations to assess the performance of Manhattan routing heuristics that we designed.

6.1.5. Power-aware replica placement

We have investigated optimal strategies to place replicas in tree networks, with the double objective to minimize the total cost of the servers, and/or to optimize power consumption. The client requests are known beforehand, and some servers are assumed to pre-exist in the tree. Without power consumption constraints, the total cost is an arbitrary function of the number of existing servers that are reused, and of the number of new servers. Whenever creating and operating a new server has higher cost than reusing an existing one (which is a very natural assumption), cost optimal strategies have to trade-off between reusing resources and load-balancing requests on new servers. We provide an optimal dynamic programming algorithm that returns the optimal cost, thereby extending known results without pre-existing servers. With power consumption constraints, we assume that servers operate under a set of \( M \) different modes depending upon the number of requests that they have to process. In practice \( M \) is a small number, typically 2 or 3, depending upon the number of allowed voltages. Power consumption includes a static part, proportional to the total number of servers, and a dynamic part, proportional to a constant exponent of the server mode, which depends upon the model for power. The cost function becomes a more complicated function that takes into account reuse and creation as before, but also upgrading or downgrading an existing server from one mode to another. We have shown that with an arbitrary number of modes, the power minimization problem is NP-complete, even without cost constraint, and without static power. Still, we have provided an optimal dynamic programming algorithm that returns the minimal power, given a threshold value on the total cost; it has exponential complexity in the number of modes \( M \), and its practical usefulness is limited to small values of \( M \). Still, experiments conducted with this algorithm showed that it can process large trees in reasonable time, despite its worst-case complexity.

6.1.6. Reclaiming the energy of a schedule

In this work, we consider a task graph to be executed on a set of processors. We assume that the mapping is given, say by an ordered list of tasks to execute on each processor, and we aim at optimizing the energy consumption while enforcing a prescribed bound on the execution time. While it is not possible to change the allocation of a task, it is possible to change its speed. Rather than using a local approach such as backfilling, we have considered the problem as a whole and studied the impact of several speed variation models on its
complexity. For continuous speeds, we gave a closed-form formula for trees and series-parallel graphs, and we cast the problem into a geometric programming problem for general directed acyclic graphs. We showed that the classical dynamic voltage and frequency scaling (DVFS) model with discrete modes leads to a NP-complete problem, even if the modes are regularly distributed (an important particular case in practice, which we analyzed as the incremental model). On the contrary, the VDD-hopping model leads to a polynomial solution. Finally, we provided an approximation algorithm for the incremental model, which we extended for the general DVFS model.

6.1.7. Workload balancing and throughput optimization

We have investigated the problem of optimizing the throughput of streaming applications for heterogeneous platforms subject to failures. The applications are linear graphs of tasks (pipelines), and a type is associated to each task. The challenge is to map tasks onto the machines of a target platform, but machines must be specialized to process only one task type, in order to avoid costly context or setup changes. The objective is to maximize the throughput, i.e., the rate at which jobs can be processed when accounting for failures. For identical machines, we have proved that an optimal solution can be computed in polynomial time. However, the problem becomes NP-hard when two machines can compute the same task type at different speeds. Several polynomial time heuristics have been designed, and simulation results have demonstrated their efficiency.

6.1.8. Comparing archival policies for BlueWaters

In this work, we focus on the archive system which will be used in the BlueWaters supercomputer. We have introduced two new tape archival policies that can improve tape archive performance in certain regimes, compared to the classical RAIT (Redundant Array of Independent Tapes) policy. The first policy, PARALLEL, still requires as many parallel tape drives as RAIT but pre-computes large data stripes that are written contiguously on tapes to increase write/read performance. The second policy, VERTICAL, writes contiguous data into a single tape, while updating error correcting information on the fly and delaying its archival until enough data has been archived. This second approach reduces the number of tape drives used for every user request to one. The performance of the three RAIT, PARALLEL and VERTICAL policies have been assessed through extensive simulations, using a hardware configuration and a distribution of I/O requests similar to these expected on the BlueWaters system. These simulations have shown that VERTICAL is the most suitable policy for small files, whereas PARALLEL must be used for files larger than 1 GB. We have also demonstrated that RAIT never outperforms both proposed policies, and that a heterogeneous policy mixing VERTICAL and PARALLEL performs 10 times better than any other policy.

6.1.9. Using Virtualization and Job Folding for Batch Scheduling

In this work we study the problem of batch scheduling within a homogeneous cluster. In this context, the problem is that the more processors the job requires the more difficult it is to find an idle slot to run it on. As a consequence the resources are often inefficiently used as some of them remain unallocated in the final schedule. To address this issue we propose a technique called job folding that uses virtualization to reduce the number of processors allocated to a parallel job and thus allows to execute it earlier. Our goal is to optimize the resource use. We propose several heuristics based on job folding and we compare their performance with classical on-line scheduling algorithms as FCFS or backfilling. The contributions of this work are both the design of the job folding algorithms and their performance analysis.

6.1.10. A Genetic Algorithm with Communication Costs to Schedule Workflows on a SOA-Grid

We propose in this work a genetic algorithm to solve the problem of scheduling a collection of workflows, identical or not, on a SOA (Service Oriented Architecture) grid. A workflow (job) is represented by a directed acyclic graph (DAG) with typed tasks. All of the grid hosts are able to process a set of typed tasks with unrelated processing costs and are able to transmit files through communication links for which the communication times are not negligible. The goal of our study is to minimize the maximum completion time (makespan) of the workflows. To solve this problem we propose a genetic approach. The contributions of this paper are both the design of a Genetic Algorithm taking the communication costs into account and its performance analysis.
6.1.11. Checkpointing policies for post-petascale supercomputers

In this work, we provided an analysis of checkpointing strategies for minimizing expected job execution times in an environment that is subject to processor failures. In the case of both sequential and parallel jobs, we gave the optimal solution for exponentially distributed failure inter-arrival times, which, to the best of our knowledge, is the first rigorous proof that periodic checkpointing is optimal. For non-exponentially distributed failures, we developed a dynamic programming algorithm to maximize the amount of work completed before the next failure, which provides a good heuristic for minimizing the expected execution time. Our work considers various models of job parallelism and of parallel checkpointing overhead. We first performed extensive simulation experiments assuming that failures follow Exponential or Weibull distributions, the latter being more representative of real-world systems. The obtained results not only corroborate our theoretical findings, but also show that our dynamic programming algorithm significantly outperforms previously proposed solutions in the case of Weibull failures. We then performed simulation experiments that use failure logs from production clusters. These results confirmed that our dynamic programming algorithm significantly outperforms existing solutions for real-world clusters.

We have also showed an unexpected result: in some cases, when (i) the platform is sufficiently large, and (ii) the checkpointing costs are sufficiently expensive, or the failures are frequent enough, then one should limit the application parallelism and duplicate tasks, rather than fully parallelize the application on the whole platform. In other words, the expectation of the job duration is smaller with fewer processors! To establish this result we have derived and analyzed several scheduling heuristics.

6.1.12. Scheduling parallel iterative applications on volatile resources

In this work we study the efficient execution of iterative applications onto volatile resources. We studied a master-worker scheduling scheme that trades-off between the speed and the (expected) reliability and availability of enrolled workers. A key feature of this approach is that it uses a realistic communication model that bounds the capacity of the master to serve the workers, which requires the design of sophisticated resource selection strategies. The contribution of this work is twofold. On the theoretical side, we assess the complexity of the problem in its off-line version, i.e., when processor availability behaviors are known in advance. Even with this knowledge, the problem is NP-hard. On the pragmatic side, we proposed several on-line heuristics that were evaluated in simulation while a Markovian model of processor availabilities.

We have started this study with the simple case of iterations composed of independent tasks that can execute asynchronously. Then we have investigated a much more challenging scenario, that of a tightly-coupled application whose tasks steadily communicate throughout the iteration. In this latter scenario, if one processor computing some task fails, all the work executed for current iteration is lost, and the computation of all tasks has to be restarted. Similarly, if one processor of the current configuration is preempted, the computation of all tasks is interrupted. Changing the configuration within an iteration becomes a much riskier decision than with independent tasks.

6.1.13. Tiled QR factorization algorithms

In this work, we have revisited existing algorithms for the QR factorization of rectangular matrices composed of $p \times q$ tiles, where $p \geq q$. We target a shared-memory multi-core processor. Within this framework, we study the critical paths and performance of algorithms such as FIBONACCI and GREEDY, and those found within PLASMA. Although neither is optimal, both are shown to be asymptotically optimal for all matrices of size $p = q^2 f(q)$, where $f$ is any function such that $\lim_{q \to \infty} f(q) = 0$. This novel and important complexity result applies to all matrices where $p$ and $q$ are proportional, $p = \lambda q$, with $\lambda \geq 1$, thereby encompassing many important situations in practice (least squares). We provide an extensive set of experiments that show the superiority of the new algorithms for tall matrices.

We have then extended this work to a distributed-memory environment, that corresponds to clusters of multi-core processors. These platforms make the present and the foreseeable future of high-performance computing. In the context of a cluster of multicores, in order to minimize the number of inter-processor communications (aka, “communication-avoiding” algorithm), it is natural to consider two-level hierarchical reduction trees.
composed of an “inter-node” tree which acts on top of “intra-node” trees. At the intra-node level, we propose a hierarchical tree made of three levels: (0) “TS level” for cache-friendliness, (1) “low level” for decoupled highly parallel inter-node reductions, (2) “coupling level” to efficiently resolve interactions between local reductions and global reductions. Our hierarchical algorithm and its implementation are flexible and modular, and can accommodate several kernel types, different distribution layouts, and a variety of reduction trees at all levels, both inter-cluster and intra-cluster. Numerical experiments on a cluster of multicore nodes confirm that each of the four levels of our hierarchical tree contributes to build up performance and build insights on how these levels influence performance and interact within each other. Our implementation of the new algorithm with the Dague scheduling tool significantly outperforms currently available QR factorization softwares for all matrix shapes, thereby bringing a new advance in numerical linear algebra for petascale and exascale platforms.

6.1.14. Scheduling malleable tasks and minimizing total weighted flow

Malleable tasks are jobs that can be scheduled with preemptions on a varying number of resources. In this work, we have focused on the special case of work-preserving malleable tasks, for which the area of the allocated resources does not depend on the allocation and is equal to the sequential processing time. Moreover, we have assumed that the number of resources allocated to each task at each time instant is bounded. Although this study concerns malleable task scheduling, we have shown that this is equivalent to the problem of minimizing the makespan of independent tasks distributed among processors, when the data corresponding to tasks is sent using network flows sharing the same bandwidth.

We have considered both the clairvoyant and non-clairvoyant cases, and we have focused on minimizing the weighted sum of completion times. In the weighted non-clairvoyant case, we have proposed an approximation algorithm whose ratio is the same as in the unweighted non-clairvoyant case. In the clairvoyant case, we have provided a normal form for the schedule of such malleable tasks, and proved that any valid schedule can be turned into this normal form, based only on the completion times of the tasks. We have shown that in these normal form schedules, the number of preemptions per task is bounded by 3 on average. At last, we have analyzed the performance of greedy schedules, and proved that optimal schedules are greedy for a special case of homogeneous instances. We conjecture that there exists an optimal greedy schedule for all instances, which would greatly simplify the study of this problem.

6.1.15. Parallelizing the construction of the ProDom database

ProDom is a protein domain family database automatically built from a comprehensive analysis of all known protein sequences. ProDom development is headed by Daniel Kahn (INRIA project-team BAMBOO, formerly HELIX). With the protein sequence databases increasing in size at an exponential pace, the parallelization of MkDom2, the algorithm used to build ProDom, has become mandatory (the original sequential version of MkDom2 took 15 months to build the 2006 version of ProDom).

When protein domain families and protein families are built independently, the result may be inconsistent. In order to solve this inconsistency problem, we designed a new algorithm, MPI_MkDom3, that simultaneously builds a clustering in protein domain families and one in protein families. This algorithm mixes the principles of MP_MkDom2 and that of the building of Hogenom. As a proof of concept, we successfully processed all the sequences included in the April 2010 version of the UniProt database, namely 6 118 869 sequences and 2 194 382 846 amino-acids.

6.2. Algorithms and Software Architectures for Service Oriented Platforms

Participants: Daniel Balouek, Nicolas Bard, Julien Bigot, Yves Caniou, Eddy Caron, Florent Chuffart, Simon Delamare, Frédéric Desprez, Gilles Fedak, Sylvain Gault, Haiwu He, Cristian Klein, Georges Markomanolis, Adrian Muresan, Christian Pérez, Vincent Pichon, Jonathan Rouzaud-Cornabas, Anthony Simonet, José Saray, Bing Tang.
6.2.1. Parallel constraint-based local search

Constraint Programming emerged in the late 1980’s as a successful paradigm to tackle complex combinatorial problems in a declarative manner. It is somehow at the crossroads of combinatorial optimization, constraint satisfaction problems (CSP), declarative programming language and SAT problems (boolean constraint solvers and verification tools). Up to now, the only parallel method to solve optimization problems being deployed at large scale is the classical branch and bound, because it does not require much information to be communicated between parallel processes (basically: the current bound).

Adaptive Search was proposed by [86], [87] as a generic, domain-independent constraint-based local search method. This meta-heuristic takes advantage of the structure of the problem in terms of constraints and variables and can guide the search more precisely than a single global cost function to optimize, such as for instance the number of violated constraints. A parallelization of this algorithm based on threads realized on IBM BladeCenter with 16 Cell/BE cores show nearly ideal linear speed-ups for a variety of classical CSP benchmarks (magic squares, all-interval series, perfect square packing, etc.).

We parallelized the algorithm using the multi-start approach and realized experiments on the HA8000 machine, an Hitachi supercomputer with a maximum of nearly 16000 cores installed at University of Tokyo, and on the Grid’5000 infrastructure, the French national Grid for the research, which contains 8612 cores deployed on 11 sites distributed in France. Results show that speedups may surprisingly be architecture and problem dependant. Work in progress considers communications between each computing resource, and a new problem (costa) has been tested for its capability to have an exponential distribution of its time to complete on a sequential resolution.

6.2.2. Service Discovery in Peer-to-Peer environments

In 2010 we experimentally validated the scalability of the Spades Based Middleware (SBAM). SBAM is an auto-stabilized P2P middleware designed for the service discovery. The context of this development is the ANR SPADES project (see Section 7.2.2). In 2011, we wanted to guaranty truthfulness of information exchanged between SBAM-agents. In this context, the implementation of an efficient mechanism ensuring quality of large scale service discovery became a challenge. In collaboration with LIP6 team we developed a self stabilized model called CoPIF and we implemented it in SBAM using synchronous message exchange between agents. Indeed, when a node has to read its neighbor states, it sends a message to each and wait all response. Despite the fact that this kind of implementation is expensive, especially on a large distributed data structure, experiment shown that our model implementation stay efficient, even on a huge prefix tree. We use this broadcast mechanism not only to check the truthfulness of the distributed data structure but also to propagate activation of services on the entire SPADES platform. For the end of 2011 and the beginning of 2012 we plan to work on experimental evaluation of a self-stabilization inspired fault tolerance mechanism. We do this through a collaboration with Myriads team at Rennes.

Moreover, in the occasion of demonstration session of IEEE P2P’2011, we introduced the feasibility of multisite resources aggregation, thanks to SBAM, we ran SBAM on up to 200 peers (we generated machine volatility in order to show the self-stabilization) on 50 physical nodes of Grid’5000 to demonstrate the scalability of multi sites, self-stabilization good performance of our P2P middleware SBAM.

6.2.3. Décrypthon

In 2011, The DIET WebBoard (a web interface to manage the Décrypthon Grid through the DIET middleware) only received bugfixes and a few new features: the possibility to use a totally customized command to call the DIET client, improved support for multiprocessor tasks, and a basic support for replication of tasks (possibility to launch “clones” of an important task, in order to increase the probability of having a successful result). We deployed the new versions of the DIET Webboard on the Décrypthon university grid whenever we made changes to it.

In 2011, we started to port the Rhénovia application (a neuron simulation program in Java and python) on the Décrypthon grid.
The “Help cure muscular dystrophy, phase 2” program that we submitted to the world community grid was still in progress, we received large amounts of result files every day. We had to do the sorting of these files, checking, compressing and moving them to a long term storage space on a regular basis. We also made statistics for the internet users: http://graal.ens-lyon.fr/~nbard/WCGStats/. The last update was on 2011 June 27th: 76.67%.

6.2.4. Scheduling Applications with a Complex Structure

Non-predictably evolving applications are applications that change their resource requirements during execution. These applications exist, for example, as a result of using adaptive numeric methods, such as adaptive mesh refinement and adaptive particle methods. Increasing interest is being shown to have such applications acquire resources on the fly. However, current HPC Resource Management Systems (RMSs) only allow a static allocation of resources, which cannot be changed after it started. Therefore, non-predictably evolving applications cannot make efficient use of HPC resources, being forced to make an allocation based on their maximum expected requirements.

In 2011, we have revisited CooRM, an RMS targeting moldable application, and extended it to CooRMv2, an RMS which supports efficient scheduling of non-predictably evolving applications. An application can make “pre-allocations” to specify its peak resource usage. The application can then dynamically allocate resources as long as the pre-allocation is not outgrown. Resources which are pre-allocated but not used, can be filled by other applications. Results show that the approach is feasible and leads to a more efficient resource usage while guaranteeing that resource allocations are always satisfied.

As future work, we plan to extend CooRMv2 for non-homogeneous clusters, for example, for supercomputers that feature a non-homogeneous network. Moreover, we would like to apply the concepts proposed by CooRMv2 to large scale resource managers such as XtremOS.

6.2.5. High Level Component Model

Most software component models focus on the reuse of existing pieces of code called primitive components. There are however many other elements that can be reused in component-based applications. Partial assemblies of components, well defined interactions between components and existing composition patterns (a.k.a. software skeletons) are examples of such reusable elements. It turns out that such elements of reuse are important for parallel and distributed applications. Therefore, we have designed High Level Component Model (HLCM), a software component model that supports the reuse of these elements thanks to the concepts of hierarchy, genericity and connectors—and in particular the novel concepts of open connection.

In 2011, we have developed two specific implementations of HLCM: L2C for for C++, MPI and CORBA based applications and GLUON++ for CHARM++ based applications in collaboration with Prof. Kale’s team at the University of Illinois at Urbana-Champaign. L2C was used to study how HLCM may simplify the development of domain decomposition applications. GLUON++ was in particular used to study the performance portability of FFT library on various kind of machines. Moreover, on going work includes the study of the benefit of HLCM for MapReduce applications.

6.2.6. Simplifying Code-Coupling in the SALOME platform

The SALOME platform is a generic platform for pre- and post-processing for numerical simulations. It is made of modules which are themselves a set of components. YACS is the module responsible for coupling applications, based on spatial and temporal relationships. The coupling of domain decomposition code, such as the coupling of several instances of Code_Aster, a thermomechanical calculation code from EDF R&D, turns out to be a complex task because of the lack of abstraction of current SALOME model.

In 2011, we have proposed and implemented some extensions to the SALOME model and platform to remove this limitation. The main extension is the ability to express the cloning of a service, which generates also the cloning of connections. The actual semantic of the cloning operation has been specified in function of the nature of the service (sequential, parallel) and of the ports (data or control flow). It has greatly simplified the expression of the coupling of several instances of Code_Aster without generating any measurable overhead at runtime: no more recompilation is needed when varying the number of coupled instances.
6.2.7. Towards Data Desktop Grid

Desktop Grids use the computing, network and storage resources from idle desktop PC’s distributed over multiple-LAN’s or the Internet to compute a large variety of resource-demanding distributed applications. While these applications need to access, compute, store and circulate large volumes of data, little attention has been paid to data management in such large-scale, dynamic, heterogeneous, volatile and highly distributed Grids. In most cases, data management relies on ad-hoc solutions, and providing a general approach is still a challenging issue.

We have proposed the BtDew framework which addresses the issue of how to design a programmable environment for automatic and transparent data management on computational Desktop Grids. BtDew relies on a specific set of meta-data to drive key data management operations, namely life cycle, distribution, placement, replication and fault-tolerance with a high level of abstraction.

Since July 2010, in collaboration with the University of Sfax, we are developing a data-aware and parallel version of Magik, an application for arabic writing recognition using the BtDew middleware. We are targeting digital libraries, which require distributed computing infrastructure to store the large number of digitalized books as raw images and at the same time to perform automatic processing of these documents such as OCR, translation, indexing, searching, etc.

In 2011, we have surveyed P2P strategies (replication, erasure code, replica repair, hybrid storage), which provides reliable and durable storage on top of hybrid distributed infrastructures composed of volatile and stable storage. Following this simulation studies, we are implementing a prototype of the Amazon S3 storage on top of BitDew, which will provide reliable storage by using both Desktop free disk space and volunteered remote Cloud storage.

6.2.8. MapReduce programming model for Desktop Grid

MapReduce is an emerging programming model for data-intense application proposed by Google, which has recently attracted a lot of attention. MapReduce borrows from functional programming, where programmer defines Map and Reduce tasks executed on large sets of distributed data. In 2010, we have developed an implementation of the MapReduce programming model based on the BitDew middleware. Our prototype features several optimizations which make our approach suitable for large scale and loosely connected Internet Desktop Grid: massive fault tolerance, replica management, barriers-free execution, latency-hiding optimization as well as distributed result checking. We have presented performance evaluations of the prototype both against micro-benchmarks and real MapReduce applications. The scalability test shows that we achieve linear speedup on the classical WordCount benchmark. Several scenarios involving lagger hosts and host crashes demonstrate that the prototype is able to cope with an experimental context similar to real-world Internet.

In collaboration with the Huazhong University of Science & Technology, we have developed an emulation framework to assess MapReduce on Internet Desktop Grid. We have made extensive comparison on BitDew-MapReduce and Hadoop using Grid5000 which show that our approach has all the properties desirable to cope with an Internet deployment, whereas Hadoop fails on several tests.

In collaboration with the Babes-Bolyai University of Cluj-Napoca, we have proposed a distributed result checker based on the Majority Voting approach. We evaluated the efficiency of our algorithm by computing the aggregated probability with which a MapReduce computation produces an erroneous result.

We have published two chapters in collective books around Cloud and Desktop Grid technologies. The first one, in collaboration with University of Madrid is an introduction to MapReduce and Hadoop, the second one, in collaboration with Virginia Tech is a presentation of two alternative implementations of MapReduce for Desktop Grids : Moon and Bitdew.

6.2.9. SpeQuloS: Providing Quality-of-Service to Desktop Grids using Cloud resources

EDGI is an FP7 European project, following the successful FP7 EDGEs project, whose goal is to build a Grid infrastructure composed of "Desktop Grids", such as BOINC or XtremWeb, where computing resources are
provided by Internet volunteers, and "Service Grids", where computing resources are provided by institutional Grid such as EGEE, gLite, Unicore and "Clouds systems" such as OpenNebula and Eucalyptus, where resources are provided on-demand. The goal of the EDGI project is to provide an infrastructure where Service Grids are extended with public and institutional Desktop Grids and Clouds.

The main limitation with the current infrastructure is that it cannot give any QoS support for applications running in the Desktop Grid (DG) part of the infrastructure. For example, a public DG system enables clients to return work-unit results in the range of weeks. Although there are EGEE applications (e.g. the fusion community’s applications) that can tolerate such a long latency most of the user communities want much smaller latencies.

In 2011, we have developed the SpeQuloS middleware to solve this critical problem. Providing QoS features even in Service Grids is hard and not solved yet satisfactorily. It is even more difficult in an environment where there are no guaranteed resources. In DG systems, resources can leave the system at any time for a long time or forever even after taking several work-units with the promise of computing them. Our approach is based on the extension of DG systems with Cloud resources. For such critical work-units the SpeQuloS system is able to dynamically deploy fast and trustable clients from some Clouds that are available to support the EDGI DG systems. It takes the right decision about assigning the necessary number of trusted clients and Cloud clients for the QoS applications. At this stage, the prototype is fully developed and validated. It supports the XtremWeb and BOINC Desktop Grid and OpenNebula, StratusLab, OpenStack and Amazon EC2 Clouds. The first versions have been delivered to the EDGI production infrastructure. We have conducted extensive simulations to evaluate various strategies of Cloud resources provisioning. Results show that SpeQuloS improve the QoS of BoTs on three aspects: it reduces the makespan by removing the tail effect, it improves the execution stability and it allows to accurately predicts the BoT completion time.

6.2.10. Performance evaluation and modeling

Simulation is a popular approach to obtain objective performance indicators of platforms that are not at one’s disposal. It may for example help the dimensioning of compute clusters in large computing centers. In many cases, the execution of a distributed application does not behave as expected, it is thus necessary to understand what causes this strange behavior. Simulation provides the possibility to reproduce experiments under similar conditions. This is a suitable method for experimental validation of a parallel or distributed application.

The tracing instrumentation of a profiling tool is the ability to save all the information about the execution of an application at run-time. Every scientific application executed computes instructions. The originality of our approach is that we measure the completed instructions of the application and not its execution time. This means that if a distributed application is executed on N cores and we execute it again by mapping two processes per core then we need N/2 cores and more time for the execution time of the application. An execution trace of an instrumented application can be transformed into a corresponding list of actions. These actions can then be simulated by SimGrid. Moreover the SimGrid execution traces will contain almost the same data because the only change is the use of half cores but the same number of processes. This does not affect the number of the completed instructions so the simulation time does not get increased because of the overhead. The Grid’5000 platform is used for this work and the NAS Parallel Benchmarks are used to measure the performance of the clusters.

Our main contribution is to propose of a new execution log format that is time-independent. This means that we decouple the acquisition of the traces from the replay. Furthermore we implemented a trace replay tool which relies on top of fast, scalable and validated simulation kernel of SimGrid. We proved that this framework applies for some of the NAS Parallel Benchmarks and we can predict their performance with a good accuracy. Moreover we are working on further improvements for solving some performance issues with the rest benchmarks. We plan to apply some new techniques about the instrumentation of the benchmarks which we have already discussed with people from the performance analysis community and also improve the trace replay tool in order to improve its accuracy. Finally we did a survey on many different tracing tools with regards to the requirements of our methodology which includes all the latest provided tools from the community.
6.2.11. Elastic Scheduling for Functional Workflows

Non-DAG (or functional) workflows are sets of task-graph workflows with non-deterministic transitions between them, that are determined at runtime by special nodes that control the execution flow. In a current work we are focusing on formalizing and evaluating an allocation and scheduling strategy for on-line non-DAG workflows. The goal of this work is to target real-world non-DAG applications and use cloud platforms to perform elastic allocations while keeping cost and stretch fairness constraints.

To address the previous problem we consider each non-DAG workflow as a set of DAG sub-workflows with non-deterministic transitions between them. Whenever an event occurs (a sub-workflow’s execution is completed, a new workflow arrives in the system, a workflow is canceled, etc.) we need to do a rescheduling. The rescheduling strategy considers the currently-running tasks as fixed. Given that the number of events increases proportional to the number of workflows in the system, there is the risk of spending too much time on the scheduling problem and not enough on the workflows themselves. As a result, the scheduling strategy that we will adopt will be a computational inexpensive one, which will give us more room for the number of possible workflows in the system.

This work is currently in the validation step through experimentation with synthetic data. In the near future we will validate against traces of real-world applications that use non-DAG workflows.

6.2.12. Self Adaptive Middleware Deployment

A computer application can be considered as a system of components that exchange information. Each component type has its specific constraints. The application, as a whole, has also its constraints. Deploying an application on a distributed system consist, among other things, to make a mapping between application components and system resources to meet each component constraints, the application constraints, and possibly those set by the the user. Previous work exists on the deployment of middleware, including DIET (with two finished PhD). However, few take into account the issue of redeployment in the event of variation (availability, load, number) of resources. We study this problem of self adaptive deployment of middleware. It consist of achieving an initial deployment, then scrutinizing some changes in the environment, and automatically adjust the deployment (if beneficial) in case of detecting a variation that degrades the performance expected. To do this, we have surveyed the fields of autonomic computing, self adaptive systems and we have defined the different problems that must be solved to achieve this goal. From this, we first define a resource model to represent the physical system, we are to define a model of middleware-based software components, have started the implementation of the resource model to achieve a simulator.

6.2.13. Virtual Machine Placement with Security Requirements

With the number of services using virtualization and clouds growing faster and faster, it is common to mutualize thousands of virtual machines (VMs) within one distributed system. Consequently, the virtualized services, pieces of software and hardware, and infrastructures share the same physical resources. This has given rise to important challenges regarding the security of VMs and the importance of enforcing non-interference between them. Indeed, cross-VM attacks are an important and real world threat. The problem is even worse in the case of adversary users hosted on the same hardware (multi-tenance). Therefore, the isolation facility within clouds needs to be strong. Furthermore, each user has different adversaries and the placement and scheduling processes need to take these adversaries into account.

First, we have worked on resource model to describe distributed system and application model to describe the composition of virtual machine. Then we have formalize isolation requirements between users, between applications and between virtual machines. We also formalized the redundancy requirement. We have created a simulator that can load our resource model and application model. Using it, we have described the Grid’5000 infrastructure and a Virtual Cluster application. We have formalized and implemented an algorithm that takes into account the requirements and place the application. Work in progress considers using Constraint Satisfaction Problems (CSP) and SAT problems to improve the quality of placement. Moreover, we study the trade-off between performance, security requirements and infrastructure consolidation. This works is part of a project on Cloud Security with Alcatel-Lucent Bell Labs and ENSI de Bourges.
6.2.14. Scheduling for MapReduce Based Applications

After a study of the state of the art regarding scheduling, especially scheduling on grid and clouds and MapReduce application scheduling, experiments were performed over the Grid’5000 and Google/IBM Hadoop platforms. We are now working on improving a previous work by Berlinska and Drozdowski which aims at providing a good static schedule of the Map and Reduce phases. A visualization tool has been developed which draws Gantt charts resulting from Berlinska and Drozdowsky’s algorithms as well as from our own scheduling heuristics.

A BlobSeer model is also developed in collaboration with the Kerdata research team that will be used for our next developments.

6.3. Parallel Sparse Direct Solvers and Combinatorial Scientific Computing

Participants: Maurice Brémond, Guillaume Joslin, Johannes Langguth, Jean-Yves L’Excellent, Mohamed Sid-Lakhdar, Bora Uçar.

6.3.1. Parallel computation of entries of the inverse of a sparse matrix

Following last year’s work on computing entries of the inverse of a sparse matrix in a serial, in-core or out-of-core environment, and that was implemented in MUMPS, we have pursued work to address this issue in a parallel environment. In such this case, it has been shown that minimizing the number of operations (or the number of accesses to the factors) and balancing the work between the processors are contradictory objectives. Several ideas have been investigated and implemented in order to deal with this issue and to reach high speed-ups. Experimental results are promising and show good speed-ups on relatively small number of processors (up to 16) when dealing with large blocks of sparse right-hand sides, while we used to experience speed-downs before.

6.3.2. Multithreaded parallelism for the MUMPS solver

Apart from using message-passing, we have in the past only exploited multicore parallelism through threaded libraries (e.g. BLAS: Basic Linear Algebra Subroutines), and a few OpenMP directives. We are currently investigating the combination of this fork-join model with threaded parallelism resulting from the task graph, which, in our context, is a tree. To do so, and in order to also target NUMA architectures, we apply ideas from distributed-memory environments to multithreaded environments. Simulations based on benchmarks followed by a first prototype implementation have validated this approach for some classes of matrices on small numbers of cores. We are currently revisiting this implementation and plan to pursue experiments on larger numbers of cores with larger classes of matrices. This starting work was done in the context of a master thesis and is the object of a starting PhD thesis. In a distributed-memory environments, it will be combined with parallelism based on message passing, where the scalability of the existing communication schemes should also be addressed. Both directions will be followed in order to face the multicore (r)evolution.

6.3.3. Low-rank approximations

Low-rank approximations are commonly used to compress the representation of data structures. The loss of information induced is often negligible and can be controlled. Although the dense internal datastructures involved in a multifrontal method, the so-called frontal matrices or fronts, are full-rank, they can be represented by a set of low-rank matrices. Applying to our context the notion of geometric clustering used by Bebendorf to define hierarchical matrices, we have shown that the efficiency of this representation to reduce the complexity of both the factorization and solve phases strongly depends on how variables are grouped. The proposed approach can be used either to accelerate the factorization and solution phases or to build a preconditioner. The ultimate goal of this work is to extend the features of the MUMPS solver to exploit low-rank properties.

This work, and the work described in the two previous paragraphs are in the context of a collaboration with ENSEEIHT-IRIT and with the partners involved in the MUMPS project (see Section 5.2 ).
6.3.4. **On partitioning problems with complex objectives**

Hypergraph and graph partitioning tools are used to partition work for efficient parallelization of many sparse matrix computations. Most of the time, the objective function that is reduced by these tools relates to reducing the communication requirements, and the balancing constraints satisfied by these tools relate to balancing the work or memory requirements. Sometimes, the objective sought for having balance is a complex function of a partition. We mention some important class of parallel sparse matrix computations that have such balance objectives. For these cases, the current state of the art partitioning tools fall short of being adequate. To the best of our knowledge, there is only a single algorithmic framework in the literature to address such balance objectives. We propose another algorithmic framework to tackle complex objectives and experimentally investigate the proposed framework.

6.3.5. **On the Use of Cluster-Based Partial Message Logging to Improve Fault Tolerance for MPI HPC Applications**

Fault tolerance is becoming a major concern in HPC systems. The two traditional approaches for message-passing applications, coordinated checkpointing and message logging, have severe scalability issues. Coordinated checkpointing protocols make all processes roll back after a failure. Message logging protocols log a huge amount of data and can induce an overhead on communication performance. Hierarchical rollback-recovery protocols based on the combination of coordinated checkpointing and message logging are an alternative. These partial message logging protocols are based on process clustering: only messages between clusters are logged to limit the consequence of a failure to one cluster. These protocols would work efficiently only if one can find clusters of processes in the applications such that the ratio of logged messages is very low. We study the communication patterns of message passing HPC applications to show that partial message logging is suitable in most cases. We propose a partitioning algorithm to find suitable clusters of processes given the communication pattern of an application. Finally, we evaluate the efficiency of partial message logging using two state of the art protocols on a set of representative applications.

6.3.6. **Integrated data placement and task assignment for scientific workflows in clouds**

We consider the problem of optimizing the execution of data-intensive scientific workflows in the Cloud. We address the problem under the following scenario. The tasks of the workflows communicate through files; the output of a task is used by another task as an input file and if these tasks are assigned on different execution sites, a file transfer is necessary. The output files are to be stored at a site. Each execution site is to be assigned a certain percentage of the files and tasks. These percentages, called target weights, are pre-determined and reflect either user preferences or the storage capacity and computing power of the sites. The aim is to place the data files into and assign the tasks to the execution sites so as to reduce the cost associated with the file transfers, while complying with the target weights. To do this, we model the workflow as a hypergraph and with a hypergraph-partitioning-based formulation, we propose a heuristic which generates data placement and task assignment schemes simultaneously. We report simulation results on a number of real-life and synthetically generated scientific workflows. Our results show that the proposed heuristic is fast, and can find mappings and assignments which reduce file transfers, while respecting the target weights.

6.3.7. **UMPa: A Multi-objective, multi-level partitioner for communication minimization**

We propose a directed hypergraph model and a refinement heuristic to distribute communicating tasks among the processing units in a distributed memory setting. The aim is to achieve load balance and minimize the maximum data sent by a processing unit. We also take two other communication metrics into account with a tie-breaking scheme. With this approach, task distributions causing an excessive use of network or a bottleneck processor which participates to almost all of the communication are avoided. We show on a large number of problem instances that our model improves the maximum data sent by a processor up to 34% for parallel environments with 4, 16, 64 and 256 processing units compared to the state of the art which only minimizes the total communication volume.
6.3.8. A Divisive clustering technique for maximizing the modularity

We present a new graph clustering algorithm aimed at obtaining clusterings of high modularity. The algorithm pursues a divisive clustering approach and using established graph partitioning algorithms and techniques to compute recursive bipartitions of the input as well as to refine clusters. Experimental evaluation shows that the modularity scores obtained compare favorably to many previous approaches. In the majority of test cases, the algorithm outperformed the best known alternatives. In particular, among 13 problem instances common in the literature, the proposed algorithm improves the best known modularity in 9 cases.

6.3.9. Constructing elimination trees for sparse unsymmetric matrices

The elimination tree model for sparse unsymmetric matrices and an algorithm for constructing it have been recently proposed [Eisenstat and Liu, SIAM J. Matrix Anal. Appl., 26 (2005) and 29 (2008)]. The construction algorithm has a worst case time complexity $O(mn)$ for an $n \times n$ unsymmetric matrix having $m$ nonzeros. We propose another algorithm that has a worst case time complexity of $O(m \log n)$.

6.3.10. Multithreaded clustering for multi-level hypergraph partitioning

Requirements for efficient parallelization of many complex and irregular applications can be cast as a hypergraph partitioning problem. The current-state-of-the art software libraries that provide tool support for the hypergraph partitioning problem are designed and implemented before the game-changing advancements in multi-core computing. Hence, analyzing the structure of those tools for designing multithreaded versions of the algorithms is a crucial task. The most successful partitioning tools are based on the multi-level approach. In this approach, a given hypergraph is coarsened to a much smaller one, a partition is obtained on the the smallest hypergraph, and that partition is projected to the original hypergraph while refining it on the intermediate hypergraphs. The coarsening operation corresponds to clustering the vertices of a hypergraph and is the most time consuming task in a multi-level partitioning tool. We present three efficient multithreaded clustering algorithms which are very suited for multi-level partitioners. We compare their performance with that of the ones currently used in today’s hypergraph partitioners. We show on a large number of real life hypergraphs that our implementations, integrated into a commonly used partitioning library PaToH, achieve good speedups without reducing the clustering quality.

6.3.11. Partitioning, ordering, and load balancing in a hierarchically parallel hybrid linear solver

PDSLin is a general-purpose algebraic parallel hybrid (direct/iterative) linear solver based on the Schur complement method. The most challenging step of the solver is the computation of a preconditioner based on an approximate global Schur complement. We investigate two combinatorial problems to enhance PDSLin’s performance at this step. The first is a multi-constraint partitioning problem to balance the workload while computing the preconditioner in parallel. For this, we describe and evaluate a number of graph and hypergraph partitioning algorithms to satisfy our particular objective and constraints. The second problem is to reorder the sparse right-hand side vectors to improve the data access locality during the parallel solution of a sparse triangular system with multiple right-hand sides. This is needed to eliminate the unknowns associated with the interface in PDSLin. We study two reordering techniques: one based on a postordering of the elimination tree and the other based on a hypergraph partitioning. To demonstrate the effect of these techniques on the performance of PDSLin, we present the numerical results of solving large-scale linear systems arising from numerical simulations of modeling accelerator cavities and of modeling fusion devices.

6.3.12. Experiments on push-relabel-based maximum cardinality matching algorithms for bipartite graphs

We report on careful implementations of several push-relabel-based algorithms for solving the problem of finding a maximum cardinality matching in a bipartite graph and compare them with fast augmenting-path-based algorithms. We analyze the algorithms using a common base for all implementations and compare their relative performance and stability on a wide range of graphs. The effect of a set of known initialization heuristics on the performance of matching algorithms is also investigated. Our results identify a variant of
the push-relabel algorithm and a variant of the augmenting-path-based algorithm as the fastest with proper initialization heuristics, while the push-relabel based one having a better worst case performance.

6.3.13. Towards a scalable hybrid linear solver based on combinatorial algorithms

The availability of large-scale computing platforms comprised of tens of thousands of multicore processors motivates the need for the next generation of highly scalable sparse linear system solvers. These solvers must optimize parallel performance, processor (serial) performance, as well as memory requirements, while being robust across broad classes of applications and systems. In this study, we present a hybrid parallel solver that combines the desirable characteristics of direct methods (robustness) and effective iterative solvers (low computational cost), while alleviating their drawbacks (memory requirements, lack of robustness). We discuss several combinatorial problems that arise in the design of this hybrid solver, present algorithms to solve these combinatorial problems, and demonstrate their impact on a large-scale three-dimensional PDE-constrained optimization problem.
GRAND-LARGE Project-Team

6. New Results

6.1. Communication avoiding algorithms for linear algebra

Participants: Laura Grigori, Simplice Donfack, Amal Khabou, Mathias Jacquelin, Sophie Moufawad.

The focus of this research is on the design of efficient parallel algorithms for solving problems in numerical linear algebra, as solving very large sets of linear equations and large least squares problems, often with millions of rows and columns. These problems arise in many numerical simulations, and solving them is very time consuming.

This research focuses on developing new algorithms for linear algebra problems, that minimize the required communication, in terms of both latency and bandwidth. We have introduced in 2008 two communication avoiding algorithms for computing the LU and QR factorizations, that we refer to as CALU and CAQR (joint work with J. Demmel and M. Hoemmen from U.C. Berkeley, J. Langou from C.U. Denver, and H. Xiang then at INRIA) [6], [9]. Since then, we have also designed a communication avoiding algorithm for rank revealing QR. In addition, we have also extended theoretical lower bounds to sparse Cholesky factorization and identified algorithms that attain these bounds and so minimize communication. The communication avoiding algorithms are now studied by several other groups, including groups at INRIA, and they start being implemented and being available in public libraries as ScaLAPACK.

During 2011, our research has focused on a study of the stability of communication avoiding LU factorization and its implementation on multicore machines. In [20] we focus on numerical properties of CALU. To decrease the communication required in the LU factorization, CALU uses a new pivoting strategy, referred to as tournament pivoting, that may lead to a different row permutation than the classic LU factorization with partial pivoting. We have further investigated the numerical stability of CALU. The reason to consider CALU is that it does an optimal amount of communication, and asymptotically less than Gaussian elimination with partial pivoting (GEPP), and so will be much faster on platforms where communication is expensive, as shown in previous work. We show that the Schur complement obtained after each step of performing CALU on a matrix A is the same as the Schur complement obtained after performing GEPP on a larger matrix whose entries are the same as the entries of A (sometimes slightly perturbed) and zeros. More generally, the entire CALU process is equivalent to GEPP on a large, but very sparse matrix, formed by entries of A and zeros. Hence we expect that CALU will behave as GEPP and it will be also very stable in practice. In addition, extensive experiments on random matrices and a set of special matrices show that CALU is stable in practice. The upper bound on the growth factor of CALU is worse than of GEPP. However, there are Wilkinson like-matrices for which GEPP has exponential growth factor, but not CALU, and vice-versa.

We present experimental results for random matrices and for a set of special matrices, including sparse matrices, for binary tree based and flat-tree-based CALU. We discuss both the stability of the LU factorization and of the linear solver, in terms of pivot growth and backward errors. The results show that in practice CALU is stable. We present the backward errors measured three ways: by \( \|PA - LU\|/\|A\| \); by the normwise backward error \( \|Ax - b\|/(\|A\|\|x\| + \|b\|) \); and by the componentwise backward error (after iterative refinement in working precision). For random matrices, all CALU’s backward errors were at most 1.9x larger than GEPP’s backward errors. We also tested "special" matrices, including known difficult examples: (1) The ratios of \( \|PA - LU\|/\|A\| \) were at most 1 in over 69% of cases (i.e. CALU was at least as stable as GEPP), and always 1.5 or smaller, except for one ratio of 4.3, in which case both backward errors were much smaller than \( 2^{-53} = \) machine epsilon. (2) The ratios of normwise backward errors were at most 1 in over 53% of cases, and always 1.5 or smaller, except for 5 ratios ranging up to 26, in which cases all backward errors were much smaller than 4x machine epsilon. (3) The ratios of componentwise backward errors were at most 1 in over 52% of cases, and always 3.2 or smaller, except for one ratio of 8.3.
In [30] we design a scheduling algorithm for efficiently executing CALU on multicore architectures. We focus on a tunable scheduling strategy that maintains load balance across cores while also maintaining data locality and low dequeue overhead. To achieve this, we use a strategy that combines static and dynamic scheduling. This approach was shown to be successful on regular mesh computations by V. Kale and B. Gropp. This tunable scheduling strategy allows us to flexibly control the percentage of tasks that can be scheduled dynamically; this gives a knob to control load balancing so that it occurs only at the point in computation when the benefits it provides outweighs the costs it induces. On NUMA machines where remote memory access is costly, the percentage of work scheduled dynamically should be small enough to avoid excessive cache misses, but large enough to keep the cores busy during idle times in the static part.

In this work, we show the effectiveness of this method in the context of already highly-optimized dense matrix factorizations. Our prior work on multi-threaded CALU was based on dynamic scheduling. The algorithm performed well on tall and skinny matrices, but became less scalable on square matrices with increasing numbers of processors. We show that the usage of this scheduling in communication avoiding dense factorization leads to significant performance gains. On a 48 core AMD Opteron NUMA machine, our experiments show that we can achieve up to 64% improvement over a version of CALU that uses fully dynamic scheduling, and up to 30% improvement over the version of CALU that uses fully static scheduling. On a 16-core Intel Xeon machine, our hybrid static/dynamic scheduling approach is up to 8% faster than the version of CALU that uses a fully static scheduling or fully dynamic scheduling. Our algorithm leads to speedups over the corresponding routines for computing LU factorization in well known libraries. On the 48 core AMD NUMA machine, our best implementation is up to 110% faster than MKL, while on the 16 core Intel Xeon machine, it is up to 82% faster than MKL. Our approach also shows significant speedups compared with PLASMA on both of these systems.

6.2. Preconditioning techniques for solving large systems of equations

Participants: Laura Grigori, Riadh Fezzanni, Sophie Moufawad.

A different direction of research is related to preconditioning large sparse linear systems of equations. This research is performed in the context of ANR PETALh project (2011-2012), which follows the ANR PETAL project (2008-2009). It is conducted in collaboration with Frederic Nataf from University Paris 6.

Several highly used preconditioners are for example the incomplete LU factorizations and Schwarz based approaches as used in domain decomposition. Most of these preconditioners are known to have scalability problems. The number of iterations can increase significantly when the size of the problem increases or when the number of independent domains is increased. This is often due to the presence of several low frequency modes that hinder the convergence of the iterative method. To address this problem, we study a different class of preconditioners, called direction preserving or filtering preconditioners. These preconditioners have the property of being identical to the input matrix on a given filtering vector. A judicious choice of the vector allows to alleviate the effect of low frequency modes on the convergence.

We consider in particular two classes of preconditioners. The first preconditioner is an incomplete decomposition that satisfies the filtering property [11]. The nested preconditioner has the same property for a specific vector of all ones. However the construction is different and takes advantage of a nested structure of the input matrix. The previous research on these methods considered only matrices arising from the discretization of PDEs on structured grids, where the matrix has a block tridiagonal structure. This structure imposes a sequential computation of the preconditioner and it is not suitable for the more general case of unsstructured grids. Hence, while very efficient, the usage of these preconditioners was very limited. At the beginning of this research we have obtained several theoretical results for these methods that demonstrate their numerical behavior and convergence properties for cases arising from the discretization of PDEs on structured grids [11]. But the main result is the development of a generalized method [48], [46] that has two important properties: it allows the filtering property to be satisfied for any input matrix; the matrix can be reordered such that its computation is highly parallel. Experimental results show that the method is very efficient for certain classes of matrices, and shows good scalability results in terms of both problem size and number of processors.
6.3. Microwave Data Analysis for petaScale computers
Participants: Laura Grigori, Mikolaj Szydlarski, Meisam Sharify.

In [47] we describe a scalable algorithm for computing an inverse spherical harmonic transform suitable for cluster of multiple CPU-GPUs. We base our implementation on hybrid programming combining MPI and CUDA. We focus our attention on the two major sequential steps involved in the transforms computation, retaining the efficient parallel framework of the original code. We detail optimization techniques used to enhance the performance of the OpenMP/CUDA-based code and compare them with those implemented in the public domain parallel package, S2HAT.

We also present performance comparisons of the multi GPU version and a hybrid, MPI/OpenMP version of the same transform. We find that one NVIDIA Tesla S1070 can accelerate overall execution time of the SHT by as much as 3 times with respect to the MPI/OpenMP version executed on one quad-core processor (Intel Nehalem 2.93 GHz) and, owing to very good scalability of both versions, 128 Tesla cards perform as good as 256 twelve-core processor (AMD Opteron 2.1 GHz).

The work presented here has been performed in the context of the Cosmic Microwave Background simulations and analysis. However, we expect that the developed software will be of more general interest and applicability.

6.4. Innovative linear system solvers for hybrid multicore/GPU architectures
Participant: Marc Baboulin.

The advent of new processor architectures (e.g. multicore, GPUs) requires the rethinking of most of the scientific applications and innovative methods must be proposed in order to take full advantage of current supercomputers [12].

To accelerate linear algebra solvers on current parallel machines, we introduced in public domain libraries a class of solvers based on statistical techniques. A first application concerns the solution of a square linear systems \( Ax = b \). We study a random transformation of \( A \) that enables us to avoid pivoting and then to reduce the amount of communication [54]. Numerical experiments show that this randomization can be performed at a very affordable computational price while providing us with a satisfying accuracy when compared to partial pivoting. This random transformation called Partial Random Butterfly Transformation (PRBT) is optimized in terms of data storage and flops count. In the solver that we developed, PRBT combined with LU factorization with no pivoting take advantage of the latest generation of hybrid multicore/GPU machines and outperform existing factorization routines from current parallel library MAGMA.

A second application is related to solving symmetric indefinite systems via \( LDL^T \) factorization for which there was no existing parallel implementation in the dense library ScALAPACK. We developed an efficient and innovative parallel tiled algorithm for solving symmetric indefinite systems on multicore architectures [59]& [1]. This solver avoids pivoting by using a multiplicative preconditioning based on symmetric randomization. This randomization prevents the communication overhead due to pivoting, is computationally inexpensive and requires very little storage. Following randomization, a tiled LDLT factorization is used that reduces synchronization by using static or dynamic scheduling. We compare Gflop/s performance of our solver with other types of factorizations on a current multicore machine and we provide tests on accuracy using LAPACK test cases.

6.5. MILEPOST GCC: machine learning enabled self-tuning compiler
Participant: Grigori Fursin [correspondant].

Tuning compiler optimizations for rapidly evolving hardware makes porting and extending an optimizing compiler for each new platform extremely challenging. Iterative optimization is a popular approach to adapting programs to a new architecture automatically using feedback-directed compilation. However, the large number of evaluations required for each program has prevented iterative compilation from widespread take-up in production compilers. Machine learning has been proposed to tune optimizations across programs systematically but is currently limited to a few transformations, long training phases and critically lacks publicly released, stable tools.
Our approach is to develop a modular, extensible, self-tuning optimization infrastructure to automatically learn the best optimizations across multiple programs and architectures based on the correlation between program features, run-time behavior and optimizations. In this paper we describe MILEPOST GCC, the first publicly-available open-source machine learning-based compiler. It consists of an Interactive Compilation Interface (ICI) and plugins to extract program features and exchange optimization data with the cTuning.org open public repository. It automatically adapts the internal optimization heuristic at function-level granularity to improve execution time, code size and compilation time of a new program on a given architecture. Part of the MILEPOST technology together with low-level ICI-inspired plugin framework is now included in the mainline GCC.

We developed machine learning plugins based on probabilistic and transductive approaches to predict good combinations of optimizations. Our preliminary experimental results show that it is possible to automatically reduce the execution time of individual MiBench programs on various machines from GRID5000, some by more than a factor of 2, while also improving compilation time and code size. We also present a realistic multi-objective optimization scenario for Berkeley DB library using MILEPOST GCC and improve execution time by approximately 17%, while reducing compilation time and code size by 12% and 7% respectively on Intel Xeon processor.

6.6. Loop Transformations: Convexity, Pruning and Optimization

Participant: Cédric Bastoul.

High-level loop transformations are a key instrument in mapping computational kernels to effectively exploit resources in modern processor architectures. However, determining appropriate compositions of loop transformations to achieve this remains a significantly challenging task; current compilers may achieve significantly lower performance than hand-optimized programs. To address this fundamental challenge, we first present a convex characterization of all distinct, semantics-preserving, multidimensional affine transformations. We then bring together algebraic, algorithmic, and performance analysis results to design a tractable optimization algorithm over this highly expressive space. The framework has been implemented and validated experimentally on a representative set of benchmarks run on state-of-the-art multi-core platforms.

6.7. Exact algorithm for the l1-compressive sensing problem using a modified Dantzig-Wolfe method

Participants: Alexandre Borghi, Jerome Darbon, Sylvain Peyronnet.

In this work, we consider the l1-Compressive Sensing problem and presents an efficient algorithm that computes an exact solution. The idea consists in reformulating the problem such that it yields a modified Dantzig-Wolfe decomposition that allows to efficiently apply all standard simplex pivoting rules. Experimental results show the superiority of our approach compared to standard linear programming methods.

6.8. Supple: a flexible probabilistic data dissemination protocol for wireless sensor networks

Participants: Aline Carneiro Viana, Thomas Hérault, Thomas LArgillier, Sylvain Peyronnet, Fatiha Zaidi.

We propose a flexible proactive data dissemination approach for data gathering in self-organized Wireless Sensor Networks (WSN). Our protocol Supple, effectively distributes and stores monitored data in WSNs such that it can be later sent to or retrieved by a sink. Supple empowers sensors with the ability to make on the fly forwarding and data storing decisions and relies on flexible and self-organizing selection criteria, which can follow any predefined distribution law. Using formal analysis and simulation, we show that Supple is effective in selecting storing nodes that respect the predefined distribution criterion with low overhead and limited network knowledge.
6.9. Non-self-stabilizing and self-stabilizing gathering in networks of mobile agents–the notion of speed

**Participants:** Joffroy Beauquier, Janna Burman, Julien Clément, Shay Kutten.

In the population protocol model, each agent is represented by a finite state machine. Agents are anonymous and supposed to move in an asynchronous way. When two agents come into range of each other (“meet”), they can exchange information. One of the vast variety of motivating examples to the population protocols model is ZebraNet. ZebraNet is a habitat monitoring application where sensors are attached to zebras and collect biometric data (e.g. heart rate, body temperature) and information about their behavior and migration patterns (via GPS). The population protocol model is, in some sense, related to cloud computing and to networks characterized by asynchrony, large scale, the possibility of failures, in the agents as well as in the communications, with the constraint that each agent is resource limited.

In order to extend the computation power and efficiency of the population protocol model, various extensions were suggested. Our contribution is an extension of the population protocol model that introduces the notion of “speed”, in order to capture the fact that the mobile agents move at different speeds and/or have different communication ranges and/or move according to different patterns and/or visit different places with different frequencies. Intuitively, fast agents which carry sensors with big communication ranges communicate with other agents more frequently than other agents do. This notion is formalized by allocating a cover time, cv, to each mobile agent v. cv is the minimum number of events in the whole system that occur before agent v meets every other agent at least once. As a fundamental example, we have considered the basic problem of gathering information that is distributed among anonymous mobile agents and where the number of agents is unknown. Each mobile agent owns a sensed input value and the goal is to communicate the values (as a multi-set, one value per mobile agent) to a fixed non-mobile base station (BS), with no duplicates or losses.

Gathering is a building block for many monitoring applications in networks of mobile agents. For example, a solution to this problem can solve a transaction commit/abort task in MANETs, if the input values of agents are votes (and the number of agents is known to BS). Moreover, the gathering problem can be viewed as a formulation of the routing problem in Disruption Tolerant Networks.

We gave different solutions to the gathering in the model of mobile agents with speed and we proved that one of them is optimal.

6.10. Making Population Protocols Self-stabilizing

**Participants:** Joffroy Beauquier, Janna Burman, Shay Kutten, Brigitte Rozoy.

As stated in the previous paragraph, the application domains of the population protocol model are asynchronous large scale networks, in which failures are possible and must be taken into account. This work concerns failures and namely the technique of self-stabilization for tolerating them.

Developing self-stabilizing solutions (and proving them) is considered to be more challenging and complicated than developing classical solutions, where a proper initialization of the variables can be assumed. This remark holds for a large variety of models and hence, to ease the task of the developers, some automatic techniques have been proposed to transform programs into self-stabilizing ones.

We have proposed such a transformer for algorithms in the population protocol model introduced for dealing with resource-limited mobile agents. The model we consider is a variation of the original one in that there is a non mobile agent, the base station, and that the communication characteristics (e.g. moving speed, communication radius) of the agents are considered through the notion of cover time.

The automatic transformer takes as an input an algorithm solving a static problem and outputs a self-stabilizing solution for the same problem. To the best of our knowledge, it is the first time that such a transformer for self-stabilization is presented in the framework of population protocols. We prove that the transformer we propose is correct and we make the complexity analysis of the stabilization time.
6.11. Self-stabilizing synchronization in population protocols with cover times

Participants: Joffroy Beauquier, Janna Burman, Shay Kutten, Brigitte Rozoy.

Synchronization is widely considered as an important service in distributed systems which may simplify protocol design. Phase clock is a general synchronization tool that provides a form of a logical time. We have developed a self-stabilizing phase clock algorithm suited to the model of population protocols with cover time. We have shown that a phase clock is impossible in the model with only constant-state agents. Hence, we assumed an existence of resource unlimited agent - the base station. The clock size and duration of each phase of the proposed phase clock tool are adjustable by the user. We provided application examples of this tool and demonstrate how it can simplify the design of protocols. In particular, it yields a solution to Group Mutual Exclusion problem.

6.12. Impossibility of consensus for population protocol with cover times

Participants: Joffroy Beauquier, Janna Burman.

We have extended the impossibility result for asynchronous consensus of Fischer, Lynch and Paterson (FLP) to the asynchronous model of population protocols with cover times. We noted that the proof of FLP does not apply. Indeed, the key lemma stating that two successive factors in an execution, involving disjoint subsets of agents, commute, is no longer true, because of the cover time property. Then we developed a completely different approach and we proved that there is no general solution to consensus for population protocols with cover times, even if there is a single possible crash. We noted that this impossibility result also applies to randomized asynchronous consensus, contrary to what happens in the classical message-passing or shared memory communication models, in which the problem is solvable inside some bounds on the number of faulty processes. Then, for circumventing these impossibility results, we introduced the phase clock oracle and the S oracle, and we shown how they allow to design solutions.

6.13. Routing and synchronization in large scale networks of very cheap mobile sensors

Participants: Joffroy Beauquier, Brigitte Rozoy.

In a next future, large networks of very cheap mobile sensors will be deployed for various applications, going from wild life preserving or environmental monitoring up to medical or industrial system control. Each sensor will cost only a few euros, allowing a large scale deployment. They will have only a few bit of memory, no identifier, weak capacities of computation and communication, no real time clock and will be prone to failures. Moreover such networks will be fundamentally dynamic. The goal of this subject is to develop the basic protocols and algorithms for rudimentary distributed systems for such networks. The studied problems are basic ones, like data collection, synchronization (phase clock, mutual exclusion, group mutual exclusion), fault tolerance (consensus), automatic transformers, always in a context of possible failures. A well known model has already been proposed for such networks, the population protocol model. In this model, each sensor is represented by a finite state machine. Sensors are anonymous and move in an asynchronous way. When two sensors come into range of each other ("meet"), they can exchange information. One of the vast variety of motivating examples for this model is ZebraNet. ZebraNet is a habitat monitoring application in which sensors are attached to zebras in order to collect biometric data (e.g., heart rate, body temperature) and information about their behavior and migration patterns. Each pair of zebras meets from time to time. During such meetings (events), ZebraNet’s agents (zebras’ attached sensors) exchange data. Each agent stores its own sensor data as well as data of other sensors that were in range in the past. They upload data to a base station whenever it is nearby. It was shown that the set of applications that can be solved in the original model of population protocols is rather limited. Other models (such as some models of Delay/Disruption-Tolerant Networks - DTNs), where each node maintains links and connections even to nodes it may interact with only intermittently, do not seem to suit networks with small memory agents and a very large (and unknown) set of anonymous agents. That is why we enhance the model of population protocols by introducing a notion
of "speed". We try to capture the fact that the mobile agents move at different speeds and/or have different communication ranges and/or move according to different patterns and/or visit different places with different frequencies. Intuitively, fast agents which carry sensors with large communication ranges communicate with other agents more frequently than other agents do. This notion is formalized by the notion of cover time for each agent. The cover time of an agent is the unknown number of events (pairwise meetings) in the whole system that occur (during any execution interval) before agent v meets every other agent at least once. The model we propose is somehow validated by some recent statistical results, obtained from empirical data sets regarding human or animal mobility. An important consequence of our approach is that the analytic complexity of the protocols designed in this model is possible, independently of any simulation or experimentation. For instance, we consider the fundamental problem of gathering different pieces of information, each sensed by a different anonymous mobile agent, and where the number of agents is unknown. The goal is to communicate the sensed values (as a multi-set, one value per mobile agent) to a base station, with no duplicates or losses. Gathering is a building block for many monitoring applications in networks of mobile agents. Moreover, the gathering problem can be viewed as a special case of the routing problem in DTNs, in which there is only one destination, the base station. Then we are able to compute the complexity of solutions we propose, as well as those of solutions used in experimental projects (like ZebraNet), and to compare them. The algorithms we present are self-stabilizing. Such algorithms have the important property of operating correctly regardless of their initial state (except for some bounded period). In practice, self-stabilizing algorithms adjust themselves automatically to any changes or corruptions of the network components (excluding the algorithm’s code). These changes are assumed to cease for some sufficiently long period. Self-stabilization is considered for two reasons. First, mobile agents are generally fragile, subject to failures and hard to initialize. Second, systems of mobile agents are by essence dynamic, some agents leave the system while new ones are introduced. Self-stabilization is a well adapted framework for dealing with such situations.

6.14. Self-Stabilizing Control Infrastructure for HPC

Participants: Thomas Hérault, Camille Coti.

High performance computing platforms are becoming larger, leading to scalability and fault-tolerance issues for both applications and runtime environments (RTE) dedicated to run on such machines. After being deployed, usually following a spanning tree, a RTE needs to build its own communication infrastructure to manage and monitor the tasks of parallel applications. Previous works have demonstrated that the Binomial Graph topology (BMG) is a good candidate as a communication infrastructure for supporting scalable and fault-tolerant RTE.

In this work, we presented and analyzed a self-stabilizing algorithm to transform the underlying communication infrastructure provided by the launching service (usually a tree, due to its scalability during launch time) into a BMG, and maintain it in spite of failures. We demonstrated that this algorithm is scalable, tolerates transient failures, and adapts itself to topology changes.

The algorithms are scalable, in the sense that all process memory, number of established communication links, and size of messages are logarithmic with the number of elements in the system. The number of synchronous rounds to build the system is also logarithmic, and the number of asynchronous rounds in the worst case is square logarithmic with the number of elements in the system. Moreover, the self-stabilizing property of the algorithms presented induce fault-tolerance and self-adaptivity. Performance evaluation based on simulations predicts a fast convergence time (1/33s for 64K nodes), exhibiting the promising properties of such self-stabilizing approach.

We pursue this work by implementing and evaluating the algorithms in the STCI runtime environment to validate the theoretical results.

6.15. Large Scale Peer to Peer Performance Evaluations

Participant: Serge Petiton.
6.15.1. Large Scale Grid Computing

Recent progress has made possible to construct high performance distributed computing environments, such as computational grids and cluster of clusters, which provide access to large scale heterogeneous computational resources. Exploration of novel algorithms and evaluation of performance is a strategic research for the future of computational grid scientific computing for many important applications [88]. We adapted [68] an explicit restarted Lanczos algorithm on a world-wide heterogeneous grid platform. This method computes one or few eigenpairs of a large sparse real symmetric matrix. We take the specificities of computational resources into account and deal with communications over the Internet by means of techniques such as out-of-core and data persistence. We also show that a restarted algorithm and the combination of several paradigms of parallelism are interesting in this context. We perform many experimentations using several parameters related to the Lanczos method and the configuration of the platform. Depending on the number of computed Ritz eigenpairs, the results underline how critical the choice of the dimension of the working subspace is. Moreover, the size of platform has to be scaled to the order of the eigenproblem because of communications over the Internet.

6.15.2. High Performance Cluster Computing

Grid computing focuses on making use of a very large amount of resources from a large-scale computing environment. It intends to deliver high-performance computing over distributed platforms for computation and data-intensive applications. We propose [99] an effective parallel hybrid asynchronous method to solve large sparse linear systems by the use of a Grid Computing platform Grid5000. This hybrid method combines a parallel GMRES(m) (Generalized Minimum RESidual) algorithm with the Least Square method that needs some eigenvalues obtained from a parallel Arnoldi algorithm. All of these algorithms run on the different processors of the platform Grid5000. Grid5000, a 5000 CPUs nation-wide infrastructure for research in Grid computing, is designed to provide a scientific tool for computing. We discuss the performances of this hybrid method deployed on Grid5000, and compare these performances with those on the IBM SP series supercomputers.

6.15.3. Large Scale Power aware Computing

Energy conservation is a dynamic topic of research in High Performance Computing and Cluster Computing. Power-aware computing for heterogeneous world-wide Grid is a new track of research. We have studied and evaluated the impact of the heterogeneity of the computing nodes of a Grid platform on the energy consumption. We propose to take advantage of the slack-time caused by the heterogeneity in order to save energy with no significant loss of performance by using Dynamic Voltage Scaling (DVS) in a distributed eigen solver [69]. We show that using DVS only during the slack-time does not penalize the performances but it does not provide significant energy savings. If DVS is applied to all the execution, we get important global and local energy savings (respectively up to 9% and 20%) without a significant rise of the wall-clock times.

6.16. High Performance Linear Algebra on the Grid

Participants: Thomas Hérault, Camille Coti.

Previous studies have reported that common dense linear algebra operations do not achieve speed up by using multiple geographical sites of a computational grid. Because such operations are the building blocks of most scientific applications, conventional supercomputers are still strongly predominant in high-performance computing and the use of grids for speeding up large-scale scientific problems is limited to applications exhibiting parallelism at a higher level.

In this work, we have identified two performance bottlenecks in the distributed memory algorithms implemented in ScaLAPACK, a state-of-the-art dense linear algebra library. First, because ScaLAPACK assumes a homogeneous communication network, the implementations of ScaLAPACK algorithms lack locality in their communication pattern. Second, the number of messages sent in the ScaLAPACK algorithms is significantly greater than other algorithms that trade flops for communication.
This year, we presented a new approach for computing a QR factorization one of the main dense linear algebra kernels of tall and skinny matrices in a grid computing environment that overcomes these two bottlenecks. Our contribution is to articulate a recently proposed algorithm (Communication Avoiding QR) with a topology-aware middleware (QCG-OMPI) in order to confine intensive communications (ScALAPACK calls) within the different geographical sites.

An experimental study conducted on the Grid5000 platform shows that the resulting performance increases linearly with the number of geographical sites on large-scale problems (and is in particular consistently higher than ScALAPACKs).

6.17. Emulation of Volatile Systems

Participants: Thomas Largillier, Benjamin Quetier, Sylvain Peyronnet, Thomas Hérault, Franck Cappello.

In the process of developing grid applications, people need to often evaluate the robustness of their work. Two common approaches are simulation, where one can evaluate his software and predict behaviors under conditions usually unachievable in a laboratory experiment, and experimentation, where the actual application is launched on an actual grid. However, simulation could ignore unpredictable behaviors due to the abstraction done and experiment does not guarantee a controlled and reproducible environment, and simulation often introduces a high level of abstraction that make the discovery and study of unexpected, but real, behaviors a rare event.

In this work, we proposed an emulation platform for parallel and distributed systems including grids where both the machines and the network are virtualized at a low level. The use of virtual machines allows us to test highly accurate failure injection since we can destroy virtual machines, and network virtualization provides low-level network emulation. Failure accuracy is a criteria that evaluates how realistic a fault is. The accuracy of our framework has been evaluated through a set of micro benchmarks and a very stable P2P system called Pastry.

We are in the process of developing a fault injection tool to work with the platform. It will be an extension of the work started in the tool Fail. The interest of this work is that using Xen virtual machines will allow to model strong adversaries since it is possible to have virtual machines with shared memory. These adversaries will be stronger since they will be able to use global fault injection strategies.

6.18. Exascale Systems

Participant: Franck Cappello.

Over the last 20 years, the open-source community has provided more and more software on which the world’s high-performance computing systems depend for performance and productivity. The community has invested millions of dollars and years of effort to build key components. Although the investments in these separate software elements have been tremendously valuable, a great deal of productivity has also been lost because of the lack of planning, coordination, and key integration of technologies necessary to make them work together smoothly and efficiently, both within individual petascale systems and between different systems. A repository gatekeeper and an email discussion list can coordinate open-source development within a single project, but there is no global mechanism working across the community to identify critical holes in the overall software environment, spot opportunities for beneficial integration, or specify requirements for more careful coordination. It seems clear that this completely uncoordinated development model will not provide the software needed to support the unprecedented parallelism required for peta/exascale computation on millions of cores, or the flexibility required to exploit new hardware models and features, such as transactional memory, speculative execution, and GPUs. We presented a rational promoting that the community must work together to prepare for the challenges of exascale computing, ultimately combing their efforts in a coordinated International Exascale Software Project.
Over the past few years resilience has become a major issue for high-performance computing (HPC) systems, in particular in the perspective of large petascale systems and future exascale systems. These systems will typically gather from half a million to several millions of central processing unit (CPU) cores running up to a billion threads. From the current knowledge and observations of existing large systems, it is anticipated that exascale systems will experience various kind of faults many times per day. It is also anticipated that the current approach for resilience, which relies on automatic or application level checkpoint/restart, will not work because the time for checkpointing and restarting will exceed the mean time to failure of a full system. This set of projections leaves the community of fault tolerance for HPC systems with a difficult challenge: finding new approaches, which are possibly radically disruptive, to run applications until their normal termination, despite the essentially unstable nature of exascale systems. Yet, the community has only five to six years to solve the problem. In order to start addressing this challenge, we synthesized the motivations, observations and research issues considered as determinant of several complimentary experts of HPC in applications, programming models, distributed systems and system management.

As a first step to address the resilience challenge, we conducted a comprehensive study of the state of the art. The emergence of petascale systems and the promise of future exascale systems have reinvigorated the community interest in how to manage failures in such systems and ensure that large applications, lasting several hours or tens of hours, are completed successfully. Most of the existing results for several key mechanisms associated with fault tolerance in high-performance computing (HPC) platforms follow the rollback-recovery approach. Over the last decade, these mechanisms have received a lot of attention from the community with different levels of success. Unfortunately, despite their high degree of optimization, existing approaches do not fit well with the challenging evolutions of large-scale systems. There is room and even a need for new approaches. Opportunities may come from different origins: diskless checkpointing, algorithmic-based fault tolerance, proactive operation, speculative execution, software transactional memory, forward recovery, etc.

We provided the following contributions: (1) we summarize and analyze the existing results concerning the failures in large-scale computers and point out the urgent need for drastic improvements or disruptive approaches for fault tolerance in these systems; (2) we sketch most of the known opportunities and analyze their associated limitations; (3) we extract and express the challenges that the HPC community will have to face for addressing the stringent issue of failures in HPC systems.
6. New Results

6.1. Algorithms and high-performance solvers

6.1.1. Dense linear algebra solvers for multicore processors accelerated with multiple GPUs

In collaboration with the Inria RUNTIME team and the University of Tennessee, we have designed dense linear algebra solvers that can fully exploit a node composed of a multicore processor accelerated with multiple GPUs. This work has been integrated in the latest release of the MAGMA package (http://icl.cs.utk.edu/magma/).

6.1.2. Hybrid direct/iterative solvers based on algebraic domain decomposition techniques

A first release of the MaPHyS package should be made available early in 2012 thanks to the developments conducted in the last year of the ADT. An approximation of the local Schur complement has been studied that is based on approximated inverse technique. This work is a natural extension of part of the PhD research of Mikko Byckling. Furthermore, during his master internship, Stojce Nakov has investigated the design of a Krylov subspace method, namely the conjugate gradient, on a run-time system in order to best exploit the computing capabilities of many-GPU nodes and manycore systems. In the framework of his starting PhD funded by TOTAL, Stojce Nakov will continue his work to design a new implementation of a hybrid linear solver (see Section 3.3) for heterogeneous manycore platforms.

6.1.3. Resilience in numerical simulations

In his master internship work, Mawussi Zounon investigated recovery strategies for core faults in the framework of parallel preconditioned Krylov solvers. The underlying idea is to recover fault entries of the iterate via interpolation from existing values available on neighbor cores. He will continue this work in the framework of his PhD funded by the ANR-RESCUE. Notice that these activities are also part of our contribution to the G8-ECS (Enabling Climate Simulation at extreme scale).

6.1.4. Full geometric multigrid method for 3D Maxwell equations

In the context of a collaboration with the CEA/CESTA center, Mathieu Chanaud continued his PhD work on a tight combination between multigrid methods and direct methods for the efficient solution of challenging 3D irregular finite element problems arising from the discretization of Maxwell equations. A parallel solver dedicated to the ODYSSEE challenge (electromagnetism) of CEA/CESTA has been implemented and integrated. The novel parallel solver was able to solve a 1.3 billion system given a 20 million unknown problem at the coarsest level. The input mesh defines the coarsest level. This mesh is further refined to defined the grid hierarchy, where matrix free smoothers are considered to reduce the memory consumption.

6.1.5. Scalable numerical schemes for scientific applications

A work is currently carried on with TOTAL (Rached Abdelkhalek PhD). The extraordinary challenge that the oil and gas industry must face for hydrocarbon exploration requires the development of leading edge technologies to recover an accurate representation of the subsurface. Seismic modeling and Reverse Time Migration (RTM) based on the full wave equation discretization, are tools of major importance since they give an accurate representation of complex wave propagation areas. Unfortunately, they are highly compute intensive. The recent development in GPU technologies with unified architecture and general-purpose languages coupled with the high and rapidly increasing performance throughput of these components made General Purpose Processing on Graphics Processing Units an attractive solution to speed up diverse applications. We have designed a fast parallel simulator that solves the acoustic wave equation on a GPU cluster. Solving the acoustic wave equation in an oil exploration industrial context aims at speeding up seismic modeling and Reverse Time Migration. We consider a finite difference approach on a regular mesh, in both
2D and 3D cases. The acoustic wave equation is solved in a constant density or a variable density domain. All the computations are done in single precision, since double precision is not required in our context. We use nvidia CUDA to take advantage of the GPU computational power. We study different implementations and their impact on the application performance. We obtain a speed up of 16 for Reverse Time Migration and up to 43 for the modeling application over a sequential code running on general purpose CPU. The defense of this thesis is planned early 2012.

For the solution of the elastodynamic equation on meshes with local refinments, we are currently collaborating with Total to design a parallel implementation of a local time refinement technique on top of a discontinuous Galerkin space discretization. This latter technique enables to manage non-conforming meshes suited to deal with multiblock approaches that capture the locally refined regions. This work is developed in the framework of Yohann Dudouit PhD thesis. A software prototype is currently developed to address these simulations.

The calculation of acoustic modes in combustion chambers is a challenging calculation for large 3D geometries. It requires the calculation of a few of the smallest eigenpairs of large unsymmetric matrices in a parallel environment. A new block Arnoldi approach is currently developed to best benefit from the continuation scheme used in this application context. This is part of the PhD research activity of Pablo Salas.

### 6.2. Efficient algorithmics for code coupling in complex simulations

The performance of the coupled codes depends on how the data are well distributed on the processors. Generally, the data distributions of each code are built independently from each other to obtain the best load-balancing. But once the codes are coupled, the naive use of these decompositions can lead to important imbalance in the coupling area. Therefore, the modeling of the whole coupling is crucial to improve the performance and to ensure a good scalability. The goal is to find the best data distribution for the whole coupled codes and not only for each standalone code. The key idea is to use a graph/hypergraph model that will incorporate information about the coupling itself. Then, we propose new algorithms to perform a coupling-aware partitioning in order to improve the load-balancing of the whole coupled simulation.

Let us consider two coupled codes, modeled by two graphs (or hypergraphs) $A$ and $B$, connected by inter-edges $I(A, B)$ that represents the coupling communications between codes. Formally, the problem consists in partitioning $A$ in $M$ and $B$ in $N$ with accounting for $I(A, B)$. This algorithm should optimize both the edge cut for each graph and the coupling communications while maintaining each graph balance. Our general strategy is divided in three main steps:

1. first, we freely partition $A$ in $M$ to obtain the partition $A/M$;
2. then, we projects this partition to $B$ according to $I(A, B)$, that provides the partition $B/M$;
3. finally, we compute the partition $B/N$ by repartitioning $B$ from $M$ existing parts into $N$.

The final repartitioning step is particularly tedious, because it must handle a variable number of processes. However, as far as we know, the state-of-the-art graph/hypergraph repartitioning tools are limited to a fixed number of processes (i.e. $M = N$). To overcome this issue, we have proposed a new repartitioning algorithm – assuming the load is constant – based on hypergraph partitioning technics with fixed vertices. Our algorithm uses an *optimal* communication pattern, that we have proved to minimize the total number of messages between the former and newer parts. Experimental results validate our work comparing it with other approaches [20]. We currently investigate how to extend our algorithm for the dynamic load-balancing of parallel adaptive codes ($A = B$), whose load evolution is variable and difficult to predict. In this case, it would be convenient to dynamically adapt the number of processes used at runtime ($M \neq N$), while minimizing migration cost during the repartitioning step. This work is currently conducted in the framework of Clément Vuchener PhD thesis.

### 6.3. Distributed Shared Memory approach for the steering of parallel simulations
As a different approach of EPSN, we conceived and developed an in-transit visualization framework for interfacing an arbitrary HPC simulation code with an interactive ParaView session using the HDF5 parallel IO library as the API. The library called H5FDdsm is coupled with a ParaView plugin ICARUS (Initialize Compute Analyze Render Update Steer).

Because our interface is based on files, stored in a distributed shared memory (DSM), we sought during this year different redistribution strategies to optimize the bandwidth and the transfers between the simulation and the ParaView servers hosting the DSM. This work showed real benefits, particularly on one of our Cray XE6 testing machines using a block cyclic redistribution. On these large HPC machines that do not support the dynamic MPI process management set of functions, we improved our connection system so that simulation and post-processing can be coupled within an MPMD job. Taking also advantage of one-sided communication models and of the Cray Gemini interconnect communication performance, our framework has been sensibly improved and should be optimal in the coming months.

The interface has also been enhanced with a steering interface that allows us to control the simulation workflow and send back not only parameters, but also complete meshes in parallel, which can then be read by the simulation using either our steering interface or HDF5 calls. This has been demonstrated with SPH-flow, a CFD code developed by Ecole Centrale de Nantes and HydrOcean, replacing dynamically and in parallel a falling wedge with a deforming sphere.

This work has been realized and is currently carried on at CSCS - Swiss National Supercomputing Centre in the framework of Jérôme Soumagne PhD thesis (under the co-supervision of Mr. John Biddiscombe) and within the NextMuSE European project 7th FWP/ICT-2007.8.0 ([17], [18], [19]).

6.4. Material physics

6.4.1. Hybrid materials

The study of hybrid materials based on a coupling between molecular dynamics (MD) and quantum mechanism (QM) simulation has been conducted in collaboration with IPREM (Pau) within the ANR CIS 2007 NOSSI (ended December 2011). These simulations are complex and costly and may involve several length scales, quantum effects, components of different kinds (mineral-organic, hydro-philic and -phobic parts). Our goal was to compute dynamical properties of hybrid materials like optical spectra. The computation of optical spectra of molecules and solids is the most consuming time in such coupling. This requires new methods designed for predicting excited states and new algorithms for implementing them. Several tracks have been investigated in the project and new results obtained as described below.

Optical spectra.

Some new improvements in our TD-DFT code have been introduced. Our method is based on the LCAO method for densities and excited states that computes electronic excitation spectra. We have worked in two directions:

- As the method introduces a regularization parameter to obtain regularized spectra we have used it to build better algorithms. In particular, we have developed a new hierarchical algorithm that builds a well adapted frequency distribution to better capture the biggest peaks (strongest oscillator strengths) in the spectrum. Moreover, a nonlinear fit method was added and used to compute the transitions and the oscillator strengths of the spectrum.

- In our algorithm, we used a coarse grain paradigm to parallelize the spectrum computation. This approach leads to a memory bottleneck for large systems. In that respect, we have explored a new parallel approach based on a fine grain paradigm (matrix-vector parallelization) to better exploit the manycore architecture of the emerging computers.

Finally, we have improved the packaging of the code to prepare a public release of the code. Our TD-DFT code will be soon available on request.
QM/MM algorithm. For structure studies or dynamical properties, we have coupled QM model based on pseudo-potentials (SIESTA code) with dynamic molecular (DL-POLY code). Therefore we have developed a new algorithm to avoid accounting twice for the forces and the quantum electric field in the molecular model. All algorithms involved in the coupling have been introduced both in SIESTA and in DL-POLY codes. The following new developments needed by the coupling have been introduced in the SIESTA code:

- We have implemented a fast evaluation of the molecular electrostatic field on the quantum grid.
- We have introduced a non periodic Poisson solver based on the parallel linear Hypre solver. This solver allows us to use computation domains as small as possible.
- We have implemented the ElectroStatic Potential (ESP) fit method to obtain more physical point charges than those given by SIESTA with the Mulliken method. These point charges are used by the MM codes to compute electrostatic forces.

Thanks to all our developments introduced in SIESTA a collaboration with the SIESTA research team has started. This enables us to have access to their private svn like repository. Preliminary results on a water dimer and a water box systems show good agreement with other methods developed in SIESTA and DL-POLY teams. All these results were presented in the final international NOSSI workshop in Biarritz on December.

6.4.2. Material failures

We have started in the context of the OPTIDIS ANR to work on dislocation simulations. The main characteristic of these simulations is that they are highly dynamical. This year, we have started the study of the state of the art on this topic in two directions. The first direction concerns the study of the algorithms used in such simulations and how we can efficiently parallize them on manycore clusters. In the second one for isotropic materials, we are investigating how to adapt our fast multipole method to compute constraints and then forces in this kind of simulations.
6. New Results

6.1. BlobSeer and Map-Reduce programming

6.1.1. BlobSeer-based cloud storage

Participants: Alexandra Carpen-Amarie, Alexandru Costan, Gabriel Antoniu, Luc Bougé.

As data volumes generated and processed by such applications increase, a key requirement that directly impacts the adoption rate of the Cloud paradigm is efficient and reliable data management. In this context, we investigate the requirements of Cloud data services in terms of data-transfer performance and access patterns and we explore the ways to leverage and adapt existing data-management solutions for Cloud workloads. We aim at building a Cloud data service both compatible with state-of-the-art Cloud interfaces and able to deliver high-throughput data storage.

To achieve this goal, we developed a file system layer on top of BlobSeer, which exposes a hierarchical file namespace enhanced with the concurrency-optimized BlobSeer primitives. Furthermore, we integrated the BlobSeer file system as a backend for Cumulus, an efficient open-source Cloud storage service. We validated our approach through extensive evaluations performed on Grid’5000. We devised a set of synthetic benchmarks to measure the performance and scalability of the Cumulus system backed by BlobSeer, showing it can sustain high-throughput data transfers for up to 200 concurrent clients.

Next, we explored the advantages and drawbacks of employing Cloud storage services for distributed applications that manage massive amounts of data. We investigated two types of applications. We relied on an atmospheric phenomena modeling application to conduct a set of evaluations in a Nimbus Cloud environment. This application is representative for a large class of simulators that compute the evolution in time set of parameters corresponding to specific points in a spatial domain. As a consequence, such applications generate important amounts of output data. We evaluated an S3-compliant Cloud storage service as a storage solution for the generated data. To this end, we employed distributed Cumulus services backed by various storage systems. The reason for targeting this approach is that storing output data directly into the Cloud as the application progresses can benefit higher-level applications that further process such simulation data. As an example, visualization tools need to have real-time access to output data for analysis and filtering purposes.

We built an interfacing module to enable the application to run unmodified in a Cloud environment and to send output data to an S3-based Cloud service. Our experiments show that distributed Cumulus backends, such as BlobSeer or PVFS, sustain a constant throughput even when the number of application processes that concurrently generate data becomes 3 times higher than the number of storage nodes.

6.1.2. Optimizing Intermediate Data Management in MapReduce Computations

Participants: Diana Moise, Gabriel Antoniu, Luc Bougé.

MapReduce applications, as well as other cloud data flows, consist of multiple stages of computations that process the input data and output the result. At each stage, the computation produces intermediate data that is to be processed by the next computing stage. We studied the characteristics of intermediate data in general, and we focused on the way it is handled in MapReduce frameworks. Our work addressed intermediate data at two levels: inside the same MapReduce job, and during the execution of pipeline applications.
We focused first on efficiently managing intermediate data generated between the “map” and “reduce” phases of MapReduce computations. In this context, we proposed to store the intermediate data in the distributed file system used as underlying backend. In this direction, we investigated the features of intermediate data in MapReduce computations and we proposed a new approach consisting in storing this kind of data in a DFS. The major benefit of this approach is better illustrated when considering failures. Existing MapReduce frameworks store intermediate data on nodes local disk. In case of failures, intermediate data produced by mappers can no longer be retrieved and processed further by reducers. The solution of most frameworks is to reschedule the failed tasks and to re-generate all the intermediate data that was lost because of failures. This solution is costly in terms of additional execution time. With our approach of storing intermediate data in a DFS, we avoid the re-execution of tasks in case of failures that lead to data loss. As storage for intermediate data, we considered BSFS as being a suitable candidate for providing for the requirements of intermediate data: availability and high I/O access. The tests we performed in this context, measured the impact of using a DFS as storage for intermediate data instead of the local-disk approach. We then assessed the performance of BSFS and HDFS when serving as storage for intermediate data produced by several MapReduce applications.

We then considered another type of intermediate data that appears in the context of pipeline MapReduce applications. In order to speed-up the execution of pipeline MapReduce applications (applications that consist of multiple jobs executed in a pipeline) and also, to improve cluster utilization, we proposed an optimized Hadoop MapReduce framework, in which the scheduling is done in a dynamic manner. We introduced several optimizations in the Hadoop MapReduce framework in order to improve its performance when executing pipelines. Our proposal consisted mainly in a new mechanism for creating tasks along the pipeline, as soon as the tasks’ input data becomes available. As our evaluation showed, this dynamic task scheduling leads to an improved performance of the framework, in terms of job completion time. In addition, our approach ensures a more efficient cluster utilization, with respect to the amount of resources that are involved in the computation.

We evaluated both approaches for intermediate data through a set of experiments on the Grid’5000 [56] testbed. Preliminary results [17] show the scalability and efficiency of our proposals, as well as additional benefits brought forward by our approach.

6.1.3. A-Brain: Perform genetic and neuroimaging data analysis in Azure clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu, Louis-Claude Canon.

Joint genetic and neuroimaging data analysis on large cohorts of subjects is a new approach used to assess and understand the variability that exists between individuals. This approach has remained poorly understood so far and brings forward very significant challenges, as progress in this field can open pioneering directions in biology and medicine. As both neuroimaging- and genetic-domain observations represent a huge amount of variables (of the order of $10^6$), performing statistically rigorous analyses on such amounts of data represents a computational challenge that cannot be addressed with conventional computational techniques.

In order to perform an accurate analysis we need to provide a programming platform and a high throughput storage. The target infrastructure is the Azure clouds. Hence we have adapted the BlobSeer storage approach for Azure, thus providing a new way to store data in clouds, that federates the local storage space from computational nodes into a uniform shared storage, called TomusBlobs. Using this storage system as a storage backend, we have built a MapReduce prototype for Azure clouds. This MapReduce system, called TomusMapReduce -TMR, is used to perform the simulation of the joint genetic and neuroimaging application. For validating the framework, a toy application that simulate the data access and computation patterns of the real application, was used. The next step, after the evaluation of the framework, that has just began, consists in replacing the toy application with the real one and the scaling of the framework in the limit allowed by the cloud provider. In addition a demo for this project is in progress, that will consists in providing a visualization tool for the framework. This will be used to intuitively represent the results for the simulation of the scientific application, this being useful both for better presenting the project to interested parties and for the researchers from bioinformatics.

6.2. Efficient VM management in clouds

Participants: Alexandru Costan, Alexandra Carpen-Amarie, Gabriel Antoniu.
Infrastructure as a Service (IaaS) cloud computing allows users to lease computational resources from the cloud provider’s datacenter for a short time by deploying virtual machines (VMs) on these resources. This model raises new challenges in the design and development of IaaS middleware. One of those challenges is the need to deploy a large number (hundreds or even thousands) of VM instances simultaneously. Once the VM instances are deployed, another challenge is to simultaneously take a snapshot of many images and transfer them to persistent storage to support fault tolerance and management tasks, such as suspend-resume and migration. With datacenters growing rapidly and configurations becoming heterogeneous, it is important to enable efficient concurrent deployment and snapshotting that are at the same time hypervisor independent and ensure a maximum compatibility with different configurations.

We addressed these challenges by proposing a virtual file system specifically optimized for virtual machine image storage \[19\]. It is based on a lazy transfer scheme coupled with object versioning that handles snapshotting transparently in a hypervisor-independent fashion, ensuring high portability for different configurations. Large-scale experiments on hundreds of nodes demonstrate excellent performance results: speedup for concurrent VM deployments ranges from a factor of 2 up to 25, with a reduction in bandwidth utilization of as much as 90 \% \[18\]. We implemented this deployment scheme in the Nimbus cloud and presented a demo illustrating it at the Grid’5000 School \[26\].

Given the dynamic nature of IaaS clouds and the long runtime and resource utilization of scientific applications, an interesting use-case for the multi-snapshotting techniques is for efficient checkpoint-restart. We introduced an approach that leverages VM disk-image multi-snapshotting and multi-deployment inside checkpoint-restart protocols running at guest level in order to efficiently capture and potentially roll back the complete state of the application, including file system modifications. This framework is specifically optimized for tightly-coupled scientific applications that were written using a message passing system (in particular MPI) and need to be ported to IaaS clouds. Our solution introduces a dedicated checkpoint repository that is able to efficiently take incremental snapshots of the whole disk attached to the virtual machine instances, thus offering support to use any checkpointing protocol that can save the state of processes into files, including application level mechanisms, where the process state is managed by the application itself, and process-level mechanisms, where the process state is managed transparently at the level of the message passing library. Experiments on the G5K testbed show substantial improvement for MPI applications over existing approaches, both for the case when customized checkpointing is available at application level and the case when it needs to be handled at process level.

We integrated this checkpointing scheme inside the Nimbus cloud with some promising preliminary results. We plan to complement the existing solution with live incremental snapshotting using asynchronous background transfers for high checkpointing efficiency and with adaptive prefetching to achieve high restart efficiency.

### 6.3. Cloud data storage management

#### 6.3.1. Autonomic storage for cloud services

**Participants:** Alexandru Costan, Alexandra Carpen-Amarie, Gabriel Antoniu, Florin Pop, Ciprian Dobre, Elena Apostol.

A means to achieve performance improvement and resource-usage optimization in cloud storage systems consists in enabling an autonomic behavior based on introspection. Self-adaptation incurs a high degree of complexity in the configuration and tuning of the system, with possible repercussions on its availability and reliability. To address these challenges we introduced in BlobSeer in \[11\] a three-layered architecture designed to identify and generate relevant information related to the state and the behavior of the system, based on the MonALISA monitoring framework. Such information is then expected to serve as an input to a higher-level self adaptation engine. These data are yielded by an (1) introspection layer, which processes the raw data collected by a (2) monitoring layer. The lowest layer is represented by the (3) instrumentation code that enables BlobSeer to send monitoring data to the upper layers.
A first approach to leverage the introspection framework aims at enhancing BlobSeer with *self-configuration* capabilities, as a means to support storage elasticity through dynamic deployment of data providers. This solution enables the data providers to scale up and down depending on the detected system’s needs. The component we designed adapts the storage system to the environment by contracting and expanding the pool of storage providers based on the system’s load. The key idea of this component is the automatic decision that has to be made on how many resources the system needs to operate normally while keeping the resources utilization down to a minimum. This problem is addressed by using a test-decided heuristic based on the monitoring data. The introspective architecture has been evaluated on the Grid’5000 testbed, with experiments that prove the feasibility of generating relevant information related to the state and the behavior of the system.

We plan to use the introspective BlobSeer to develop a distributed data aggregation service. Its primary goal will be to serve as a repository backend for complex analysis and automatic mining of scientific data. Another direction that will be explored is to use the introspective BlobSeer as a cloud-based storage layer for sensitive context data, collected from a vast amount of sources: from smartphones to sensors located in the environment. Clouds are perfect candidates to handle the storage and aggregation of such data for even larger context-aware applications. Such solutions rely on more relaxed storage capabilities than traditional relational databases (eventual consistency suffices for example). This, combined with the high concurrency support and the flexible storage schema make BlobSeer a suitable candidate for the storage layer. We plan to develop a new layer on top of BlobSeer targeting context aware applications. At the logical level, this layer will provide transparency, mobility, real-time guarantees and access based on meta-information. At the physical layer, the most important capability will rely on BlobSeer’s elasticity to scale up and down according to real-time usage, in order to reduce the costs within the Cloud.

### 6.3.2. Managing data access on Clouds through security policies

**Participants:** Alexandru Costan, Alexandra Carpen-Amarie, Gabriel Antoniu.

With the emergence of Cloud computing, there has been a great need to provide an adequate security level in such environments, as they are vulnerable to various attacks. Malicious behaviors such as Denial of Service attacks, especially when targeting large-scale data management systems, cannot be detected by typical authentication mechanisms and are responsible for drastically degrading the overall performance of such systems.

In [14] we proposed a generic security management framework allowing providers of Cloud data management systems to define and enforce complex security policies. The generality of this approach comes from the flexibility both in terms of supporting custom security scenarios and interfacing with different Cloud storage systems. This security framework is designed to detect and stop a large array of attacks defined through an expressive policy description language and to be easily interfaced with various data management systems. We introduced a modular architecture consisting of three components. The **Policy Management** module represents the core of the framework, where security policies definition and enforcement takes place. This module is completely independent of the Cloud system, as its input only consists in user activity events monitored from the system. The **User Activity History** module is a container for monitoring information describing users’ actions. It collects data by employing monitoring mechanisms specific to each storage system and makes them available for the Policy Management module. The **Trust Management** module incorporates data about the state of the Cloud system and provides a trust value for each user based on his past actions. The trust value identifies a user as a fair or a malicious one. Furthermore, the trust values enable the system to take custom actions for each detected policy violation, by taking into account the history of each user.

As a case study, we applied the proposed framework to BlobSeer. We defined a specific policy to detect DoS attacks in BlobSeer and we evaluated the performance of our framework through large scale experiments on the Grid’5000 testbed. The results show that the Policy Management module meets the requirements of a data storage system in a large-scale deployment: it was able to deal with a large number of simultaneous attacks and to restore and preserve the performance of the target system.

As a next step we will focus on more in-depth experiments involving the detection of various types of attacks in the same time. Moreover, we will investigate the limitations of our Security Management framework, with
respect to the accuracy of the detection in the case of more complex policies, as well as the probability and
the impact of obtaining false positive or false negative results. Another research direction is to further develop
the Trust Management component of the security management framework and study the impact it has on the
Policy Enforcement decisions for complex scenarios.

6.4. Storage architecture and adaptive consistency for clouds

Participants: Houssem-Eddine Chihoub, Gabriel Antoniu.

As more and more applications are becoming data-intensive, the design of a scalable storage architecture
providing a huge file sharing and fine grain access with high throughput under heavy concurrency is a timely
and relevant challenge.

In [24], we introduce a storage architecture for Cloud computing. This architecture proposes efficient and
scalable storage support for both VM images and application data. Our architecture relies on BlobSeer [12], a data sharing platform optimized for concurrent accesses, as a basic storage backend enhanced in term of quality of service and efficiency by GloBeM tool [61] that rely on behavior modeling and monitoring to
avoid bad case scenarios. The architecture uses this approach to have a better and efficient storage VM images,
that allow a faster image deployment and efficient versioning and snapshotting. Furthermore, the architecture
provides a platform to store, manage, and share cloud application data allowing several key features to clouds
such as storage elasticity.

The main aim is to provide high availability and good scalable performance at low cost. This is justified by the
growing need of data-intensive applications for managing huge sets of data replicated over several data centers.
In order to provide good performance and high availability, data replication is mandatory. But this generates
the issue of replicas consistency as shown by the CAP theorem [52]. To achieve the aforementioned goals,
relaxing consistency rules is unavoidable. On the other hand, opting for weaker consistency, all the time, can
be too costly.

In current work we leverage the trade-off between consistency and availability and performance. We are
investigating an approach that changes the consistency level at runtime considering system and application
needs. In order to choose the most suitable consistency level, our approach monitor the storage system and
collect useful information, such as network load and applications access patterns, that enable the system to
estimate the amount of expected stale reads.

6.5. Modelling cloud storage performance

Participants: Daniel Higuero, Louis-Claude Canon, Alexandru Costan, Gabriel Antoniu.

The objective of this research direction is to provide comprehensive performance models for storage systems.
Their role is to capture how the system components interact for different usage patterns (number of reads or
writes). The objective is to determine the incurred costs in terms of storage space and efficiency for a given
workload.

One application of this model consists in dynamically adjusting the parameters of the storage system as
required in an autonomic approach. For this purpose, it is necessary to identify the characteristics of the
storage system for meeting a given level of requirements. Progress has been made on this part during the
3-month visit of Daniel Higuero (University Carlos III, Madrid). A preliminary performance model currently
predicts the available bandwidth when multiple concurrent transfers occur. This model serves as a basis for a
dimensioning strategy that is formulated through a linear program.

This approach has further been complemented with an offline analysis of several traces of the BlobSeer storage
system when it is used as a backend for MapReduce applications. Mining this information in an automated
fashion allowed to detect the different trade-offs that influence a BlobSeer deployment: time required to
execute the application vs. number of machines used by the storage system, communication costs vs. space
usage. The final goal is to tune BlobSeer for specific applications. The proposed strategy is currently being
evaluated.
Future directions are directed towards refining the proposed model. Several parameters significantly impact the performance of storage systems such as the redundancy mechanism, the data placement strategy or disk-related effects. As a first step, experiments for assessing the quality of finer models will be designed. Ultimately, we aim to capture the I/O variability of storage systems, in particular in the context of the cloud.

This work will enable new collaborations. It is planned to work on the models mentioned above with the Mescal INRIA team in the context of a collaboration between the MapReduce ANR project and the Songs ANR project. Moreover, in the framework of the MapReduce project, we expect to work on a performance model for designing decision algorithms that are required by the component-based MapReduce framework that is developed in the GRAAL/Avalon INRIA team. Finally, the GRAAL/AVALON team works on scheduling algorithms that could beneficially profit from a storage performance model.

### 6.6. Scalable I/O and visualization for post-petascale HPC simulations

**Participants:** Matthieu Dorier, Gabriel Antoniu.

In the context of the Joint INRIA/UIUC Laboratory for Petascale computing (JLCP), we are addressing the new challenges related to I/O, data analysis and visualization for extreme-scale simulations. As HPC resources approaching millions of cores become a reality, a growing challenge in maintaining high performance is the presence of high variability in the effective throughput of codes performing I/O operations. Since I/O is mainly performed for the purpose of subsequent data analysis and visualization, another way to limit the impact of I/O performance on scientific discovery consists in enabling in-situ visualization. This brings again new challenges such as how to efficiently couple large-scale simulations with visualization software.

We started the development of Damaris, a middleware targeting multicore SMP nodes to efficiently address the problems mentioned above. Damaris has been evaluated with the CM1 atmospheric simulation [43], one of the targeted application for the BlueWaters project. To show the capability of our approach to efficiently hide I/O jitter and related costs, experiments have been carried on the French Grid’5000, on a Power5 cluster at NCSA and on the Kraken Cray XT5 supercomputer (currently 11th in the Top500) with up to 9K cores. By gathering data into large files while avoiding synchronization between processes, our solution brings several benefits:

1. it increases the sustained write throughput of the simulation by a factor of almost 15;
2. it provides almost 70% overall application speedup on 9,000 cores;
3. it fully hides I/O-related costs;
4. it enables a 600% compression ratio without any additional overhead, leading to a major reduction of storage requirements.

A poster [16] presenting some of these results has been accepted at ICS’11 and awarded the second price at the ACM Student Research Competition. All these results are presented in a research report [25] pending for publication.

Current work addresses the efficient coupling of large-scale simulations and visualization tools through Damaris. We have been able to get access to the Jaguar supercomputer hosted at ORNL and we are planning very-large scale experiments on up to 100,000 cores to show the benefits of our Damaris approach.

### 6.7. Scalable array-oriented active storage

**Participants:** Viet-Trung Tran, Gabriel Antoniu, Luc Bougé.
The recent explosion in data sizes manipulated by distributed scientific applications has prompted the need to develop specialized storage systems capable to deal with specific access patterns in a scalable fashion. In this context, a large class of applications focuses on parallel array processing: small parts of huge multi-dimensional arrays are concurrently accessed by a large number of clients, both for reading and writing. However, many established storage solutions such as parallel file systems and database management systems expose data access models (e.g., file systems, structured databases) that are too general and do not exactly match the nature requirements of the application. This forces the application developer to either adapt to the exposed data access model or to use an intermediate layer that performs a translation. In either case, the mismatch leads to suboptimal data management: the one-storage-solution-fits-all-needs has reached its limits.

Thus, there is an increasing need to specialize the I/O stack to match the requirements of the application. The objective of this research is to design Pyramid: an array-oriented active storage system optimizing for applications that represent and manipulate data as huge multi-dimensional arrays. However, a specialized storage system that deals with such an access pattern faces several challenges at the level of data/metadata management, we carefully design the system with the following principles: (1) we introduce a dedicated array-oriented data access model that offers support for active storage and versioning; (2) we enrich striping techniques specifically optimized for multi-dimensional arrays with a distributed metadata management scheme that avoids potential I/O bottlenecks observed with centralized approaches.

We evaluated Pyramid through a set of experiments on the Grid’5000 [56] testbed that aims to evaluate both the performance and the scalability of our approach under concurrent accesses. Preliminary evaluation in our recent papers [22], [13] shows promising results: our prototype demonstrates good performance and scalability under concurrency, both for read and write workloads.
6. New Results

6.1. Perfect simulation

We have proposed a new approach for sampling the stationary distribution of general Markov chains that only needs to consider two trajectories. We show that this new approach is particularly effective when the state space can be partitioned into pieces where envelopes can be easily computed [26]. We further show that most Markovian queuing networks have this property and we propose efficient algorithms for some of them, in particular when the rates of events range over several orders of magnitude [45]. We also provided a novel approach for efficient sampling of queues with phase type servers [37] (this paper has received the best paper award at ASMTA 2011) and Markov chains with infinite state spaces (but with a known bounding process). Perfect sampling has been used for model checking of probabilistic models in [14].

6.2. Economic models for clouds

Recently introduced spot instances in the Amazon Elastic Compute Cloud (EC2) offer low resource costs in exchange for reduced reliability; these instances can be revoked abruptly due to price and demand fluctuations. Mechanisms and tools that deal with the cost-reliability trade-offs under this scheme are of great value for users seeking to lessen their costs while maintaining high reliability. We study how mechanisms, namely, checkpointing and migration, can be used to minimize the cost and volatility of resource provisioning. Based on the real price history of EC2 spot instances, we compare several adaptive checkpointing schemes in terms of monetary costs and improvement of job completion times. We evaluate schemes that apply predictive methods for spot prices. Furthermore, we also study how work migration can improve task completion in the midst of failures while maintaining low monetary costs. Trace-based simulations show that our schemes can reduce significantly both monetary costs and task completion times of computation on spot instance [25].

6.3. Game theory and networks

We studied the traffic routing problem in networks whose users try to minimize their latencies by employing a distributed learning rule inspired by the replicator dynamics of evolutionary game theory. The stable states of these dynamics coincide with the network’s (Wardrop) equilibrium points. Despite this abundance of stable states, we find that (almost) every solution trajectory converges to an equilibrium point at an exponential rate. When network latencies fluctuate unpredictably we show that the time-average of the traffic flows of sufficiently patient users is still concentrated in a neighborhood of evolutionarily stable equilibria and we estimate the corresponding stationary distribution and convergence times [42].

We also analyzed the distributed power allocation problem in parallel multiple access channels (MAC) by studying an associated non-cooperative game which admits an exact potential function. We show that the parallel MAC game admits a unique equilibrium almost surely. Furthermore, if the network’s users employ a distributed learning scheme based on the replicator dynamics, we show that they converge to equilibrium from almost any initial condition, even though users only have local information at their disposal [41].

Using a large deviations approach we calculate the probability distribution of the mutual information of MIMO channels in the limit of large antenna numbers. We calculate the full distribution, including its tails which strongly deviate from the Gaussian behavior near the mean. This calculation provides us with a tool to obtain outage probabilities analytically at any point in the parameter space, as long as the number of antennas is not too small [20].
6.4. Mean field analysis for networks

We have studied the deterministic limits of Markov processes made of several interacting objects. While most classical results assume that the limiting dynamics has Lipschitz properties, we show that these conditions are not necessary to prove convergence to a deterministic system.

We show that under mild assumptions, the stochastic system converges to the set of solutions of a differential inclusion and we provide simple way to compute the limiting inclusion. When this differential inclusion satisfies a one-sided Lipschitz condition, there exists a unique solution of this differential inclusion and we show convergence in probability with explicit bounds.

This extends the applicability of mean field techniques to systems exhibiting threshold dynamics such as queuing systems with boundary conditions or controlled dynamics. This is illustrated by applying our results to several types of systems: fluid limits of priority queues, best response dynamics in games, push-pull queues with a large number of sources and a large number of servers and self-adapting computing systems [65].

6.5. Idleness and failure prediction in large infrastructures

We have proposed a method to discover statistical models of availability in large distributed systems and applied it to run an enlightening study of SETI@home [19]. This was also used to make long-term availability predictions for groups of desktop grid resources [21]. We have used statistically based models of heterogeneous failures in parallel systems and assessed their tolerance [39]. A similar approach was used to design correlated resource models of Internet end hosts [38], [17].

6.6. Scheduling and Game Theory

A stochastic model of failures has been used to optimize the scheduling of checkpoints on desktop grids [28].

We have also shown that non-cooperative scheduling can be considered harmful in collaborative volunteer computing environments [33].

Optimal scheduling and route selection have been investigated using a novel approach based on Lagrangian optimization. This result is inspired from flow control in multi-path networks and was used for multiple mag- of-tasks application scheduling on grids [61].

In the similar context of broker-based networks of non-observable parallel queues, we provide lower bounds on the minimum response time. We introduce the “Price of Forgetting” (PoF), the ratio between the minimum response times achieved by a probabilistic broker and a broker with memory, that is shown to be unbounded or arbitrarily close to one depending on the coefficient of variation of the service time distributions. We also put our results in the context of game theory revisiting the “Price of Anarchy” (PoA) of parallel queues: It can be decomposed into the product of the PoA achieved by a probabilistic broker (already well understood) and the PoF [10].

6.7. Validity study of flow-based network models.

Researchers in the area of distributed computing conduct many of their experiments in simulation. While packet-level simulation is often used to study network protocols, it can be too costly to simulate network communications for large-scale systems and applications. The alternative chosen in SimGrid and a few other simulation frameworks is to simulate the network based on less costly flow-level models. Surprisingly, in the literature, validation of these flow-level models is at best a mere verification for a few simple cases. Consequently, although distributed computing simulators are widely used, their ability to produce scientifically meaningful results is in doubt. In [9], [70] we focus on the validation of state-of-the-art flow-level network models of TCP communication, via comparison to packet-level simulation. While it is straightforward to show cases in which previously proposed models lead to good results, instead we systematically seek cases that lead to invalid results. Careful analysis of these cases reveal fundamental flaws and also suggest improvements. One
contribution of this work is that these improvements lead to a new model that, while far from being perfect, improves upon all previously proposed models. A more important contribution, perhaps, is provided by the pitfalls and unexpected behaviors encountered in this work, leading to a number of enlightening lessons. In particular, this work shows that model validation cannot be achieved solely by exhibiting (possibly many) “good cases.” Confidence in the quality of a model can only be strengthened through an invalidation approach that attempts to prove the model wrong.

6.8. Visualization

We have proposed a methodology for detecting resource usage anomalies in large scale distributed systems. The methodology relies on four functionalities: characterized trace collection, multi-scale data aggregation, specifically tailored user interaction techniques, and visualization techniques. We have shown the efficiency of this approach through the analysis of simulations of the volunteer computing Berkeley Open Infrastructure for Network Computing architecture (BOINC). Three scenarios have been analyzed in [48], [23]: analysis of the resource sharing mechanism, resource usage considering response time instead of throughput, and the evaluation of input file size on Berkeley Open Infrastructure for Network Computing architecture. The results show that our methodology enables to easily identify resource usage anomalies, such as unfair resource sharing, contention, moving network bottlenecks, and harmful short-term resource sharing. Triva, the resulting software, has been demonstrated at the SuperComputing conference.

We also have investigated how to use trace-based visualization to understand applications I/O performance [49] and how to visually compare two traces [70] and highlight differences.

6.9. Experimental methodology

In the scientific experimentation process, an experiment result needs to be analyzed and compared with several others, potentially obtained in different conditions. Several tools are dedicated to the control of the experiment input parameters and the experiment replay. In parallel, concurrent and distributed systems, experiment conditions are not only restricted to the input parameters, but also to the software environment in which the experiment was carried out. It is therefore essential to be able to reconstruct this type of environment. This can quickly become complex for experimenters, particularly on research platforms dedicated to scientific experimentation, where both hardware and software are in constant rapid evolution. We study the concept of the reconstructability of software environments and propose a tool for dealing with this problem in [64].

We have also started investigating the systematic use of Design of Experiments to computer studies (see [61]). Nonetheless such approach provides results that are much more trustworthy than what is generally done in the parallel and distributed computing community but it also enables to shorten the experiments cycle and to use less computing resources.

6.10. Multi-core platforms

We have used memory access traces to map threads on hierarchical multi-core platforms [13]. We have also used software transactional memory to analyze and trace applications running on multi-core architectures [30].

An approach based on machine learning was used to map threads on transactional memory applications in [31].

The impact of CPU and memory affinity on multi-core platforms was investigated in [46] using numerical scientific multi-threaded applications as a typical case study. This resulted in improvement of the performance of parallel systems using a NUMA-aware load balancer [68].

We have also carried a performance evaluation of WiNoCs for parallel workloads based on collective communications [43] as well as for Infiniband networks [40].
6.11. High performance computing

We have developed a runtime system, named SGPU 2, that enable large applications to run on clusters of hybrid nodes [44].

BigDFT is a parallel simulator of the matter at the nano scale. It uses Daubechies Wavelets for High Performance Electronic Structure Calculations [16]. This tool is shown to make efficient use of massive parallel hybrid architectures [57].

6.12. Input-Output

Atmospheric models usually demand high processing power and generate large amounts of data. As the degree of parallelism grows, the I/O operations may become the major impacting factor of their performance. In [27], we evaluate the Ocean-Land-Atmosphere Model (OLAM) on the PVFS file system in order to point the I/O characteristics of the application. We show that storing the files on PVFS has lower performance than using the local disks of the cluster nodes due to file creation and network concurrency. Additionally, we study the performance of a new version of OLAM that used MPI associated with OpenMP and show that the combined strategy presents I/O times 20 times shorter than the original MPI-only version and 9 times shorter on total execution time. Finally, a survey on I/O Characterization of several applications is given in [51].
6. New Results

6.1. Kaapi

New version of Kaapi, called X-Kaapi, has been released. The kernel is written in C for hypothetical required from embedded system. On top of the kernel, several APIs co-exist: a template based C++ library called Kaapi++; a C API; a Fortran API; and a compiler that transform a source code annotated with pragma directive to a source code with calls to the runtime library function. The compiler works with C and a subset of C++.

http://kaapi.gforge.inria.fr

6.2. Multi-criteria optimization

The main idea is the development of a methodology to generate a reasonable set of approximated Pareto’ solutions (closed to the best achievable solutions). Especially, we have applied this methodology to better take into account users’ criteria than the other existing methods offer. We have also studied the problem of selection of best algorithms in a portfolio. This research axis is currently enforced by the INRIA postdoc position of Joachim Lepping where we have started to include a learning process to select the best algorithm on a given instance.

6.3. Stochastic models for optimizing checkpoint protocol

After our past studied on design of origin checkpoint protocols, we have proposed a new stochastic performance model of the parallel execution in presence of failures. Thanks to this formulation, we are able to optimize several criteria (the time lost due to failure; the expected completion time) by making right decision of the date of each checkpoint. The model is general and it does not take into account the failure distribution law and accept variable checkpoint time estimation, which is important for dynamic parallelism applications.

6.4. Work stealing scheduling algorithm taking care of communication

On some applications, the amount of data transfers can be high. To minimize the amount of data transfers during the execution, Jean-Noel Quintin has developed an algorithm called WSCOM which uses the DAG structure of the application. For each steal request, the work-stealing algorithm tries to balance the load between the thief and the stolen processor. Thus, WSCOM tries to divide the work on the stolen processor into two parts with a small number of edges between the two parts. This cutting is done with a negligible overhead at each steal request. This algorithm has been implemented in a tool called DSMake. This tool executes the set of tasks described by a Makefile on a distributed platform. In addition, I have developed a simulator to validate algorithm performance and its behavior. We compared WSCOM and several static list-scheduling algorithms. The comparison shows that WSCOM outperforms list-scheduling algorithms, on clusters with some network congestion.

Besides, based on SIPS analysis of work stealing, Stefano Mor in his thesis compared the influence of the choice of the stolen tasks on the number of steal operations, distinguishing unsuccessful and successful steals. While standard bounds are related to unsuccessful steals, they are pessimistic with respect to the number of successful steals that define intensive data communications.
6.5. Homomorphic coding for soft error resilience

We extended our results for fault-tolerant modular computations in two directions. To improve the correction rate of Reed-Solomon codes, power-decoding techniques consist in augmenting the number of syndrom equations by raising the received word to successive powers. The correction is done by a generalization of Berlekamp-Massey algorithm acting on multiple sequences. This method is, if not equivalent, at least very close to the list-decoding proposed by Sudan in its first version, in particular, error correction rates are identical. We improve the power-decoding method by reformulation into a vector rational function reconstruction, with benefit from fast polynomial matrix arithmetic. Besides, for basic exact linear algebra computations (eg dense linear system), we designed interactive protocols between a trusted platform and a non trusted one for resilience to soft-errors.

6.6. Chimeric algorithms design

To reach provable multicriteria performance, we used the coupling of various algorithms that adapt in several contexts: recursive cascading of both sequential and parallel algorithms with work-stealing; coupling specific algorithms on heterogeneous platforms (eg CPU/GPU); interactive distributed computations; fault-tolerant computations by coupling both a trustfully platform with low computation bandwidth and and an unreliable computing platform with high bandwidth. A unification work is currently developed for the design of a chimeric algorithms that is composed of the parts of multiple algorithms, interactively cascaded to achieve provable multicriteria performance.
RUNTIME Project-Team

6. New Results

6.1. High-Performance Intra-node Collective Operations

Participants: Brice Goglin, Stéphanie Moreaud.

- KNEM is known to improve the performance of point-to-point intra-node MPI communication significantly [60], [18].
- We designed an extended RMA interface in KNEM that suits the needs of point-to-point, collective and RMA operations.
- We showed that the native use of KNEM in MPI collective implementations enabled further optimization by combining the knowledge of collective algorithms with the mastering of KNEM region management and copies [35].
- This work was initiated in the context of our collaboration with the MPICH2 team and is now also pursued within the OPEN MPI project in collaboration with the University of Tennessee in Knoxville.

6.2. I/O-Affinity-aware MPI Communications

Participants: Brice Goglin, Stéphanie Moreaud.

- We demonstrated in the past that the locality of I/O devices within modern computing nodes has the significant impact of the MPI communication performance [11] (Non-Uniform I/O Access, NUIOA).
- A first way to deal with such affinities would be to privilege I/O-intensive processes by placing them near the network interfaces. However, determining the communication-intensiveness may be tricky. Also, some applications have uniform communication patterns. The other way to deal with I/O affinities is to modify the implementation of communication operations given a predetermined task placement.
- We demonstrated that the implementation of collective operations should take I/O affinities into account. Deciding which steps and leaders should be involved in the algorithms based on NUIOA effects led us to improve broadcast performance significantly [34], [18].

6.3. High-Performance Point-to-Point Communications

Participants: Alexandre Denis, Raymond Namyst.

- NEWMADELEINE is our communication library designed for high performance networks in clusters. We have worked on optimizations on low-level protocols so as to improve point-to-point performance.
- We have proposed [29] auto-tuning mechanisms for most parameters of a communication library: rendez-vous threshold, multi-rail ratio, optimization strategies.
- We have proposed a communication protocol [33] for InfiniBand that completely amortizes the cost of memory registration, through the use of a superpipeline that overlaps communication and memory copies. We have modeled the behavior of the network and proposed auto-tuning mechanism to adapt the protocol to the hardware properties.
6.4. Improve code-coupling performance in the SALOME platform

**Participants:** Alexandre Denis, Sébastien Barascou.

- SALOME platform is an open source software developed by EDF, CEA, and OpenCascade. It is an open simulation platform with pre-processing, post-processing, interoperability with CAD models, integration with computation kernels.
- YACS is the workflow engine used for code coupling applications in SALOME. It leverages CORBA for communications between kernels. We have ported YACS atop PadicoTM, our communication platform for grids. It enables CORBA connections to use InfiniBand networks. Benchmarks show a significant improvement in code coupling performance.

6.5. Hardware topology-aware MPI applications

**Participants:** Emmanuel Jeannot, Guillaume Mercier, François Tessier.

- We have expanded our previous work dealing with MPI process placement. Indeed, our approach relied on tools and techniques which were outside the scope of the MPI standard itself. In order to allow the users to utilize our work in a portable way, we enhanced some routines of the MPI standard. We worked mainly with the MPICH2 implementation but we are also working on an OPEN MPI version as well.
- Instead of modifying the binding of the MPI processes onto the physical cores on the underlying architecture, we chose to create a new communicator for which the logical topology organization is optimized for the hardware. This work has been published in [37] and show interesting performance improvements for some class of MPI applications.
- The problem of process placement, which can be reduced to a NP-hard graph partitioning problem, can be dealt with several famous applications like Scotch or ParMETIS. To evaluate these solutions with TREEMATCH, we ran several benchmarks using NAS Parallel Benchmarks and a real CFD application. On the one hand we study the quality of processes permutation (which will impact the execution time) and on the other hand the computation time of the permutation. These results will allow us to conclude about the pertinence of what graph partitioner can be used to bind processes on process units or to do a dynamic processes reordering.

6.6. Mastering Heterogeneous Platforms

**Participants:** Cédric Augonnet, Olivier Aumage, Ludovic Courtès, Nathalie Furmento, Andra Hugo, Raymond Namyst, Samuel Thibault, Pierre-André Wacrenier.

- We continued our work on extending STARPU to master exploitation of Heterogeneous Platforms.
- We have extended the STARPU scheduler into managing **parallel tasks** which permit a better exploitation of CPUs and load balancing with GPUs.
- We have designed over STARPU a lightweight DSM over MPI, which permits to seamlessly execute STARPU applications over an MPI cluster of GPU-enhanced nodes.
- We have been developing a GCC plug-in which extends the C language with pragmas and attributes that make writing STARPU applications a lot easier.
- We have brought to STARPU support for automatically converting data between CPU and GPU formats (typically arrays of structures vs structures of arrays). We are now optimizing it.
- We have added an OpenCL interface to STARPU, SOCL [42], which permits to execute unmodified OpenCL applications over STARPU.
- We have introduced in STARPU theoretical bound support [27]: from a record of the set of tasks submitted by the application, STARPU uses linear programming to give the execution time of an ideal scheduling, which can then be compared with the actual results.
- We have continued collaboration with the University of Tennessee, Knoxville for STARPU support in the state-of-the-art dense linear algebra library, Magma, in particular LU [26] and QR [27] factorizations. We have also collaborated with the University of Mons [41] and Linköping [32].
- Cédric Augonnet defended his PhD on STARPU [17].
6.7. Development of a flexible heterogeneous system-on-chip platform using a mix of programmable processing elements and hardware accelerators

**Participants:** Paul-Antoine Arras, Emmanuel Jeannot, Samuel Thibault.

- Today’s embedded applications are increasingly demanding in terms of computational power, especially in real-time digital signal processing (DSP) where tight timing requirements are to be fulfilled. More specifically, when it comes to video decoding (e.g. H.264/AVC) not only has it been almost impossible for some time to run such codecs on a stand-alone embedded processor, but it now also becomes quite impractical to execute them on homogeneous multicore platforms. In this context, STMicroelectronics is developing a scalable heterogeneous system-on-chip template called P2012 and aimed at meeting the latest codecs’ requirements.

- This year, the privileged axis of research was directed towards dataflow-based models, which benefit from such strong, well-known properties as analyzability, schedulability and expressivity. Furthermore, dataflow programming has already been used extensively in DSP, yielding a number of dedicated software synthesis tools. We have proposed a first version of the programming model that will be evaluated later.

6.8. Sparse GMRES on heterogeneous platforms in oil extraction simulation

**Participants:** Olivier Aumage, Corentin Rossignon, Samuel Thibault.

- We started a study on sparse matrix factorization and system resolution on heterogeneous platforms in collaboration with Pascal Hénon from company Total, in the context of oil extraction simulation. Sparse matrix computations are notoriously difficult to efficiently run on heterogeneous platforms in the general case due to the irregular memory access patterns they generate.

- However, in the specific context of this study, Corentin Rossignon showed as part of his Master Thesis [56] that the sparsity layout of matrices generated by such oil extraction simulation problems can lead to a much higher level of efficiency on heterogeneous platforms thanks when using a suitable sparse internal representation together with carefully written operators such as the sparse matrix-vector product together with the StarPU heterogeneous scheduler.

- Corentin Rossignon is now starting a Phd. Thesis in partnership with Total to build on these promising results.

6.9. Programming models for heterogeneous platforms

**Participants:** Olivier Aumage, Cyril Roelandt, Samuel Thibault, Ludovic Courtès.

- As part of Project FP3C with Japan, we started a study on to explore the use of StarPU as possible target runtime system for the XcalableMP language and compiler developed by Prof. Sato’s team from University of Tsukuba. XcalableMP is a pragma-based language designed for parallelising application on clusters of multicore processors. The compiler is responsible to expand XcalableMP pragma into complex work mapping, communication and data redistribution commands.

- The study of porting XcalableMP on top of StarPU was conducted by Cyril Roelandt during his Master Thesis [55], starting from the idea that computing node with one or more attached accelerating expansion cards can be seen as a distributed platform. The results of the study showed that on the one side, the power of the XcalableMP language itself is very interesting for the goal of simplifying the port of applications on heterogeneous platforms. However, a current assumption of the XcalableMP model is that the compiler does not insert implicit commands and behaviour except at the exact location of pragma annotations, which limit the range of optimizations available to the dynamic scheduler and memory manager of StarPU. We will thus continue to collaborate with Prof. Sato’s team within the FP3C to see how these limitations could be reduced or lifted when using XcalableMP with StarPU.
In an effort to make it easier for C programmers to benefit from StarPU, the team-project has been working on extensions to the C language allowing important StarPU concepts to be expressed concisely. These C extensions are provided as a plug-in for the GNU Compiler Collection (GCC\(^1\)), and is now distributed as part of StarPU.

The GCC plug-in extends the syntax and semantics of C and related languages (C++, Objective-C) using attributes and pragmas. Attributes are used, for instance, to declare StarPU tasks and their implementations for the available targets (CPU, OpenCL, CUDA, etc.) Pragmas are used notably to provide programmers a way to describe data buffers that are passed to tasks, which in turn allows the StarPU run-time support to manage data transfers between main memory and GPUs as it sees fit. Finally, tasks are invoked like regular C functions.

In addition to easing application development, the GCC plug-in, thanks to its higher-level view of the program structure, is able to report certain classes of errors at compile-time, which would otherwise lead to run-time errors.

This project has been led by Ludovic Courtès of Inria’s Development and Experimentation Department (SED) at Bordeaux, as part of a joint development action with the SED.

6.10. Parallel Concha

Participants: Olivier Aumage, Marie-Christine Counilh.

- Within the ADT Ampli project, we contributed to the Concha CFD library developed by R. Becker’s Inria Team Concha in Pau. Together with R. Becker, E. Bergounioux and D. Trujillo from Concha Team, and François Rue from SED Bordeaux we designed and experimented with the MPI parallelization and the hybrid MPI+OpenMP parallelization of the library.
- The MPI parallelization is now finalized. The OpenMP level has been successfully tested on the Vanka smoother and is now being spread in the library. We will thus continue to contribute to this parallelization work, in particular with respect to the support of 3D simulation cases.

6.11. Scientific Application Analysis and Experiments

Participants: Olivier Aumage, Denis Barthou, Andres Charif-Rubial, François Tessier, Ludovic Stordeur.

- Within the context of the ANR ProHMPT project, we contributed a thorough analysis of hot spots, data structure usages and locality issues in memory accesses of an aerodynamics application from partner CEA CESTA.
- In accordance with these results, a new version of this application has been written by the CESTA Team with redesigned, locality-friendly data structures and simplified loop scheme. This new version performs much better than the previous one on both 2D and 3D cases.
- We also conducted tests about the port of selected kernels of this application on accelerated heterogeneous platforms. The results of these tests were disappointing with the first version of the application due to the layout of the main data structures that led to a lot of memory transfers between the central memory and the accelerated memory.
- We are now working on conducting these experiments with the redesigned version of the application whose new data structures should dramatically reduce the amount of data transfers.

\(^1\)See [http://gcc.gnu.org/](http://gcc.gnu.org/), for more information on GCC.
6.12. Virtualization of GPUs for OpenCL
Participants: Sylvain Henry, Alexandre Denis, Denis Barthou.

- We propose a new approach for OpenCL programming, using a unique virtual accelerator instead of using the physical accelerator. Placement on the real hardware is handled by the runtime instead of the user, improving productivity and performance scalability. This proposition relies on OpenCL standard but changes the way its API is used.
- We have shown on some simple examples how this approach, using StarPU as a runtime, enables executions with a better load balance and performance. We are working on how to generalize this to more complex benchmarks. This work has been presented in Renpar[42] workshop.

6.13. Automatically Adapting Task Grain for Hybrid Architectures
Participants: Sylvain Henry, Alexandre Denis, Denis Barthou.

- Given a parallel task graph, a runtime such as StarPU can place each task on different hardware. However, there is still the need to adapt the number of tasks, the granularity of these tasks, according to the target hardware. Due to architectures with CPUs and GPUs, it is potentially interesting to have tasks of different granularities. We explore transformations that enable to either automatically split tasks into small ones, or given some user knowledge on the tasks, decide how and when to split a large task into small ones.
- This work starts from a high-level representation of the code, using an explicit data-flow graph.

6.14. Performance modeling for power consumption reduction on the SCC
Participants: Bertrand Putigny, Brice Goglin, Denis Barthou.

- We build a model to predict performance of HPC code on the SCC ship. This model can predict runtime of regular code as well as power consumption for different frequency.
- This allows users to choose either to optimize power consumption, power efficiency or raw performance.
- This work has been published in an Intel Symposium [38].

6.15. Modeling cache coherence protocol overhead
Participants: Bertrand Putigny, Denis Barthou, Brice Goglin.

- We are building a fine grained cache model to understand common cache coherence issue.
- This model is built on a set of micro-benchmarks and can also be used to improve find some bottlenecks in memory bound code. Our set of micro-benchmarks can also be used as a test bed for new architectures [54].

6.16. Memory Performance Analysis and Tool for OpenMP codes
Participants: Andres Charif-Rubial, Denis Barthou.

- We propose a performance analysis of OpenMP codes, based on memory accesses and cache hierarchies.
- This analysis relies on memory traces for multi-threaded codes and on static analysis of binary code. Memory traces are obtained through MAQAO by static binary rewriting and are compressed online, building polyhedral iteration domains. The static analysis, mostly induction variable detection on binary code, provides the same analysis whenever possible, removing the need in some cases for dynamic instrumentation.
- The analysis focuses on a number of issues in multi-threaded executions: thread affinity issues, false sharing, cache pollution.
- This work is in collaboration with Exascale Computing Lab.
6.17. Data-layout Optimization for Stencil codes on multi-cores and GPUs

Participants: Julien Jaeger, Denis Barthou.

- We develop a new approach for stencil code generation, optimizing data-layout for multi-threaded, SIMD code on multicore and CUDA code on GPU. The transformation handles different stencil parameters, and memory constraints.
- The code generated reaches high levels of performance, outperforming related works for multicore and with similar performance on GPUs. This work is submitted to publication and was first presented in a workshop [52].
6. New Results

6.1. Network Economics

**Participants:** Pierre Coucheney, Hai Tran Hoang, Bruno Tuffin, Jean-Marc Vigne.

While pricing telecommunication networks was one of our main activities for the past few years, we are now dealing with the more general topic of *network economics*. We have tackled it from different sides: i) investigating how QoS or QoE can be related to users’ willingness to pay, ii) investigating the consequences and equilibrium due to competition among providers in different contexts, iii) studying the economic aspect of interdomain relationships, iv) looking at the economics of applications, for example adword auctions for search engines, v) investigating the economics of security in telecommunications, vi) studying the network neutrality issue.

On the first item, in [70], [29], we have studied how utility functions can be related to QoE recent research. Indeed, a logarithmic version of utility usually serves as the standard example due to its simplicity and mathematical tractability. We argue that there are much more (and better) reasons to consider logarithmic utilities as really paradigmatic, at least when it comes to characterizing user experience with specific telecommunication services. We justify this claim and demonstrate that, especially for Voice-over-IP and mobile broadband scenarios, there is increasing evidence that user experience and satisfaction follow logarithmic laws. Finally, we go even one step further and put these results into the broader context of the Weber-Fechner Law, a key principle in psychophysics describing the general relationship between the magnitude of a physical stimulus and its perceived intensity within the human sensory system.

A notable part of our activity has been related to competition among telecommunication providers, mainly within the framework of the ANR CAPTURES project. The goal is to improve most of the pricing models analysis which only deal with a single provider while competition (that is observed in the telecommunication industry) can drive to totally different outcomes. A general view of some of our results is summarized in [67]. A general model of competition in loss networks is described and analyzed in [22] as a two-levels game: at the smallest time scale, users’ demand is split among providers according to the Wardrop principle, depending on the access price and available QoS (depending itself on the level of demand at the provider), and at the largest time scale, providers play a pricing game, trying non-cooperatively to maximize their revenue. A striking result is that this game leads to the same outcome than if providers were cooperatively trying to maximize social welfare: the so-called *price of anarchy* is equal to one. An additional (higher) level of game is analyzed in [23] (but using another type of negative externality for users, based here on delay), at which providers play on the technologies to implement, based on the infrastructure and licence (if any) costs, anticipating what would be the resulting price war outcome and revenue for given profiles of sets of technologies. This type of study may help a regulator to decide a licence cost, in order to drive the resulting Nash equilibrium to a better point in terms of social or user welfare. A specific situation we have analyzed is the case for a competitive market operated by a Mobile Network Operator (MNO) and a Mobile Virtual Network Operator (MVNO) [46]. The resource that is leased by the MNO to the MVNO is spectrum. MNO and MVNO compete posting subscription prices and the mobile users may choose to subscribe to one operator, or not to subscribe. The scenario is modeled by a three-level game comprising: a bargaining game, which models the spectrum leasing by the MNO; a competition game, which models the price competition between the MNO and the MVNO, and a subscription game, which models the subscription choice by the mobile users, and the outcome of which may be either not to subscribe, to subscribe to the MNO or to subscribe to the MVNO. We assess which conditions lead to an equilibrium where the competition does take place and the amount of the spectrum that should be leased to maximize user or social welfare.
Another important activity is around interdomain issues, with a network like the Internet being made of thousands of autonomous systems. Intermediate domains need some (economic in our case) incentives for forwarding the traffic of other domains. In [33], we have described the problem, provided a state of the art and highlighted the difficulties that must be solved. In [32], we have designed a decentralized algorithm based on double-sided auctions to allocate (and charge) the resource usage.

But network economics is not only about ISPs, it also deals with the application side. In order to make money many service providers base their revenue on advertisement. Search engines for example get revenue thanks to adword auctions, where commercial links are proposed and charged to advertisers as soon as the link is clicked through. Most search engines have chosen (or switched to) a revenue-based ranking and charging scheme instead of a bid-based one. In [53] we investigate the relevance of that scheme when advertisers’ valuation comes from a random distribution, showing that depending on the search engine’s click-through-rate, revenue-based does not always outperform bid-based in terms of revenue to the search engine. But in this adword auction context too, there exist very few works dealing with serach engines in competition for advertisers. We have developed a two-level game where at the largest time scale search engines decide which allocation rule to implement, between revenue-based and bid-based; and at the lowest time-scale advertisers decide how to split their advertising budget between the two search engines, depending on the benefits this will bring to them. The game at the largest time scale is solved using backward induction, the search engines anticipating the reactions of advertisers [54], [52]. We describe the advertisers best strategies and show how to determine, depending on parameters, an equilibrium on the ranking rule strategy for search engines; this may explain Yahoo!’s move to switch from bid-based to revenue-based ranking to follow Google’s strategy.

We similarly have looked at the competition aspects linked to security. We have reviewed the interactions and strategies of attackers and defenders [68]. But we have also looked at the economics of network security, when network users can choose among different security solutions to protect their data, offered by competitive security providers [51]. The interactions among users are modeled as a noncooperative game, with a negative externality coming from the fact that attackers target popular systems to maximize their expected gain.

A new issue we are investigating is the network neutrality debate coming from the increasing asymmetry between Internet Service Providers (ISPs), mainly due to some prominent and resource consuming content providers which are usually connected to a single ISP. We have described and analyzed in [69] the respective arguments of neutrality proponents and opponents, and are currently completing the analysis of several promising game-theoretic models on this issue.

### 6.2. Dependability and extensions

**Participants:** Raymond Marie, Gerardo Rubino, Samira Saggadi, Bruno Tuffin.

We maintain a permanent research activity in different domains related to dependability, performability and vulnerability analysis of communication systems. Our focus is on evaluation techniques using both the Monte Carlo and the Quasi-Monte Carlo approaches. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only available tool to solve complex problems in the area, and rare event simulation requires a special attention, in order to be able to efficiently analyze the model, that is, to be able to use good estimators having, in particular, a sufficiently small relative variance. Novel results in simulation can be decomposed into two subsets: results on rare event simulation, and those on Randomized Quasi-Monte Carlo methods.

The effectiveness of randomized quasi-Monte Carlo (RQMC) techniques is examined in [26] to estimate the integrals that express the discrete choice probabilities in a mixed logit model, for which no closed form formula is available. These models are used extensively in travel behavior research. We consider popular RQMC constructions, but our main emphasis is on randomly-shifted lattice rules, for which we study how to select the parameters as a function of the considered class of integrands. We compare the effectiveness of all these methods and of standard Monte Carlo (MC) to reduce both the variance and the bias when estimating the log-likelihood function at a given parameter value.
The main part of our activity in this simulation area in 2011 has been on rare event simulation though. The two major simulation families or rare event estimations are importance sampling and splitting. In [63], we have provided a recent view of those methods, while in [64] we have overviewed how the zero-variance importance sampling can be approximated in classical reliability problems.

The problem of estimating the probability that a given set of nodes is connected in a graph (or network) where each link is failed with a given probability has received a lot of attention from us in 2011. We have proposed in [21] a new Monte Carlo method, based on dynamic Importance Sampling. The method generates the link states one by one, using a sampling strategy that approximates an ideal zero-variance importance sampling scheme. The approximation is based on minimal cuts in subgraphs. In an asymptotic rare-event regime where failure probability becomes very small, we prove that the relative error of our estimator remains bounded, and even converges to 0 under additional conditions, when the unreliability of individual links converges to 0. The empirical performance of the new sampling scheme is illustrated by examples. The method is even sped up in [50] by applying series-parallel reductions at each step of the algorithm.

The same problem is also analyzed in [15] by a novel method that exploits a generalized splitting (GS) algorithm. We show that the proposed GS algorithm can accurately estimate extremely small unreliabilities and we exhibit large examples where it performs much better than existing approaches. Remarkably, it is also flexible enough to dispense with the frequently made assumption of independent edge failures. In [17], another splitting approach is explored for the same problem, with very good results. It consists of a standard splitting procedure applied to the so-called Creation Process that can be associated with the initial static model. The paper discusses both a method for splitting this process, and an experimental analysis of the covering of the resulting estimator, showing its good behavior on different classes of test problems. Last, in [16], always for the same static reliability problem, we proposed a new procedure belonging to the RVR family (Recursive Variance Reduction) where a new estimator based both in computed minpaths and mincuts of the graph, together with series-parallel reductions, allows to obtain very good accuracy in many rare events situations.

Concerning the risk on spares for life-time maintenance purposes which is due to uncertainties on the mean up time, an extended version of a presentation made in 2010 has been published in [24].

6.3. Performance evaluation

Participants: Laura Aspirot, Raymond Marie, Gerardo Rubino, Bruno Sericola.

An important problem arising when dimensioning a P2P system is to understand the evolution of the peers’ population with time. The number of units being usually large, the standard stochastic models used to represent this kind of system (e.g. a Markovian stochastic process) are difficult to use in practice. Instead, it is popular today to move to deterministic continuous-state (fluid) models whose dynamics is governed by differential equations. It is then of interest to analyze the conditions under which the latter are the limit, in some sense, of the former. We started to develop this program in [36] by focusing on some popular models of P2P systems, and analyzed when and how the deterministic model is the limit of the stochastic one when the number of peers goes to infinity.

In [60], we continued to explore the concept of power of a queueing model proposed by Kleinrock in the 80s. Kleinrock’s idea was to build a metric combining two “competing” ones, the mean throughput and the mean response time, for the system in equilibrium. The power is defined as the ratio of normalized versions of those metrics. We discuss different ways of adapting this concept to more general queueing systems such as queueing networks. In this research line, [60] opens the way for a definition of efficiency, which is currently analyzed in the team.
In [30], we expose a clear methodology to analyze maximum level and hitting probabilities in a Markov
driven fluid queue for various initial condition scenarios and in both cases of infinite and finite buffers. Step by
step we build up our argument that finally leads to matrix differential Riccati equations for which there exists a
unique solution. The power of the methodology resides in the simple probabilistic argument used that permits
to obtain analytic solutions. We illustrate our results by a comprehensive fluid model that we solve exactly.

In [65], we analyze the transient behavior of a fluid queue driven by a general ergodic birth and death process
using spectral theory in the Laplace transform domain. These results are applied to the stationary regime and
to the busy period analysis of that fluid queue.

Finally, in [71] we present a global view of the performance evaluation area in computer and communication
systems, an extended and reviewed version of a talk given in 2010.

6.4. Quantitative aspects of distributed systems

Participants: Bruno Sericola, Romaric Ludinard.

This work is a collaboration with the Inria team-project Asap. We proposed in [20] a fully decentralized
algorithm to provide each of the nodes of a distributed system with a value reflecting its connectivity quality.
Comparing these values between nodes, enables to have a local approximation of a global characteristic of the
graph. Our algorithm relies on an anonymous probe visiting the network in a unbiased random fashion. Each
node records the time elapsed between visits of the probe which is called the return time of the random walk.
Computing the standard deviation of such return times enables to approximate the conductance of the graph.
Typically, this information may be used by nodes to assess their position, and therefore the fact that they are
critical, in a graph exhibiting low conductance.

We continue our collaboration with the Inria team-projects Adept and Ipso. It is well-known that peer-to-peer
overlays networks can only survive Byzantine attacks if malicious nodes are not able to predict what will
be the topology of the network for a given sequence of join and leave operations. In [13] and in [35], we
investigate adversarial strategies by following specific games. Our analysis demonstrates first that an adversary
can very quickly subvert DHT-based overlays by simply never triggering leave operations. We then show that
when all nodes (honest and malicious ones) are imposed on a limited lifetime, the system eventually reaches
a stationary regime where the ratio of polluted clusters is bounded, independently from the initial amount of
corruption in the system. These results have been obtained using Markov models. In [14] and [34], we
consider 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. We
obtain its distribution and its expectation and we propose an algorithm to compute these quantities. We also
exhibit the asymptotic behavior of the system when the number of Markov chains goes to infinity. Actually,
this problem comes from the analysis of large-scale distributed systems and we show how our results apply to
this domain.

6.5. QoE (Quality of Experience)

Participants: Sebastián Basterrech, Yassine Hadjadj-Aoul, Sofiene Jelassi, Adlen Ksentini, Gerardo Rubino,
Kamal Singh, César Viho.

We continue the development of the PSQA technology (Pseudo-Subjective Quality Assessment) in the area of
Quality of Experience (QoE). PSQA is today a stable 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 than 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, if necessary. Observe that perceived quality
is the main component of QoE. 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, application, service, the module must be built from scratch. But once built, it works automatically and very efficiently, allowing if necessary to use it in real time.

At the heart of the PSQA approach there is the statistical learning process necessary to develop measuring modules. So far we have been using Random Neural Networks (RNNs) as our learning tool (see [82] for a general description), but recently, we have started to explore other approaches. For instance, in the last ten years a new computational paradigm was presented under the name of Reservoir Computing (RC) [78] covering the main limitations in training time for recurrent neural networks while introducing no significant disadvantages. Two RC models have been developed independently and simultaneously under the name of Liquid State Machine (LSM) [81] and Echo State Networks (ESN) [78] and constitute today one of the basic paradigms for Recurrent Neural Networks modeling [79]. The main characteristic of the RC model is that it separates two parts: a static sub-structure called reservoir which involves the use of cycles in order to provide dynamic memory in the network, and a parametric part composed of a function such as a multiple linear regression or a classical single layer network. The reservoir can be seen as a dynamical system that expand the input stream in a space of states. The learning part of the model is the parametric one. In a recent collaboration with the Applied Computational Intelligence Research Unit, Artificial Neural Networks Group of the University of the West of Scotland during the first half of the year, we developed an algorithm based on a combination of topology preserving maps such as the Self-Organising Map [80] and the Scale Invariant Map [77] to improve the performance of RC models. The obtained results are presented in two papers: [37] and [38].

In [42] we developed a PSQA version for evaluating the perceived quality in the context of SVC video coding. The tool is based on the use of the RNN model. The main difficulties in defining this tool is regarding the relation between the SVC layers, since the enhanced layers require the information of the base layer in order to be decoded.

In [61], we developed a tool for evaluating the perceived quality of an application distributing streamed video using HTTP (and thus, TCP). The difficulties here are focused around the possible playout interruptions and the quality variations due to the use of adaptive bitrate techniques. Our procedure belongs to the no-reference family of learning ones, and it is also based on the use of the RNN tool.

In [41] we compared PSQA used for the video evaluation to other no-reference tools as well as two objective evaluation tools. We showed that PSQA outperforms the majority of the other tools, in terms of high correlation with human evaluation. This version will be used as the main metric for evaluating the QoE in the future internet architecture proposed by the FP7 Alicante project.

We have also being developing single-ended parametric-model speech-quality assessors of VoIP conversations over future networks. To do that, a careful identification and accurate characterization of quality-degrading factors over next-generation networks has been done. The recent progress and challenges for accurate assessment of voice quality over evolving VoIP systems has been detailed in the survey paper [19]. In [18], we study the perceived effects of packet loss processes, which are the principal source of quality degradation over IP networks. In reality, the perceived effect of a given packet loss process is highly related to the distribution of missing packets. Basically, the higher the burstiness of packet loss processes, the greater the perceived quality degradation. Recently, several assessors of speech quality sensitive to packet loss burstiness have been proposed in the literature. A comprehensive comparison study of bursty-packet-loss-aware artificial assessors has been conducted in [18]. An extended and more elaborated version has been published in [47]. Moreover, novel artificial quality assessors that consider transient loss of connectivity incurred by mobile users over mobile transport system have been developed. A paper describing our developed tools and performance results is under preparation. Recently, we started to work on new analytical models of packet losses and delays of packet-based voice conversations over wireless ad-hoc networks. The developed models will be used to design specialized artificial quality assessors of multimedia services over wireless ad-hoc networks. Moreover,
we are working on the enhancement of a voice quality assessor version of PSQA, by considering the features of removed speech signals.

### 6.6. Wireless networks

**Participants:** Nizar Bouabdallah, Yassine Hadjadj-Aoul, Adlen Ksentini, Raymond Marie, Bruno Sericola, César Viho.

We continue working on wireless networking. The focus mainly concerns wireless distribution of audio and video, which require strict Quality of Service (QoS) support.

In [27], we investigated the main challenges when the goal is to constitute an efficient Radio Resource Management (RRM) framework. The existing solutions of RRM were classified based on the considered decision-making technique. Moreover, we investigated in [28] how QoE can help for designing efficient RRM for wireless networks. A resource allocation mechanism is proposed in [62]. In [59] we proposed a novel network selection mechanism for heterogeneous wireless networks that take QoE into consideration for decision-making. The main idea is to use QoE of ongoing users in candidate networks as an indicator to select the best network for connection. Besides, in order to provide efficient interworking between the different access players, we first defined some issues related to the interworking operation between the satellite and terrestrial domains. We suggested some solutions and discussed their potential in [31].

We also investigated in [74] solutions that ensure the scalability of mobile networks, which are facing a rapid increase of data traffic. We devise methods that enable User Equipments, both in idle and active mode and while being on the move, to always have optimal Packet Data Network (PDN) connections (i.e., IP addresses) in such decentralized networks. We demonstrated the effectiveness of such approach in current mobile and wireless networks. In these systems, minimizing energy consumption is becoming more and more crucial. In [66], we devised a PID (Proportional Integral Derivative)-based controller permitting to reduce the amount of wasted energy by determining an optimal schedule between the sleep and wakeup periods of the wireless interface during the VoIP communication while keeping the perceived quality at the desired level.

Based on our previous research on proactive routing for wireless ad-hoc networks, we have published a book chapter in [73], focusing on modeling the resilience of routing information for ad hoc networks where topology information is uncertain.

We continue our collaboration with the Inria team-projects Pops (Lille), D-Net (Lyon), Reso (Lyon) and the NPA (Networks and Performance Analysis) research group of LIP6 (Paris) on fast self-stabilization in large scale wireless networks. In these systems, distributed self-organization is more convenient than centralized planification. Self-stabilization protocols are a useful technique to provide self-organization but their stabilizing time is related to the size of the network. In [25], we show that a clustering algorithm, known for its good robustness properties, is actually self-stabilizing. We propose several enhancements to the scheme in order to reduce the stabilization time and thus improved the stability in a dynamic environment. The key technique to these enhancements is a localized self-stabilizing algorithm for directed acyclic graph construction. We provide extensive studies (both theoretical and experimental) that show that our approach enables efficient yet adaptive clustering in wireless multihop networks.

### 6.7. Sensor networks

**Participants:** Nizar Bouabdallah, Sofiane Moad.

Wireless Sensor Networks (WSNs) are composed of tiny sensor nodes, which are capable of sensing and processing data from inaccessible environments and communicating them to the end-user for further analysis. WSNs are characterized by the limited capacity of their sensor node batteries, making energy efficiency a critical issue. Once a WSN is deployed, sensor nodes must self-organize and live as long as possible, based only on their initial energy stores. Consequently, techniques minimizing energy consumption are required to improve network lifetime. Our research on WSNs [72] revolves around two main directions: 1) clustering, and 2) radio diversity. Regarding clustering, we first developed a Connectivity Degree-Based Energy Efficient
Clustering Protocol for WSNs (CDEEC) [55], resulting in better topology management and decreased energy consumption compared to the well-known clustering protocol HEED. Then, we integrated a compression mechanism within a cluster-based architecture to develop a Compression Cluster-based scheme in a Spatial Correlated Region protocol (CC_SCR) [56], with the goal of further decreasing the energy consumption. In the direction of radio diversity, we first proposed the WETX metric [58] which uses a minimum-energy radio while routing, then we proposed the BL metric [57], on top of WETX, which allows energy-balancing inside a network in order to further extend its lifetime. The validation of our contributions was carried out with analytical analysis, and simulation using TOSSIM.

6.8. Scalable Video Coding (SVC) transmission over IP and Broadcast networks

Participants: Majd Ghareeb, Adlen Ksentini, César Viho, Yassine Hadjadj-Aoul.

One of the multimedia market trends is audiovisual service (TV or VoD) anywhere, at any time. To support such service, a Video Service Provider has to manage, store, and distribute content towards multiple kinds and scales of terminals, and over different and transient access technologies to reach the end user. To solve such issues, video scalability seems to be the most relevant solution. It encodes the video in multiple separated layers, which enable a large number of users with heterogeneous capability to view any desired video stream, at anytime, and from anywhere. One of the most well known scalable standards is the Scalable Video Coding (SVC) extension of H.264/MPEG-4 AVC video compression. Our researches in this topic are related to how to optimize and enhance SVC transmission over IP and broadcast networks.

With the aim at keeping a high perceived video quality using SVC, MultiPath Video Streaming (MPVS) over Video Distribution Network (VDN) comes as a promising solution to overcome the limitations of the classical single path and IP-level video streaming approaches. In [45] and [43] we proposed different approaches that couple the three SVC scalability modes (Spatial, Temporal, SNR), with the path diversity provided by VDN. Our method adapts to both the heterogeneity of end-users using the scalable video coding as well as to network bandwidth fluctuations by observing the changes of the available bandwidth over the multiple overlay paths, and updating the streaming strategy accordingly. In [44] we enhanced the precedent solutions by using the PSQA-SVC version [42] in order to get the end-user feedback in terms of QoE, which helps adapting the streaming strategy. In [48] we designed a new protocol optimizing the energy consumption when transmitting video streams. We propose to exploit the SVC coding to adapt dynamically the received video quality to the instantaneous wireless nodes’ characteristics. This is achieved through determining the number of the transmitted/received enhancements layers of an SVC video based on the wireless node context.

In [49] we proposed to support SVC over DVBT2 networks, by associating the layering architecture of both technologies in order to tackle users mobility. This association allows mobile receivers with good physical channels to decode all the SVC layers and benefit from high video quality. Meanwhile, users with poor channel conditions can at least decode the base layer and benefit from acceptable video quality. Further, we introduced a novel QoS-based adaptive mechanism for SVC layers decoding. The proposed approach selects dynamically the number of layers to decode, at the receiver side, so as to maximize the users’ perceived quality. Thus, no feedbacks or signaling messages are needed to implement the proposed algorithm. This makes it compliant with unidirectional technologies such as DVB-T2.
6. New Results

6.1. Fundamental results and algorithms: distributed planning

Participants: Eric Fabre, Loïg Jézéquel.

A planning problem consists in organizing some actions in order to reach an objective. Formally, this is equivalent to finding a path from an initial state to a goal/marketed state in a huge automaton. The latter is specified by a collection of resources, that may be available or not (which defines a state), and actions that consume and produce resources (which defines a transition). In the case of optimal planning, actions have a cost, and the objective is to find a path of minimal cost to the goal.

Our interest in this problem is threefold. First, it is naturally an instance of a concurrent system, given that actions have local effects on resources. Secondly, it is a weak form of an optimal control problem for a concurrent/distributed system. Finally, we are interested in distributed solutions to such problems, which is an active topic in the planning community under the name of “factored planning.”

Our previous contributions to the domain was the first optimal factored planning algorithm [61]. The main idea is to represent a planning problem as a network of interacting weighted automata, the objective being to jointly drive all of them to a target state, while minimizing the cost of their joint trajectory. We have developed and tested [68] a distributed algorithm to solve this problem, based on a weighted automata calculus, and that takes the shape of a message passing procedure. Components perform local computations, exchange messages with their neighbors, in an asynchronous manner, and the procedure converges to the path that each component should follow. The optimal global plan is thus given as a tuple of (compatible) local plans, i.e. a partial order of actions.

In 2011, we have extended this framework in two directions. In terms of modelling, first. In most planning problems, some actions consume/produce resources, but also are enabled by the presence of other resources, that they only read but not consume. We have proposed to model this feature under the form of networks of automata with read arcs. Interactions then take the form of synchronous actions, as previously, but also the form of readings: a component may only be allowed to fire some local transition if another component is in a specific state. Our distributed planning approach has then been extended to this new model of distributed systems [38].

The second improvement is algorithmic. So far, our distributed optimal planning algorithm computes all possible distributed plans, in a factored form. This contrasts with the philosophy of planning algorithms, that look for one plan only, and organize the computations to quickly reach the best plan. In other words, most planning algorithms are based on a common ground known as the A-star algorithm, a depth-first search procedure in a graph, guided by a heuristic function that estimates the remaining cost to reach the goal. We have developed a truly distributed version of this algorithm, to perform a search on a product graph. Each component runs an A-star procedure to find a path to its goal, taking into account the costs of its neighbors in order to guarantee that all components converge to local plans that are compatible and jointly optimal.

6.2. Fundamental results and algorithms: communication with messages and scenarios

Participants: Loïc Hélouët, Rouwaida Abdallah, Claude Jard, Blaise Genest.

In this paragraph, we collect our fundamental results regarding the models and algorithms we use for communicating systems, and in particular, scenarios.
A major challenge with models communicating with messages (e.g.: scenarios) is to exhibit good classes of models allowing users to specify easily complex distributed systems while preserving the decidability of some key problems, such as diagnosis, equality and intersection. Furthermore, when these problems are decidable for the designed models, the second challenge is to design algorithms to keep the complexity low enough to allow implementation in real cases.

This year, we have considered analysis for a timed extension of scenarios called Time-constrained MSCs and implementation techniques that take scenarios as an input model and output an equivalent distributed implementation.

The first part of our work is the study of Time-Constrained MSC graphs (TC-MSGS for short). Time-constrained MSCs (TC-MSCs) are simply MSCs decorated with constraints on the respective occurrence dates of events. The semantics of a TC-MSC $T$ is a dated MSC, that is a MSC where events are associated with an occurrence date. For a given TC-MSC, there can be an infinite set $L(T)$ of dated MSCs satisfying its constraints. Note however that some time-constraints in a TC-MSC may not be satisifiable, and hence $L(T)$ can simply be empty. TC-MSCs can be extended by composition mechanisms such as TC-MSC graphs. TC-MSC graphs are simply automata labeled by TC-MSC. Each path $\rho$ of a TC-MSC $G$ is associated with a TC-MSC $T_\rho$ obtained by concatenation of TC-MSC along $\rho$. The language $L(G) = \bigcup_{\rho \text{ path of } G} L(T_\rho)$ of a TC-MSC Graph is then the union of all dated MSCs associated to paths of $G$. Because of inconsistent timing constraints, some path may have no possible realization (i.e $L(T_\rho = \emptyset$). One can even design a MSC Graph $G$ such that $L(G) = \emptyset$ - such TC-MSC graph is clearly inconsistent-. It has been shown [64] that checking whether $L(G) = \emptyset$ is an undecidable problem in general, but can be decided for the restricted subclass of regular TC-MSC graphs (that have the expressive power of event-count timed automata). We have proposed two restrictions allowing for the decision of emptiness. The first one is $K$-drift boundedness, which imposes for a fixed integer $K$ that for every $T_\rho$ there exists one dated realization such that for every pair of events $e,f$ appearing in the same transition of $G$, the dates of $e$ and $f$ differ by at most $K$. We have shown that $K$-drift boundedness is decidable in a symbolic and efficient way, and that for $K$-drift bounded TC-MSC graphs, emptiness is decidable [52]. This extends decidability results beyond regular specifications. The second restriction is $K$-non-zenoness, which imposes that for a fixed $K$, for every path $\rho$ of $G$, there exists one realization such that at every date $d$, at most $K$ events occur between date $d$ and $d+1$. When a TC-MSC graph is $A$-drift-bounded and $B$-non-zeno, then $L(G)$ has a regular set of representants, which opens the way for more involved model-checking applications [51].

The second part of our work is the study of realistic implementation of scenarios. The main idea is to propose distributed implementation (communicating state machines) of High-level MSCs that do not contain deadlocks, and behave exactly as the original specification. It is well known that a simple projection of a HMSC on each of its process to obtain communicating finite state machines results in an implementation with more behaviors than the original specification. An implementation of a HMSC $H$ is considered as consistent if and only if it exhibits the same prefix closed set of behaviors as $H$. We have studied how such projection with additional local controllers allows the distributed synthesized behavior to remain consistent with the original specification. This work has been implemented in our scenario prototype (see the Software section). As usually for scenarios, the synthesis algorithm works for a particular syntactic class of scenarios, namely the class of local HMSCs. Roughly speaking, in local HMSC, a decision to behave according to a scenario or another is always taken by a single participant. The deciding process need not be the same at each choice. This class is a sensible restriction of HMSCs, as distributed choices can not be implemented without additional synchronization among processes [53].

Last, we have extended existing results on diagnosis from scenarios [15]. We have shown that when a distributed implementation is instrumented with software probes that publish their observations while the system is running, and when the system is modeled as a High-level MSC, then diagnosis can be expressed as a new HMSC the executions of which are all explanations of the observation. The construction of diagnosis can be performed offline or online, and we have considered the conditions under which online diagnosis can run with finite memory.
6.3. Fundamental results and algorithms: timed models

Participants: Claude Jard, Aurore Junier, Akshay Sundararaman.

Our works in that subject concern Time Petri Nets (TPNs) and Network Calculus. With TPNs, we are particularly interested in symbolic unfoldings (extended with parameters). Possible applications are supervision of distributed timed systems [8] and testing of concurrent systems (work done in collaboration with Stefan Haar, INRIA-LSV in Cachan).

The article [26] was made during the internship of the master degree of Aurore Junier under the supervision of Anne Bouillard (ENS Paris). It uses a (min, plus)-algebra to define a worst-case delay bound for networks where flows have fixed priorities.

After that, we studied a well-known problem: detection of congestion and failure in networks. The idea was to find an efficient and deterministic method that is very reactive and takes little memory space. Such a method does not exists for now and is an important issue for Alcatel-Lucent. We achieved a solution that solves this problem based on the analysis of flows behaviour. This work is part of the work done within the Alcatel-Lucent-Inria joint lab and a patent is being established.

We are also studying the way buffers of routers can increase. The objective is to find a method that can detect if sizes of buffers can dangerously increase on a defined topology. We start by looking at Link State Advertisements (LSA) in the OSPF protocol. We represent the topology and a part of the protocol by a Time Petri Net and try to infer parameters ensuring stability.

6.4. Fundamental results and algorithms: dynamic epistemic logic

Participants: Guillaume Aucher, François Schwarzentruber.

Dynamic Epistemic Logic (DEL) deals with the representation and the study of knowledge and belief change in a multi-agent setting. The core representative task of this logical framework can be split up in three parts: 1/ the initial global state of the distributed system, 2/ an event occurring in this system, 3/ a product update taking as argument these two representations and yielding a new representation of the new global state of the distributed system. Therefore, we can express uniformly within the DEL framework epistemic statements about:

(i) what is true about an initial state
(ii) what is true about an event occurring in this initial state
(iii) what is true about the resulting state after the event has occurred.

We axiomatized within the DEL framework what we can infer about (iii) given (i) and (ii), what we can infer about (ii) given (i) and (iii), and what we can infer about (i) given (ii) and (iii). Given three logical formulas \( \phi, \phi', \phi'' \) describing respectively (i), (ii) and (iii), we also showed how to build three formulas that capture respectively all the information which can be inferred about (iii) from \( \phi \) and \( \phi' \), all the information which can be inferred about (ii) from \( \phi \) and \( \phi'' \), and all the information which can be inferred about (i) from \( \phi' \) and \( \phi'' \). We showed how our results extend to other modal logics than the minimal modal logic \( K \). These results are to appear in [9] and [10]. In [19], we also provided a tableau method deciding whether such inferences are valid. We implemented it in LOTRECscheme and showed that this decision problem is \text{NEXPTIME}-complete. This work contributes to the proof theory and the study of the computational complexity of DEL which have rather been neglected so far.

Application to fault localization in IMS network (see the UNIVERSELF project) has started. The various agents involved in an IMS network (clients, assistance, administrators...) have a partial view of the network and so need to communicate their partial knowledge of the network to each other in order to localize the fault in the network (each communication having possibly a different cost). One of the main problems is to determine which communication should occur and which agent should be queried so that the fault is eventually localized. This problem can naturally be expressed in the DEL framework. We have shown how the initial state of an IMS network representing the knowledge of each agent can be represented by a particular kind of epistemic
model (i) and how the desired state where the fault is localized can be expressed by a logical formula (iii). The problem amounts to determining which communication or sequence of communications should occur (ii) so that one passes from the initial epistemic model (i) to another epistemic model where the fault is localized (iii), and also to determine if such a communication or sequence of communications is possible. We have focused so far on the case of a single communication, but we plan to extend it to a sequence of communications. Further theoretical work still needs to be done to address the issue of sequential communication.

In parallel to this work, we also axiomatized different notions of knowledge and belief which are defined by means of a ‘sphere’ semantics. This work is the result of an invited contribution and is to appear in [48].

6.5. Fundamental results and algorithms: statistical model checking

Participants: Sean Sedwards, Cyrille Jégourel, Axel Legay.

Our work on statistical model checking (SMC) avoids an explicit representation of the state space by building a statistical model of the executions of a system and giving results within confidence bounds. The key challenges of this approach are to reduce the length (simulation steps and cpu time) and number of simulation traces necessary to achieve a result with given confidence. Rare properties pose a particular problem in this respect, since they are not only difficult to observe but their probability is difficult to bound. A further goal is to make a tool where the choice of modelling language and logic are flexible.

We have developed the prototype of a compact, modular and efficient SMC platform which we have named PLASMA (PLatform for Statistical Model checking Algorithms). PLASMA incorporates an efficient discrete event simulation algorithm and features an importance sampling engine that can reduce the necessary number of simulation runs when properties are rare. We have found that PLASMA performs significantly better than PRISM (the de facto reference probabilistic model checker) when used in a similar mode: PLASMA's simulation algorithm scales with a lower order and can handle much larger models. When using importance sampling, PLASMA's performance with rare properties is even better.

6.6. Fundamental results and algorithms: quantitative model checking and quantitative specification Theories

Participants: Uli Fahrenberg, Axel Legay.

Model checking of systems deals with the question whether a given model of a computer system satisfies the properties one might want to require of it. This is a well-established and successful approach to formal verification of safety-critical computer systems.

When the models of the systems contain quantitative information, the model checking problem becomes complicated by the fact that in most cases, quantitative properties of the systems do not need to be satisfied exactly. Indeed, the model or the properties might be subject to measurement error, or probabilistic information might only be an approximation. In this case, it is of little use to know whether or not a model satisfies a specification precisely; what is needed instead is a notion of satisfaction distance: a measure which can assess to which extent a quantitative model satisfies a quantitative specification.

In other words, what is needed is a notion of satisfaction which is robust in the sense that small deviations in the model or the specification only lead to small changes in the outcome of the model checking question. We have published work on such distances in the papers [37], [34].

For more elaborate reasoning about distributed systems or systems-of-systems, an important role is played by specification theories. Such systems are often far too complex to reason about, or model-check, as a whole, and additionally they might be composed of a large number of components which are implemented by different vendors. Hence one needs methods for compositional reasoning, which allow to infer properties of a system from properties of its components, and for incremental design, which allow to synthesize and refine specifications in a step-wise manner.
Such specification theories are by now well-established e.g. in the incarnations of interface theories and modal transition systems. Additionally to defining a formalism for describing and model-checking specifications, they provide notions of refinement of specifications, logical conjunction of specifications, and structural composition and quotient.

When the models and specifications contain quantitative information, all the above notions need to be made robust. One needs to introduce a quantitative version of refinement, and the operations on specifications need to be continuous with respect to refinement distance: compositions of specifications with small refinement distance need themselves to have small refinement distance. We have published work on these issues in the papers [21], [35]; additionally, two other papers within this research area are currently under submission.

6.7. Specific studies: Web services orchestrations

Participants: Ajay Kattepur, Albert Benveniste, Claude Jard.

Web services orchestrations and choreographies refer to the composition of several Web services to perform a co-ordinated, typically more complex task. We decided to base our study on a simple and clean formalism for WS orchestrations, namely the Orc formalism proposed by Jayadev Misra and William Cook [71].

Main challenges related to Web services QoS (Quality of Service) include: 1/ To model and quantify the QoS of a service. 2/ To establish a relation between the QoS of queried Web services and that of the orchestration (contract composition); 3/ To monitor and detect the breaching of a QoS contract, possibly leading to a reconfiguration of the orchestration. Typically, the QoS of a service is modeled by a contract (or Service Level Agreement, SLA) between the provider and consumer of a given service. To account for variability, in previous years, we proposed soft probabilistic contracts specified as probabilistic distributions involving the different QoS parameters; we studied contract composition for such contracts; we developed probabilistic QoS contract monitoring; and we studied the monotonicity of orchestrations; an orchestration is monotonic if a called service improves its performance, then so does the overall orchestration.

This year, in the framework of the Associated Team FOSSA with the University of Texas at Austin (John Thywissen (PhD), Jayadev Misra and William Cook), we have extended our approach to general QoS parameters, i.e., beyond response time. We now encompass composite parameters, which are thus only partially, not totally, ordered. We have developed a general algebra to capture how QoS parameters are transformed while traversing the orchestration and we have extended our study of monotonicity. Finally, we have developed corresponding contract composition procedures. John Thywissen (from UT Austin) and Ajay Kattepur have started extending the Orc language and execution engine to support QoS according to our theory. This extension mainly consists in 1/ providing a rich type system to declare QoS domains and related algebra, and 2/ providing a new operator for Orc that allows for selecting competing returns from different sites on the basis of their QoS. A journal paper is under revision.

A key task in extending Orc for QoS was to extend the Orc engine so that causalities between the different site calls are made explicit at run time while execution progresses. This benefits from our previous work on Orc semantics, but a new set of rules has been proposed to generate causalities in an efficient way, by covering new features of the language. This is joint work of Claude Jard, Ajay Kattepur and John Thywissen from Austin. A publication is in preparation.

Besides this main line of work, other topics have been addressed by Ajay Kattepur as part of his thesis.

- In [41], we study variability of composite services. We model variability as a feature diagram (FD) that captures all valid configurations of its orchestration. Then, we apply pair-wise testing to sample the set of all possible configurations to obtain a concise subset. Finally, we test the composite service for selected pairwise configurations for a variety of QoS metrics such as response time, data quality, and availability. Using two case studies, Car crash crisis management and e-Health management, we demonstrate that pairwise generation effectively samples the full range of QoS variations in a dynamic orchestration. The pairwise sampling technique eliminates over 99% redundancy in configurations, while still calling all atomic services at least once.
• Web services orchestrations conventionally employ exhaustive comparison of runtime quality of service (QoS) metrics for decision making. The ability to incorporate more complex mathematical packages is needed, especially in case of workflows for resource allocation and queuing systems. By modeling such optimization routines as service calls within orchestration specifications, techniques such as linear programming can be conveniently invoked by non-specialist workflow designers. Leveraging on previously developed QoS theory, we propose the use of a high-level flexible query procedure for embedding optimizations in languages such as Orc. The Optima site provides an extension to the sorting and pruning operations currently employed in Orc. Further, the lack of an objective technique for consolidating QoS metrics is a problem in identifying suitable cost functions. We use the analytical hierarchy process (AHP) to generate a total ordering of QoS metrics across various domains. With constructs for ensuring consistency over subjective judgements, the AHP provides a suitable technique for producing objective cost functions. Using the Dell Supply Chain example, we demonstrate the feasibility of decision making through optimization routines, specially when the control flow is QoS dependent. This work was published in [39].

• With web services quality of service (QoS) modeled as random variables, the accuracy of sampled values for precise service level agreements (SLAs) come into question. Samples with lower spread are more accurate for calculating contractual obligations, which is typically not the case for web services QoS. Moreover, the extreme values in case of heavy-tailed distributions (e.g., 99.99 percentile) are seldom observed through limited sampling schemes. To improve the accuracy of contracts, we propose the use of variance reduction techniques such as importance sampling. We demonstrate this for contracts involving demand and refuel operations within the Dell supply chain example. Using measured values, efficient forecasting of future deviation of contracts may also be performed. A consequence of this is a more precise definition of sampling, measurement and variance tolerance in SLA declarations. This work was published in [40].

6.8. Specific studies: active documents and web services

Participants: Albert Benveniste, Loïc Hélouët, Benoît Masson.

Active Documents have been introduced by the GEMO team at INRIA Futurs, headed by Serge Abiteboul, mainly through the language Active XML (or AXML for short). AXML is an extension of XML which allows to enrich documents with service calls or sc’s for short. These sc’s point to web services that, when triggered, access other documents; this materialization of sc’s produces in turn AXML code that is included in the calling document. One therefore speaks of dynamic or intentional documents. In the past years, we have collaborated with the GEMO team to study a distributed version of their language.

This year, we have addressed the problem of distributed documents from a different point of view. Starting from our knowledge of distributed active XML (DAXML), we have first proposed a Petri Net semantics for a subset of DAXML [43], and then considered compositionality issues [54]. Compositionality in services can be addressed in several ways: first one have to ensure that modules that provide services an modules that use them agree on the data that they exchange. This notion is called composability of modules. However, composability does not ensure that a service always terminates (i.e. it returns a result to the caller) when it is invoked with appropriate data. Composability plus termination of services is called compatibility. We have shown that under some restrictions on the recursion in active documents, on the data, and upon the assumption that services use positive guards, composability is decidable. This work has also helped us isolate the core idea behind active documents, and propose a model for them called Docnets. Docnets are dynamic Petri nets which places are typed, which transitions are guarded computable type transformations, and which can receive new tokens from their environment. Docnet modules compose well, and if their closure by type transformation is finite, their compatibility is decidable. This work led to a publication [43].

Within the context of the DST associated team, we have proposed a new model, that combines arbitrary numbers of finite workflows, hence allowing for the definition of sessions. Sessions is a central paradigm in web-based systems. As messages exchange between two sites need not follow the same route over the net, a site can not rely on the identity of machines to uniquely define a transaction. This unique identification is
essential, as commercial site, for instance, need to manage several interactions at a given time. The current trend, as in BPEL, is to associate a unique identifier to each session. Modeling realistic session hence often forces to include session counters, and hence render most of properties undecidable. The session formalism studied in 2011 can be seen as a mix of BPEL and ORC elements, but was designed to keep several properties decidable. The strength of this formalism is to allow designing systems that use sessions without the obligation to provide identifiers. The formalism has the expressive power of reset Petri nets for which coverability is decidable. This is sufficient to decide whether a set of agent can be found in some bad configuration during the lifetime of a system. This joint work with Ph. Darondeau from the S4 Team, and with M. Mukund from the Chennai Mathematical Institute led to a publication in the ATVA conference [28].

Our last work on Web-services was the development of an experimental platform. During his post-doc, Benoît Masson has designed a distributed Active XML engine, which can be distributed over a network. We have built a lightweight experimentation platform, made of four linux machines, that run DAXML services and communicate with one another. Simultaneously, R. Abdallah has designed a synthesis tool to generate REST services from High-level Message Sequence Charts. These services were successfully tested on the platform.

6.9. Specific studies: security and privacy

Participants: Guillaume Aucher, Blaise Genest.

We have worked on three parallel lines of research related to security and privacy. The first line deals with problems of delegation and revocation in distributed systems. The second line deals with problems of compliance of a system with respect to a privacy regulation expressed in a language combining epistemic, deontic and dynamic modalities. The third line tackles the minimal information needed at runtime to e.g. break in a (stochastic) system.

6.9.1. Delegation and revocation in distributed systems

Together with Steve Barker from King’s College London, Guido Boella from the University of Torino, Valerio Genovese and Leon van der Torre from the University of Luxembourg, we defined a (sound and complete) propositional dynamic logic to specify and reason about delegation and revocation schemes in distributed systems. This logic describes formally a family of delegation and revocation models that are based on the work of [65]. We extended our logic to accommodate an epistemic interpretation of trust. What emerges from this work is a rich framework of formally well-defined delegation and revocation schemes that accommodates an important trust component. In particular, we showed how to automatically reason about whether an agent is authorized to do an operation on an object and about the authorization policy resulting from the execution of a sequence of actions. We used our logical framework to give a formal account of eight different types of revocation schemes informally introduced in previous literature. This work is published in [18].

6.9.2. Privacy policy with modal logic: the dynamic turn

As explained in Section 6.4, we want to define a logical language to specify privacy policies which is close to the natural language. In general, privacy policies can be defined either in terms of permitted and forbidden knowledge, or in terms of permitted and forbidden actions. For example, it may be forbidden to know the medical data of a person, or it may be forbidden to disclose these data. Implementing a privacy policy based on permitted and forbidden knowledge is relatively easy, since we can add a filter on the system checking the outgoing messages. Such a filter is an example of a security monitor. If the system attempts to send a forbidden message, then the security monitor blocks the sending of that message. However, the price to pay for this relatively straightforward implementation is that it is difficult to decide which actions are permitted or forbidden so that a piece of information is not disclose. We are therefore interested in privacy policies expressed in terms of permitted and forbidden knowledge. Expressing a privacy policy in terms of permitted and forbidden knowledge is relatively easy, since it lists the situations, where, typically, it may not be permitted to know some sensitive information. Implementing a privacy policy based on permitted and forbidden knowledge is quite difficult, since the system has to reason about the relation between permitted knowledge and actions. The challenge is that the exchange of messages changes the knowledge, and the
security monitor therefore needs to reason about these changes. This inference problem is already non trivial with a static privacy policy, and becomes challenging when privacy policies can change over time. Together with Guido Boella and Leon van der Torre, we therefore introduced a dynamic modal logic that permits not only to reason about permitted and forbidden knowledge to derive the permitted actions, but also to represent explicitly the declarative privacy policies together with their dynamics. The logic can be used to check both regulatory and behavioral compliance, respectively by checking that the permissions and obligations set up by the security monitor of an organization are not in conflict with the privacy policies, and by checking that these obligations are indeed enforced. We also showed that the complexity of the model checking problem is quadratic in the size of the model and the formula and provided the corresponding model-checking algorithms. This work is published in [11].

6.9.3. Minimal information needed

Together with Nathalie Bertrand from Vertecs, we tackle the problem of the minimal information a user needs at runtime to achieve a simple goal, modeled as reaching an objective with probability one [25]. The natural question is then to minimize the additional information the user needs to fulfill her objective. This optimization question gives rise to two different problems, whether we consider to minimize the worst case cost, or the average cost. On the one hand, concerning the worst case cost, we show that efficient techniques from the model checking community can be adapted to compute the optimal worst case cost and give optimal strategies for the users. On the other hand, we show that the optimal average price (a question typically considered in the AI community) cannot be computed in general, nor can it be approximated in polynomial time even up to a large approximation factor. Following this negative results, we investigate with P.S. Thiagarajan’s group at NUS, Singapore basic algorithms of the AI community to infer the exact probability in (compact) stochastic systems. We proposed in [45] a simple parametrized extension of the usual Factored Frontier algorithm in order to choose the desired accuracy of the algorithm, at the cost of additional but manageable computations. We showed its benefit when dealing with biological pathways.

6.10. Specific studies: network maintenance

Participants: Eric Fabre, Carole Hounkonnou.

This work represents part of our activities within the research group “High Manageability,” supported by the common lab of Alcatel-Lucent Bell Labs (ALBLF) and INRIA. It concerns a methodology for the graceful shut down and restart of routers in OSPF networks, one of the core protocols of IP networks. A methodology has been proposed to safely switch off the software layer of a router while still maintaining this router in the forwarding plane: the router still forwards packets, but is not able to adapt its routing table to changes in network conditions or topology. Nevertheless, it is possible to check whether this frozen router is harmless or can cause packet losses, through a centralized or distributed algorithm. And if ever it puts the network at risk, minimal patches can be set up temporarily until the router comes back to normal activity. This avoids running twice a global OSPF update at all nodes (once for shutdown of the equipment, one for restart). There is a patent project on this activity, that we don’t detail more here.

6.11. Specific studies: network and service diagnosis

Participants: Eric Fabre, Carole Hounkonnou.

This work represents part of our activities within the research group “High Manageability,” supported by the common lab of Alcatel-Lucent Bell Labs (ALBLF) and INRIA. It is also supported by the UniverSelf EU integrated project, and conducted in relation with Orange Labs.

The objective is to develop a framework for the joint diagnosis of networks and of the supported services. We are aiming at a model-based approach, in order to tailor the methods to a given network instance and to follow its evolution. We also aim at active diagnosis methods, that collect and reason on alarms provided by the network, but that can also trigger tests or the collection of new observations in order to refine a current diagnosis.
In 2011, the main effort was dedicated to a key and difficult part of this approach: the definition of a methodology for self-modelling. This consists in automatically building a model of the monitored system, by instantiating generic network elements. There are several difficulties to address:

- The model must capture several layers, from the physical architecture up to the service architecture and its protocols. As a case-study, we have chosen VoIP services on an IMS network, deployed over a wired IP network.
- The model should be hierarchical, to allow for multiscale reasoning, and to reflect the intrinsic hierarchical nature of the managed network.
- The model should be generic, i.e. obtained by assembling component instances coming from a reduced set of patterns, just like a text is obtained by assembling words.
- The model should be adaptive, to capture the evolving part of the network (e.g. introduction of new elements) but also its intrinsically dynamic nature (e.g. opened/closed connections).
- The model should display the hierarchical dependency of resources, specifically the fact that lower-level resources are assembled to provide a support to a higher level resource or functionality.
- The model should allow progressive discovery and refinement: for a matter of size, it is not possible to first build a model of the complete network and then monitor it; one must adopt an approach where the model is build on-line, and where the construction is guided by the progress of the diagnosis algorithms.

The first elements of a methodology achieving these objectives have been designed in 2011. The next efforts will aim at refining the grammar of this model, for our specific case study, and at developing the dedicated diagnosis algorithms. For the latter, we envision a new setting of hierarchical and generic Bayesian networks, in order to capture the dependencies between network elements at different granularities.
DNET Team

6. New Results

6.1. Exposure to diffusion in dynamic networks

In many contexts, complex networks are subject to diffusion phenomenon, like spreading of epidemics in human groups or the diffusion of information in social networks. Often, the underlying network is dynamic, that is, its links change along time. Clearly, the dynamics of links has an influence on the diffusion phenomenon taking place over the network. A first step to understand these relationship is to determine which nodes of the dynamic network are more likely to be reached by a diffusion process. We designed new notions of exposure in order to do it, based either on contacts, paths or flows in a dynamic network. In particular, the notion of dynamic flow, which we introduced, has given interesting preliminary results. We computed the exposition scores of nodes of real world dynamic networks and showed that it is correlated to the likelihood of nodes to be affected by a diffusion in the classical SI model.

6.2. Aggregation of temporal graph series

A very natural and extensively used way to represent a dynamic network, where links change along time, is to build a graph series : the series of snapshots of the network taken at different time of its evolution. The way to do so is to aggregate all the contact information on a time window into a single graph : that is, we put an edge between u and v in the graph if they are in contact at least once during the considered time window. Doing so for disjoint windows of equal length which cover the whole period of study, we obtain a series of graphs representing the dynamics of the network. A question remain : how one should choose the length of the aggregation window? The problem is critical since depending on the choice made, the properties of the dynamic network are different and the conclusion derived from its analysis may change. We design a systematic method to estimate the maximum possible aggregation length. Up to our knowledge, this is the first method addressing the problem. It is based on activity rate of dynamic paths in the dynamics. On a dynamic path, only some time steps are used to move within the network. When the aggregation time is short, the activity rate of paths is close to zero and it tends to 1 when this time grow until the whole period of experiments. Between the two behaviors, we showed that there is a phase transition that we interpret as the moment when the properties of the dynamics are distorted because of the too long aggregation time.

6.3. Characterizing changes in dynamic networks

Very often, dynamic networks are described as time series of graphs. Many works focus on analyzing or capturing into models the properties of the graphs of the series. This approach has a clear limitation : it looses the relationships between the different graphs of the series, which however contain a key information on the dynamics. In order to get more insight in the relationships between the graphs of the series, we analyzed the structure of we call the difference graphs. The difference graph of two consecutive graphs $G_1$, $G_2$ of the series is the graph whose edge set is the symmetric difference of the edge sets of $G_1$ and $G_2$. In other words, this is exactly the graph of the pairs whose adjacency relationship changed from $G_1$ to $G_2$. We showed that the structure of difference graphs is very particular : their edges are concentrated around a small number of vertices. This shows that the changes between two graphs of the series are not spread everywhere in the network, but are due to changes of the neighborhood of only a small number of nodes of the network. We could show this fact by computing a graph parameter called Minimum Vertex Cover (MVC), which is Np-complete to compute. Using a preprocessing step, we could compute the exact value of this parameter for all difference graphs of real world series. We obtained that the value of the MVC on difference graphs is very small compared to the expected value on a random graph with same density. We also showed that the most common models of dynamic networks do not capture this property of concentration of edges in the difference graphs of the series. Our result shed light on the way dynamic networks evolve and open the way to significant improvement of existing models.
6.4. Community detection: dynamic and overlapping

Overlapping community detection is a popular topic in complex networks. Comparing to disjoint community structure, overlapping community structure is more reasonable to describe networks at a macroscopic level. Overlaps shared by communities play an important role in combining different communities. We propose two different approaches to detect overlaps: fuzzy community detection and overlapping community detection. The former estimates membership degree of node belonging to community, and the latter allows node to be shared by communities. In this paper, a fuzzy detection and a clique optimization are introduced. Experimental studies in synthetic networks show fuzzy detection yields meaningful information in stability and hierarchy of communities. And clique optimization is efficient in capturing overlapping nodes. Applications in real networks whose community structure is not well-known find that overlapping clusters found by our fuzzy detection can provide different views than general overlapping nodes in characterize overlaps.

Although community detection has drawn tremendous amount of attention across the sciences in the past decades, no formal consensus has been reached on the very nature of what qualifies a community as such. We take an orthogonal approach by introducing a novel point of view to the problem of overlapping communities. Instead of quantifying the quality of a set of communities, we choose to focus on the intrinsic community-ness of one given set of nodes. To do so, we propose a general metric on graphs, the cohesion, based on counting triangles and inspired by well established sociological considerations. The model has been validated through a large-scale online experiment called Fellows in which users were able to compute their social groups on Facebook and rate the quality of the obtained groups. By observing those ratings in relation to the cohesion we assess that the cohesion is a strong indicator of users subjective perception of the community-ness of a set of people.

6.5. Cross-Layer Optimization for Software Layer to Physical Device layer Mapping

We develop a generic method for mapping software state machines used in protocol stacks and communication layers directly to hardware communication devices using their specifications. The proposed method can handle power modes and timing constraints imposed by hardware devices in order to optimize the software code running on top of the device. This property allows the use of the hardware device in its lowest power consumption mode while making sure that real time constraints are met. To validate the merit of the proposed method, the generated code and power consumption gain, we evaluate the optimizations that can be done on a BMAC medium access control layer used in wireless sensor networks using a large scale experimental testbed. The results show that an average energy consumption gain of up to 60% at the radio level can be achieved.
4. New Results

4.1. Understanding graph representations

4.1.1. Distributed algorithms without knowledge of global parameters

Participants: Amos Korman, Jean-Sébastien Sereni, Laurent Viennot.

Many fundamental local distributed algorithms are non-uniform, that is, they assume that all nodes know good estimations of one or more global parameters of the network, e.g., the maximum degree \( \Delta \) or the number of nodes \( n \). In [28], we introduce a rather general technique for transforming a non-uniform algorithm into a uniform one with same asymptotic complexity.

4.1.2. Asymptotic modularity

Participants: Fabien de Montgolfier, Mauricio Soto, Laurent Viennot.

Modularity has been introduced as a quality measure for graph partitioning by Newman and Girvan. It has received considerable attention in several disciplines, especially complex systems. In order to better understand this measure from a graph theoretical point of view, we study in [32], [31] the asymptotic modularity of a variety of graph classes.

4.1.3. Internet Structure

Participants: Fabien de Montgolfier, Mauricio Soto, Laurent Viennot.

In [33], [1], we study the measurement of the Internet according to two graph parameters: treewidth and hyperbolicity.

4.1.4. Multipath Spanners

Participants: Cyril Gavoille, Quentin Godfroy, Laurent Viennot.

Motivated by multipath routing, we introduce in [23], [39] a multi-connected variant of spanners.

4.1.5. \( \delta \)-hyperbolicity

Participants: Victor Chepoi [CNRS LIF, University of Marseille, France], Feodor Dragan [University of Ohio, USA], Bernard Estrellon [CNRS LIF, University of Marseille, France], Michel Habib [CNRS LIAFA, University of Paris Diderot, France], Yann Vaxes [University of Florence, Italy], Yang Xiang [University of Ohio, USA].

\( \delta \)-Hyperbolic metric spaces have been defined by M. Gromov in 1987 via a simple 4-point condition: for any four points \( u, v, w, x \), the two larger of the distance sums \( d(u, v) + d(w, x), d(u, w) + d(v, x), d(u, x) + d(v, w) \) differ by at most \( 2\delta \). They play an important role in geometric group theory, geometry of negatively curved spaces, and have recently become of interest in several domains of computer science, including algorithms and networking. In [5] paper, we study un-weighted \( \delta \)-hyperbolic graphs. Using the Layering Partition technique, we show that every \( n \)-vertex \( \delta \)-hyperbolic graph with \( \delta \geq 1/2 \) has an additive \( O(\delta \log n) \)-spanner with at most \( O(\delta n) \) edges and provide a simpler, in our opinion, and faster construction of distance approximating trees of \( \delta \)-hyperbolic graphs with an additive error \( O(\delta \log n) \). The construction of our tree takes only linear time in the size of the input graph. As a consequence, we show that the family of \( n \)-vertex \( \delta \)-hyperbolic graphs with \( \delta \geq 1/2 \) admits a routing labeling scheme with \( O(\delta \log^2 n) \) bit labels, \( O(\delta \log n) \) additive stretch and \( O(\log^2 (4\delta)) \) time routing protocol, and a distance labeling scheme with \( O(\log^2 n) \) bit labels, \( O(\delta \log n) \) additive error and constant time distance decoder.
4.1.6. Perfect Phylogeny

4.1.6.1. Perfect Phylogeny Is NP-hard

Participants: Michel Habib [CNRS LIAFA, University of Paris Diderot, France], Juraj Stacho [University of Haifa, Israel].

We answer in the affirmative [24], to the question posed by Mike Steel as a $100 challenge: “Is the following problem NP-hard? Given a ternary phylogenetic $X$-tree $T$ and a collection $Q$ of quartet subtrees on $X$, is $T$ the only tree that displays $Q$?” As a particular consequence of this, we show that the unique chordal sandwich problem is also NP-hard.

4.1.6.2. Compatibility of Multi-states Characters

Participants: Michel Habib [CNRS LIAFA, University of Paris Diderot, France], Thu-Hien To [CNRS LIAFA, University of Paris Diderot, France].

Perfect phylogeny consisting of determining the compatibility of a set of characters is known to be NP-complete. We propose in [25], a conjecture on the necessary and sufficient conditions of compatibility: Given a set $C$ of $r$-states full characters, there exists a function $f(r)$ such that $C$ is compatible if $f$ every set of $f(r)$ characters of $C$ is compatible. According to numerous references, $f(2) = 2$, $f(3) = 3$ and $f(r) \geq r$. Some conjectured that $f(r) = r$ for any $r \geq 2$. In this paper, we present an example showing that $f(4) \geq 5$. Therefore it could be the case that for $r \geq 4$ characters, the problem behavior drastically changes. In a second part, we propose a closure operation for chordal sandwich graphs. The later problem is a common approach of perfect phylogeny.

4.1.7. Graph sandwich

Participants: Arnaud Durand [CNRS LIAFA, University of Paris Diderot, France], Michel Habib [CNRS LIAFA, University of Paris Diderot, France].

Graph sandwich problems were introduced by Golumbic et al. (1994) in [12] for DNA physical mapping problems and can be described as follows. Given a property $\Pi$ of graphs and two disjoint sets of edges $E_1, E_2$ with $E_1 \subseteq E_2$ on a vertex set $V$, the problem is to find a graph $G$ on $V$ with edge set $E_s$ having property $\Pi$ and such that $E_1 \subseteq E_s \subseteq E_2$. In [8] paper, we exhibit a quasi-linear reduction between the problem of finding an independent set of size $k \geq 2$ in a graph and the problem of finding a sandwich homogeneous set of the same size $k$. Using this reduction, we prove that a number of natural (decision and counting) problems related to sandwich homogeneous sets are hard in general. We then exploit a little further the reduction and show that finding efficient algorithms to compute small sandwich homogeneous sets would imply substantial improvement for computing triangles in graphs.

4.1.8. Diameter of Real-World Undirected Graphs

Participants: Pierluigi Crescenzi [University of Florence, Italy], Roberto Grossi [University of Pisa, Italy], Michel Habib [CNRS LIAFA, University of Paris Diderot, France], Lorenzo Lanzi [University of Florence, Italy], Andrea Marino [University of Florence, Italy].

In [16], we propose a new algorithm for computing the diameter of undirected unweighted graphs. Even though, in the worst case, this algorithm has complexity $O(nm)$, where $n$ is the number of nodes and $m$ is the number of edges of the graph, we experimentally show (on almost 200 real-world graphs) that in practice our method works in linear time. Moreover, we show how to extend our algorithm to the case of undirected weighted graphs and, even in this case, we present some preliminary very positive experimental results.

4.1.9. Parsimonious flooding in dynamic graphs

Participants: Hervé Baumann [CNRS LIAFA, University of Paris Diderot, France], Pierluigi Crescenzi [University of Florence, Italy], Pierre Fraigniaud [CNRS LIAFA, University of Paris Diderot, France].
An edge-Markovian process with birth-rate \( p \) and death-rate \( q \) generates infinite sequences of graphs \( (G_0, G_1, G_2, \ldots) \) with the same node set \([n]\) such that \( G_t \) is obtained from \( G_{t-1} \) as follows: if \( e \notin E(G_{t-1}) \) then \( e \in E(G_t) \) with probability \( p \), and if \( e \in E(G_{t-1}) \) then \( e \notin E(G_t) \) with probability \( q \). In [2], we establish tight bounds on the complexity of flooding in edge-Markovian 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 \( t' > t \). These bounds complete previous results obtained by Clementi et al. Moreover, we also show that flooding in dynamic graphs can be implemented in a parsimonious manner, so that to save bandwidth, yet preserving efficiency in term of simplicity and completion time. For a positive integer \( k \), we say that the flooding protocol is \( k \)-active if each node forwards an information only during the \( k \) time steps immediately following the step at which the node receives that information for the first time. We define the reachability threshold for the flooding protocol as the smallest integer \( k \) such that, for any source \( s[n] \), the \( k \)-active flooding protocol from \( s \) completes (i.e., reaches all nodes), and we establish tight bounds for this parameter. We show that, for a large spectrum of parameters \( p \) and \( q \), the reachability threshold is by several orders of magnitude smaller than the flooding time. In particular, we show that it is even constant whenever the ratio \( p/(p + q) \) exceeds \( \log n/n \). Moreover, we also show that being active for a number of steps equal to the reachability threshold (up to a multiplicative constant) allows the flooding protocol to complete in optimal time, i.e., in asymptotically the same number of steps as when being perpetually active. These results demonstrate that flooding can be implemented in a practical and efficient manner in dynamic graphs. The main ingredient in the proofs of our results is a reduction lemma enabling to overcome the time dependencies in edge-Markovian dynamic graphs.

4.2. Distributed computational complexities

4.2.1. Local Distributed Decision

Participants: Pierre Fraigniaud [CNRS LIAFA, University of Paris Diderot, France], Amos Korman [CNRS LIAFA, University of Paris Diderot, France], David Peleg [Weizmann Institute of Science, Israel].

Inspired by sequential complexity theory, in [20] we focus on a complexity theory for distributed decision problems. We first study the intriguing question of whether randomization helps in local distributed computing, and to what extent. Our main result provides a sharp threshold for the impact of randomization on decision hereditary problems. In addition, we investigate the impact of non-determinism on local decision, and establish some structural results inspired by classical computational complexity theory. Specifically, we show that non-determinism does help, but that this help is limited, as there exist languages that cannot be decided non-deterministically. Perhaps surprisingly, it turns out that it is the combination of randomization with non-determinism that enables to decide all languages in constant time. Finally, we introduce the notion of local reduction, and establish some completeness results.

4.2.2. Asynchronous Wait-free Decision

Participants: Pierre Fraigniaud [CNRS LIAFA, University of Paris Diderot, France], Sergio Rajsbaum [Maths. Institute, University of Mexico, Mexico], Corentin Travers [Technion, Israel].

In order to capture the core of asynchronous distributed decision model, we address in [22] the wait-free model with crash failures. The set of tasks whose input is a pair \( (s, t) \) and deciding whether \( t \in \Delta(s) \), i.e. whether \( t \) is a valid output for \( s \), has been proven to be decidable in this model.

4.2.3. Mobile Distributed Decision

Participants: Pierre Fraigniaud [CNRS LIAFA, University of Paris Diderot, France], Andrzej Pelc [UQO, University of Quebec, Canada].

In [21], we partially answer the question of decidability of any language for mobile agents in a 2D environment like telecom networks or robots. It is proven that, for every agent, verifying whether (i) he/she is alone or not and (ii) he/she is able to capture the environment, is associated with the question of pertaining to an equivalence class of a map. A positive answer helps in the non-deterministic decision for any language for mobile agent.
4.2.4. Approximating the Statistics of various Properties in Randomly Weighted Graphs

Participants: Yuval Emek [University of Tel Aviv, Israel], Amos Korman [CNRS LIAFA, University of Paris Diderot, France], Yuval Shavitt [University of Tel Aviv, Israel].

In [19], we consider the setting of randomly weighted graphs. Under this setting, weighted graph properties typically become random variables and we are interested in computing their statistical features. Unfortunately, this turns out to be computationally hard for some weighted graph properties albeit the problem of computing the properties per se in the traditional setting of algorithmic graph theory is tractable. For example, there are well known efficient algorithms that compute the diameter of a given weighted graph, yet, computing the expected diameter of a given randomly weighted graph is \#P-hard even if the edge weights are identically distributed. In this paper, we define a family of weighted graph properties and show that for each property in this family, the problem of computing the \( k \)'th moment (and in particular, the expected value) of the corresponding random variable in a given randomly weighted graph \( G \) admits a fully polynomial time randomized approximation scheme (FPRAS) for every fixed \( k \). This family includes fundamental weighted graph properties such as the diameter of \( G \), the radius of \( G \) (with respect to any designated vertex) and the weight of a minimum spanning tree of \( G \).

4.2.5. New bounds for the controller problem

Participants: Yuval Emek [University of Tel Aviv, Israel], Amos Korman [CNRS LIAFA, University of Paris Diderot, France].

In [10], we establish two new lower bounds on the message complexity of the controller problem. We first prove a simple lower bound stating that any \((M, W)\)-controller must send \( \Omega(N \log \frac{M}{W+1}) \) messages. Second, for the important case when \( W \) is proportional to \( M \) (this is the common case in most applications), we use a surprising reduction from the (centralized) monotonic labeling problem to show that any \((M, W)\)-controller must send \( \Omega(N \log N) \) messages. In fact, under a long lasting conjecture regarding the complexity of the monotonic labeling problem, this lower bound is improved to a tight \( \Omega(N \log^2 N) \).

4.2.6. Online computation with advice

Participants: Yuval Emek [University of Tel Aviv, Israel], Pierre Fraigniaud [CNRS LIAFA, University of Paris Diderot, France], Amos Korman [CNRS LIAFA, University of Paris Diderot, France], Adi Rosén [CNRS LIAFA, University of Paris Diderot, France].

In [9], we consider a model for online computation in which the online algorithm receives, together with each request, some information regarding the future, referred to as advice. We are interested in the impact of such advice on the competitive ratio, and in particular, in the relation between the size \( b \) of the advice, measured in terms of bits of information per request, and the (improved) competitive ratio. In this paper we propose the above model and illustrate its applicability by considering two of the most extensively studied online problems, namely, metrical task systems (MTS) and the \( k \)-server problem.

4.2.7. Tight Bounds For Distributed MST Verification

Participants: Liah Kor [Weizmann Institute of Science, Israel], Amos Korman [CNRS LIAFA, University of Paris Diderot, France], David Peleg [Weizmann Institute of Science, Israel].

In [26], we establishes tight bounds for the Minimum-weight Spanning Tree (MST) verification problem in the distributed setting. Specifically, we provide an MST verification algorithm that achieves simultaneously \( \tilde{O}(|E|) \) messages and \( O(\sqrt{n} + D) \) time, where \( |E| \) is the number of edges in the given graph \( G \) and \( D \) is \( G \)'s diameter. On the negative side, we show that any MST verification algorithm must send \( \tilde{O}(|E|) \) messages and incur \( \tilde{O}(\sqrt{n} + D) \) time in worst case. Our upper bound result appears to indicate that the verification of an MST may be easier than its construction.

4.2.8. Distributed verification and hardness of distributed approximation

Participants: Atish Das Sarma [Google research, USA], Stephan Holzer [ETH, Zurich, Switzerland], Liah Kor [Weizmann Institute of Science, Israel], Amos Korman [CNRS LIAFA, University of Paris Diderot, France], Danupon Nanongkai [Nanyang Technological University, Singapore], David Peleg [Weizmann Institute of Science, Israel], Roger Wattenhofer [ETH, Zurich, Switzerland].
In [30], we initiate a systematic study of distributed verification, and give almost tight lower bounds on the running time of distributed verification algorithms for many fundamental problems such as connectivity, spanning connected subgraph, and $s-t$ cut verification. We then show applications of these results in deriving strong unconditional time lower bounds on the hardness of distributed approximation for many classical optimization problems including minimum spanning tree, shortest paths, and minimum cut. Many of these results are the first non-trivial lower bounds for both exact and approximate distributed computation and they resolve previous open questions. Moreover, our unconditional lower bound of approximating minimum spanning tree (MST) subsumes and improves upon the previous hardness of approximation bound of Elkin [STOC 2004] as well as the lower bound for (exact) MST computation of Peleg and Rubinovich [FOCS 1999]. Our result implies that there can be no distributed approximation algorithm for MST that is significantly faster than the current exact algorithm, for any approximation factor. Our lower bound proofs show an interesting connection between communication complexity and distributed computing which turns out to be useful in establishing the time complexity of exact and approximate distributed computation of many problems.

4.3. Peer to Peer Networks Performance

**Participants:** Fabien Mathieu, François Baccelli.

In [3], we present and discuss possible architectures for P2P systems to manage overlays that try to cope with the underlying network.

In [40], [29], we discuss theoretical performance issues that arise from using “Live Seeding”, a technique that can be employed to leverage the capacity of a P2P/Hybrid Live Streaming Systems by utilizing the capacities of idle peers.

In [38], we propose a new paradigm for P2P networks, where the bandwidth bottleneck is not the access node anymore. This new model is versatile enough to be used in the context of classical networks with congestion control, wireless networks, or semantic networks.

4.4. Fault Tolerance in Distributed Networks

4.4.1. Verification of population protocols

**Participants:** Hugues Fauconnier, Carole Gallet-Delpporte.

In [15], we address the problem of verification by model-checking of the basic population protocol (PP) model of Angluin et al. This problem has received special attention in the last two years and new tools have been proposed to deal with it. We show that the problem can be solved by using the existing model-checking tools, e.g., Spin and Prism. In order to do so, we apply the counter abstraction to get an abstraction of the PP model which can be efficiently verified by the existing model-checking tools. Moreover, this abstraction preserves the correct stabilization property of PP models. To deal with the fairness assumed by the PP models, we provide two new recipes. The first one gives sufficient conditions under which the PP model fairness can be replaced by the weak fairness implemented in Spin. We show that this recipe can be applied to several PP models. In the second recipe, we show how to use probabilistic model-checking and, in particular, Prism to take completely in consideration the fairness of the PP models. The correctness of this recipe is based on existing theorems involving finite discrete Markov chains. An abstract of this paper has been also published in [34].

4.4.2. Failure Detection

**Participants:** Hugues Fauconnier, Carole Gallet-Delpporte.

What does it mean to solve a distributed task? In Paxos, Lamport proposed a definition of solvability in which every process is split into a proposer that submits commands to be executed, an acceptor that takes care of the command execution order, and a learner that receives the outcomes of executed commands. The resulting perspective of computation in which every proposed command can be executed, be its proposer correct or faulty, proved to be very useful when processes take steps on behalf of each other, i.e., in simulations.
Most interesting tasks cannot be solved asynchronously, and failure detectors were proposed to circumvent these impossibilities. Alas, when it comes to solving a task using a failure detector, we cannot leverage simulation-based techniques. A process cannot perform steps of failure detector-based computation on behalf of another process, since it cannot access the remote failure-detector module.

In [17], we propose a new definition of solving a task with a failure detector in which computation processes that propose inputs and provide outputs are treated separately from synchronization processes that coordinate using a failure detector. In the resulting framework, any failure detector is shown to be equivalent to the availability of some $k$-set agreement. As a corollary, we obtain a complete classification of tasks, including ones that evaded comprehensible characterization so far, such as renaming.

Shared objects like atomic register, test-and-set, cmp-and-swap are classical hardware primitives that help to develop fault-tolerant distributed applications. In order to compare shared objects, in [41], we consider their implementations in message passing models. With the minimal failure detector for each object, we get a new hierarchy that has only two levels. This paper summarizes recent works and results on this topic.

In [7], we first define the basic notions of local and non-local tasks for distributed systems. Intuitively, a task is local if, in a system with no failures, each process can compute its output value locally by applying some local function on its own input value (so the output value of each process depends only on the process’ own input value, not on the input values of the other processes); a task is non-local otherwise. All the interesting distributed tasks, including all those that have been investigated in the literature (e.g., consensus, set agreement, renaming, atomic commit, etc.) are non-local.

In this paper we consider non-local tasks and determine the minimum information about failures that is necessary to solve such tasks in message-passing distributed systems. As part of this work, we also introduce weak set agreement — a natural weakening of set agreement — and show that, in some precise sense, it is the weakest non-local task in message-passing systems.

4.4.3. Adversary disagreement and Byzantine agreement

Participants: Hugues Fauconnier, Carole Gallet-Delporte.

At the heart of distributed computing lies the fundamental result that the level of agreement that can be obtained in an asynchronous shared memory model where $t$ processes can crash is exactly $t + 1$. In other words, an adversary that can crash any subset of size at most $t$ can prevent the processes from agreeing on $t$ values. But what about all the other $2^{2n-1} - (n+1)$ adversaries that are not uniform in this sense and might crash certain combination of processes and not others? In [6], we present a precise way to classify all adversaries. We introduce the notion of disagreement power: the biggest integer $k$ for which the adversary can prevent processes from agreeing on $k$ values. We show how to compute the disagreement power of an adversary and derive $n$ equivalence classes of adversaries.

So far, the distributed computing community has either assumed that all the processes of a distributed system have distinct identifiers or, more rarely, that the processes are anonymous and have no identifiers. These are two extremes of the same general model: namely, $n$ processes use $\ell$ different authenticated identifiers, where $1 \leq \ell \leq n$. In [18], we ask how many identifiers are actually needed to reach agreement in a distributed system with $t$ Byzantine processes.

We show that having $3t + 1$ identifiers is necessary and sufficient for agreement in the synchronous case but, more surprisingly, the number of identifiers must be greater than $\frac{n+3t}{2}$ in the partially synchronous case. This demonstrates two differences from the classical model (which has $\ell = n$): there are situations where relaxing synchrony to partial synchrony renders agreement impossible; and, in the partially synchronous case, increasing the number of correct processes can actually make it harder to reach agreement. The impossibility proofs use the fact that a Byzantine process can send multiple messages to the same recipient in a round. We show that removing this ability makes agreement easier: then, $t + 1$ identifiers are sufficient for agreement, even in the partially synchronous model.
4.4.4. Fast and compact self stabilizing verification, computation, and fault detection of an MST

Participants: Amos Korman [CNRS LIAFA, University of Paris Diderot, France], Shay Kutten [Technion, Israel], Toshimitsu Masuzawa [Osaka University, Japan].

In [27], we address the impact of optimizing the memory size on the time complexity, and show that this carries at most a small cost in terms of time in the context of MST. Specifically, we present a self stabilizing distributed verification algorithm whose time complexity is $O(\log^2 n)$ in synchronous networks, or $O(\Delta \log^2 n)$ in asynchronous networks, where $\Delta$ denotes the largest degree of a node. More importantly, the memory size at each node remains optimal - $O(\log n)$ bits throughout the execution. This answers an open problem posed by Awerbuch and Varghese (FOCS 1991). We also show that $\Omega(\log n)$ time is necessary if the memory size is restricted to $O(\log n)$ bits, even in synchronous networks. We demonstrate the usefulness of our verification scheme by using it as a module in a new self stabilizing MST construction algorithm. This algorithm has the important property that, if faults occur after the construction ended, they are detected by some nodes within $O(\log^2 n)$ time in synchronous networks, or within $O(\Delta \log^2 n)$ time in asynchronous networks. The rest of the nodes detect within $O(D \log n)$ time, where $D$ denotes the diameter. Moreover, if a constant number of faults occur, then, within the required detection time above, they are detected by some node in the $O(\log n)$ locality of each of the faults. The memory size of the self stabilizing MST construction is $O(\log n)$ bits per node (optimal), and the time complexity is $O(n)$. This time complexity is significantly better than the best time complexity of previous self stabilizing MST algorithms, that was $\Omega(n^2)$ even without the above localized fault detection property. The time complexity of previous algorithms that used $O(\log n)$ memory size was $O(n|E|)$.

4.5. Discrete Optimization Algorithms

4.5.1. Estimating Satisfiability

Participants: Yacine Boufkhad, Thomas Hugel.

In [4], the problem of estimating the proportion of satisfiable instances of a given CSP (constraint satisfaction problem) can be tackled through weighting. It consists in putting onto each solution a non-negative real value based on its neighborhood in a way that the total weight is at least 1 for each satisfiable instance. We define in this paper a general weighting scheme for the estimation of satisfiability of general CSPs. First we give some sufficient conditions for a weighting system to be correct. Then we show that this scheme allows for an improvement on the upper bound on the existence of non-trivial cores in 3-SAT obtained by Maneva and Sinclair (2008) to 4.419. Another more common way of estimating satisfiability is ordering. This consists in putting a total order on the domain, which induces an orientation between neighboring solutions in a way that prevents circuits from appearing, and then counting only minimal elements. We compare ordering and weighting under various conditions.

4.5.2. Eigenvectors of three term recurrence Toeplitz matrices and Riordan group

Participant: Dominique Fortin.

Eigenvalues of tridiagonal (including main) Toeplitz matrices are analytically known under some regular distance to the main diagonal. Any eigenvector may be easily computed then, through a backward process; instead, in [11], we give an analytical form for each component through the reciprocation of the underlied trinomial. More generally, the connection to the Riordan group follows some bilinear iterative process.

4.5.3. Piecewise Convex Maximization problems and algorithms

Participants: Dominique Fortin, Ider Tseveendorj.

In [14], we provide a global search algorithm for maximizing a piecewise convex function $F$ over a compact $D$. We propose to iteratively refine the function $F$ at local solution $y$ by a virtual cutting function $p_y(\cdot)$ and to solve
max\{\min \{F(x) - F(y), p_y(x)\} \mid x \in D\} instead. We call this function either a patch, when it avoids returning back to the same local solutions, or a pseudo patch, when it possibly yields a better point. It is virtual in the sense that the role of cutting constraints is played by additional convex pieces in the objective function. We report some computational results, that represent an improvement on previous linearization based techniques.

It is well known that maximization of any difference of convex functions could be turned into a convex maximization; in [13], we aim at a piecewise convex maximization problem instead. Despite, it may seem harder, sometimes the dimension may be reduced by 1 and the local search improved by using extreme points of the closure of the convex hull of better points. We show that it is always the case for both binary and permutation problems and give, as such instances, piecewise convex formulations for the maximum clique problem and the quadratic assignment problem.

in [12], we consider mathematical programming problems with the so-called piecewise convex objective functions. A solution method for this interesting and important class of nonconvex problems is presented. This method is based on Newton’s law of universal gravitation, multicriteria optimization and Helly’s theorem on convex bodies. Numerical experiments using well known classes of test problems on piecewise convex maximization, convex maximization as well as the maximum clique problem show the efficiency of the approach.
6. New Results

6.1. Massive mobile dense wireless networks

Participants: Cédric Adjih, Aline Carneiro Viana, Emmanuel Baccelli, Philippe Jacquet, Pascale Minet, Paul Mühlethaler, Yasser Toor.

6.1.1. Executive summary

Scaling properties of mobile ad hoc network lead to an increase of global capacity when the network density increases or when the packets can be stored for a while in mobile nodes instead of being immediately retransmitted.

Gupta and Kumar have shown in 2000 that the transport capacity per node in a multihop ad hoc network decreases in \( \frac{1}{\sqrt{N \log N}} \), \( N \) being the number of nodes in the network. Therefore the global capacity of the network increases in \( \frac{\sqrt{N}}{\sqrt{\log N}} \). This is a surprising result since in wired network a collection of nodes connected to a single communication resource has a transport capacity that just remains constant (i.e. the average per node capacity decreases in \( \frac{1}{N} \)).

Therefore adding space to a multihop wireless network increases the capacity: this is the space capacity paradox.

When nodes randomly move, it turns to be more advantageous to store packets for a while on mobile routers instead of forwarding them immediately like hot potatoes. When the mobile router moves closer to the destination, then it can delivers packets on a much smaller number of hops. Of course the delivery delay is much longer, but the network capacity also increases by slowing non urgent packets. This is the time capacity paradox: by slowing packets, nodes mobility increases network capacity. This was hinted the first time by Grossglauser and Tse in 2002.

The great challenge is to find the good protocol and tunings that allow to adjust the delivery delay from zero to infinity in order to get a continuous increase in capacity. The challenge is two-sided: one has to keep the delivery delay between reasonable bounds and one has to consider realistic mobility models.

Existing protocols for Mobile Ad Hoc Networks (MANET) are highly efficient in routing data between mobile nodes that belong to the same connected component (cf. the protocols which have received the RFC status by the manet group of IETF). What about a disconnected network where source and destination may be located in two different connected components? In this case usual routing protocols drop packet due to host unreachable as no end-to-end route exists at that time.

A simple idea is to allow the router that has no available route to the destination to keep the packet in buffer until the conditions become more appropriate for forwarding. The forwarding conditions will change because of mobility: the router can move closer to the destination so that they belong to the same connected component and the packet can be delivered.

Indeed, the network may be continuously partitionned due to high mobility, and the traditional approach to allow a mobile node to wait for the network to be fully connected (i.e. form a unique component) or to wait to be in range of the destination may lead to unacceptable delays. Furthermore, concrete applications, such as Defence and Disaster-Relief, cannot always rely on such assumptions.

Nevertheless, even if the communicating nodes may never be within the same connected component, it is important to observe that a “communication path” may be available through time using intermediate nodes that are temporarily within reach of each other while moving, hence making such networks viable for critical applications. Depending on the nature of the environment, these networks are now commonly referred as Intermittently Connected MANET and Delay Tolerant Networks.
In between stands the problem of the fully connected network that forms a single connected component, but for which maintaining full knowledge of the topology would simply make the network collapse under its huge control traffic. In fact this is the main problem that wireless network engineering has to face, in most experiments the generation of control traffic is the main source of disruption.

6.1.2. Scientific achievements

6.1.2.1. Scaling and spatial capacity in non uniform wireless networks

We found a more precise instance of Gupta- Kumar result by using a simple but realistic network model based on slotted ALOHA with Poisson traffic. It turns out that when the traffic density increases then the average node neighborhood area shrinks so that the average encircled traffic load remains constant with an analytical expression.

In their original model Gupta and Kumar assume that the traffic density is constant, which is far from realistic. However we have derived similar generalized results when the traffic density is not uniform. In this case, the heavier is the local traffic, the smaller are the local neighborhood and the larger is the number of hops needed to cross the congested region. Therefore the shortest paths (in hop number as computed by OLSR) will have a natural tendency to avoid congested area. The path tend to follow trajectory that have analogy in non linear optic with variable indices.

6.1.2.2. Time capacity and node mobility

We have defined a protocol that takes advantage of node mobility in a general way. In short the packet stay with its host router as long as the latter does not evade too fast from its next hop (computed via a shortest path protocol such as OLSR). In the way we understand “too fast” stands the tuning parameters we discussed above. There is no need to have node geographical location and to physically measure motion vector, since everything can be done via the analysis of the dynamic of neighborhood intersections. We analytically derived performance evaluation under random walk mobility models. We plan to simulate the protocol in a real mobility scenario. This algorithm has application in Intelligent Transport System.

6.1.2.3. Overhead reduction in large networks

The first limitation of multihop wireless network is the size of the overhead per node that increases linearly with the size of the network. This is a huge improvement compared to classic internet protocols which have quadratic overhead increases. Nevertheless this till limit the network size to some thousands. We have analyzed the performance of OLSR with Fisheye feature that significantly reduce the overhead with respect to distance. In theory the overhead reduction allows to network size of several order of magnitude. Anyhow the tuning of the overhead attenuation with distance must be carefully done when the network is mobile, in order to avoid tracking failure. We showed that an overhead reduction within square root of the network size achieve this goal. An alternative way to overhead reduction is ad hoc hierarchical routing and Distributed Hashing Table. Work has just begun in this area.

6.1.2.4. Coloring in wireless networks

Coloring is used in wireless networks to improve communication efficiency, mainly in terms of bandwidth, energy and possibly end-to-end delays. Nodes access the medium according to their color. It is the responsibility of the coloring algorithm to ensure that interfering nodes do not have the same color. First, we established complexity results about the h-hop coloring problem. Second, we focused on wireless sensor networks with grid topologies. We proposed the Vector-Based Coloring Method, denoted VCM, a new method that is able to provide an optimal periodic coloring for any radio transmission range and for any h-hop coloring, h>=1. Third, we also designed OSERENA "Optimized SchEduling RoutEr Node Activity", a distributed coloring algorithm optimized for dense wireless networks.

6.1.2.5. Complexity results about the h-hop coloring problem

In the paper we published at the WMNC 2011 conference, we define the h-hop node coloring problem, with h any positive integer, adapted to two types of applications in wireless networks. We specify both general mode for general applications and strategic mode for data gathering applications. We prove that the associated decision problem is NP-complete.
6.1.2.6. Grid coloring and the Vector-Based Coloring Method

In 2011, we also focused on wireless sensor networks with grid topologies. How does a coloring algorithm take advantage of the regularity of grid topology to provide an optimal periodic coloring, that is a coloring with the minimum number of colors? We propose the Vector-Based Coloring Method, denoted VCM, a new method that is able to provide an optimal periodic coloring for any radio transmission range and for any h-hop coloring, h>=1. This method consists in determining at which grid nodes a color can be reproduced without creating interferences between these nodes while minimizing the number of colors used. We compare the number of colors provided by VCM with the number of colors obtained by a distributed coloring algorithm with line and column priority assignments. We also provide bounds on the number of colors of optimal general colorings of the infinite grid, and show that periodic colorings (and thus VCM) are asymptotically optimal. Finally, we discuss the applicability of this method to a real wireless network.

6.1.2.7. Opportunistic routing

The model of wireless networks based on dynamic graph does not well assess the real processes in a wireless network. In particular the range of transmission can greatly vary between packets, the graph keeping only the average range. Opportunistic routing consists into taking advantage of temporary extension of the transmission range in order to gain several hops.

We have strong established theoretical performance limits in opportunistic routing. The limits are based on realistic interference scenarios in slotted Aloha. We have also investigated the impact of mobility on this theoretical limits.

We have designed an opportunistic routing protocol whose performance are within a small margin of the theoretical limits.

We have also conducted studies to support intelligent and adaptive forwarding, which allows a good trade-off between reliability and resource-efficiency. We then design a new protocol, called GrAnt, a new prediction-based forwarding protocol for complex and dynamic delay tolerant networks (DTNs). The proposed protocol uses the Ant Colony Optimization (ACO) metaheuristic with a greedy transition rule. This allows GrAnt to select the most promising forwarder nodes or allow for the exploitation of previously found good paths. The main motivation for using ACO is to take advantage of its population-based search and the rapid adaptation of its learning framework. Considering data from heuristic functions and pheromone concentration, the GrAnt protocol includes three modules: routing, scheduling, and buffer management.

6.1.2.8. Intermittent and delay tolerant networks

We consider the problem of routing in these networks, with the sole assumption that the speed of the node mobility is less than the speed of transmitting a packet to a neighbour. We compare this problem with sound propagation in liquid. We show that various pattern of mobility and network clustering can be described by a single parameter such as the information speed propagation.

We introduce new algorithms that route a packet toward a remote destination. The different algorithms vary depending on the buffering and the capacity capabilities of the network (i.e. if one or more copies of a packet can be sent and/or be kept). All algorithms are based on link aging rumors across connected components. The packet bounces from connected components to connected components, thanks to node mobility. We establish several analytical properties using an analogy with the sound propagation in liquid where molecules creates temporary connected components where sounds travel very fast.

Previous models assumed that the propagation of information path evolves like in a dynamic Erdos-Renyi graph leading to an epidemic flooding in O(log n) or O(1), n being the number of nodes in the network. We disprove the Erdos-Renyi model by showing via space-time considerations that the set of information path from a source to a destination is in fact much smaller than the path set in the Erdos-Renyi model. This lead to a much larger minimal delay in square root of n instead of log n. This correspond to a bounded maximal information propagation speed, whose estimate depends on the mobility model and the node density, and is root of multivariate explicit Bessel formulas.
Additionally, we have also considered the problem of data collection in global sensing and intermittently connected systems while avoiding the use of costly infrastructures (e.g., 3G). Motivated by the observation that node encounters are sufficient to build a connected relationship graph, we propose to take advantage of such inherent interactions to transform some mobile devices into delegates. We use then opportunistic delegation as a data traffic offload solution by investigating two main questions: (i) How to gain insights into social mobile networking scenarios?, (ii) How to utilize such insights to design solutions to alleviate overloaded 3G networks?. Our solution leverages usage of mobile applications requiring large data transfers by channeling the traffic to a few, socially important users in the network called VIP delegates. Mobile collectors need then only to meet delegates that, in turn, are responsible for gathering data from a subset of standard producers. We first investigate several delegation strategies based on the relative importance of nodes in their social interactions. Second, by considering a prediction strategy that estimates the likelihood of two nodes meeting each other, we investigate how the delegation strategies perform on predicted traces.

6.1.2.9. Network Coding

We study network coding for multi-hop wireless networks. We focus on the case of broadcasting where one source transmits information to all nodes in the network. Our goal is energy-efficient broadcast, that is, minimizing the total number of transmissions for broadcasting to the entire network. Note that this is a different problem for the classical problem of capacity maximization; and assuming we are far from the network capacity limit, hence in fact, we could assume interference-free transmissions.

Our previous results, they have shown that network coding (and a simple coding strategy) was able to reach optimality for asymptotically large and dense networks, with asymptotically 100% of the received transmissions being useful (innovative). We extended the results with the combined use of connected dominating sets and network coding: we were able to quantify (and bound) the benefits of network coding in networks where the area of the network stays fixed, and only the density increases.

We have proved that the performance of wireless random network coding are optimal in the following network model: the Erdos-Renyi random graph model and the unit disk random graph model. In particular we show in the Erdos-Renyi the network coding capacity rate outperform any Connected Dominating Set strategy by a factor of order \( \log n \). In the unit disk model we gain is larger than 60%. The result is based on the analysis of the connectivity stretch ratio of the random graphs. The connectivity stretch ratio is the ratio of the smallest degree over the connectivity number, and the connectivity stretch ratio tends to one in the two graph models.

6.1.3. Collaborations

- Professor Bernard Mans, Macquarrie University, Sydney, Australia,
- TREC INRIA team,
- Professors Anelise Munaretto and Myriam Regattieri Delgado from Federal Technological University of Parana (UTFPR), Brazil,
- CNRS researcher Marcelo Dias de Amorim, LIP6/UPMC, France,
- Mathias Boc, CEA LIST, France,
- Computer Science Department, Sapienza University of Rome, Italy,
- University of St. Andrews, UK,
- Professor Leila Saidane, ENSI, Tunisia.

6.2. New generation of OLSR, new services and protocols

Participants: Cédric Adjih, Aline Carneiro Viana, Emmanuel Baccelli, Thomas Clausen, Philippe Jacquet, Pascale Minet, Ichrak Amdouni, Ridha Soua, Erwan Livolant, Paul Mühllethaler, Yasser Toor.

6.2.1. Executive summary

The user of a mobile network very quickly experience problems with quality of service: links fade, connectivity disrupts, delays accumulate.
In a wireless network, the set of neighbors which with one node can communicate depends on transmission range, and numerous factors, and in addition the transmission range is often lower than the interference range (the range within which a node prevents correct transmissions of other nodes). Thus bandwidth reservation, a crucial step of quality of service, is an important and difficult problem.

The services and protocols that need careful adaptation are

- Connectivity continuity
- Bandwidth reservation
- Delay routing
- Connectivity control
- Autoconfiguration
- Security
- Energy efficiency
- Localization

The connectivity continuity is the most important problem. Trivial in the wired world where a link failure is a rare event, it becomes problematic in the mobile world where link failure caused by mobility are frequent and normal. The first experiments of mobile ad hoc networks with regular internet protocols miserably failed simply because either the protocol was to slow to recover link failure, or when tuned appropriately was generating such a huge overhead that the network collapsed under its own weight. A new generation of routing protocols has arised that allow a suitable control of connectivity in mobile networks. Among them the Optimized Link State Routing combines the optimization of overhead for mobile networks and the full internet legacy. It naturally provides path redundancy which accelerate link failure recovery.

The most important lesson that must be retained is that most of these optimization become NP complete, which is a significant complication compared to their counterpart in the classical wired world. The reason for the NP-completeness is two-sided: on one side the co-interferences make impossible an optimization link by link, on the other side, the large dispersion of performance measurement makes simple heuristic ineffective. As an example, routing with respect to shortest delay average does not guarantee smallest probability of high delay.

Since the bandwidth is scarce, any multimedia application such as video streaming is resource demanding. For example a TV broadcast that uses a mesh network will rapidly exhaust the bandwidth if all connections are point to point. In this case multicast protocols that allows to gather all these point to point connections in a single flow is a need.

There are two classes of multicast protocols: the tree based protocols and the network coding protocols. In the first class the protocols take advantage of the relatively small size of the recipient node set. One can show equivalent results of Gupta and Kumar scaling properties but in the multicast plan when the ratio of recipient versus network size is a fundamental parameter. When this ratio tends to one the performance naturally worsen.

When the recipient set is the whole network, one can apply the network coding scheme with random packet combination. In network coding the packets are no longer isolated: relay nodes makes linear combination of packets and transmitted mixed packets. In theory the performance of network coding is better than isolated packet multicast. In practice network coding is simpler to operate does not need topology management such as spanning trees or Connected Dominating Set. The reason for this is highly non intuitive, as if packet superposition was acting like state superposition in quantum mechanic, leading to non expected results.

Quality of service has become the central requirement that users expect from a network. High throughput, service continuity are critical issue for multimedia application over the wireless internet where the bandwidth is more scarce than in the wired world. A significant issue in the ad-hoc domain is that of the integrity of the network itself. Routing protocols allow, according to their specifications, any node to participate in the network - the assumption being that all nodes are behaving well and welcome. If that assumption fails - then the network may be subject to malicious nodes, and the integrity of the network fails. An important security service over mobile networks is to ensure that the integrity of the network is preserved even when attacks are launched against the integrity of the network.
6.2.2. Scientific achievements

6.2.2.1. Optimized Link State Routing (OLSR)

The routing protocol OLSR is universally known in the mobile wireless community (more than 475,000 hits on Google). It has numerous implementations and is used in many wireless networks. It is a proactive protocol with full internet legacy which is based on partial topology information exchange, that non the less provide optimal path with additive metrics (such as BGP/OSPF). It is an experimental RFC within IETF and soon will become a full standard under the name OLSRv2.

6.2.2.2. OSPF extension for wireless mesh networking

Long a near-future myth, ad hoc networks are now becoming a reality as a variety of wireless mesh networks are being deployed. Wireless mesh networks are a specific kind of ad hoc network, where terminals are essentially fixed. Even in such cases, which somewhat resembles usual networks, specific routing protocols have nevertheless to be employed, to cope with the characteristics of wireless, multi-hop communications. Such characteristics include scarce bandwidth over inherently unreliable, versatile, semi-broadcast links, and absence of a central authority in general. One of the main difficulties in this context is to cope with contradictory requirements such as, on one hand, dealing with bandwidth scarcity, which typically requires decreasing control traffic, while on the other hand, dealing with unreliable, versatile links which typically requires increasing control traffic. The two prominent routing protocols that have been developed for ad hoc networks and studied over the past decade, are the IETF standards AODV and OLSR. AODV is based on a reactive scheme (i.e. on-demand flooding to discover a path to a new destination), while OLSR is based on a proactive scheme, which is essentially an optimization of link state routing (i.e. pre-provisioning of paths to all possible destinations). OLSR is to date the most deployed such protocol, as it powers numerous wireless mesh community networks that currently flourish in various cities throughout Europe and North America. Based on this experience, the integration of ad hoc networking in the "standard" networking body is going further in several directions. One direction is the IEEE 802.11s standardization effort, which uses AODV and OLSR-derived algorithms to provide wireless mesh routing capabilities below IP. Another direction, spearheaded by the IETF, is the extension of IP routing standards such as OSPF to support ad hoc routing: in this realm we recently spun RFC 5449, as well as a series of academic publications on the subject. The idea behind extending OSPF to support ad hoc networks comes from a simple observation: OSPF is algorithmically quite similar to OLSR, as both are based on a proactive, link state approach. As on the other hand OSPF is a well-understood, widely deployed, industry-standard protocol, employing it to integrate ad hoc networks with existing infrastructure is considered by users as an easy migration path.

6.2.2.3. Multi-metric routing

Quality of service involves finding routes between two nodes in the network that satisfies a number of constraints. These constraints could be the requested bandwidth, the maximum delay, the minimum loss probability, the reliability of links, etc. This problem is NP-Complete because it combines additive metrics in the optimization problem. Hipercom proposed heuristics for finding routes that respect up to four metrics when calculating routes between source and destination. Another QoS issue is the creation of models that estimate the actual value of a metric. For example, computing the available bandwidth or the transfer delay on a link, etc. is very complex in a non-deterministic medium access such as WiFi. To resolve this problem, we developed a model for estimating the available bandwidth in a wireless network. This model is based on considering interfering links in cliques, after which we provide the maximum capacity that could be deployed in a clique. We may still enhance the model by adding a scaling factor to the clique equations in order to become more accurate when compared to real measurements.

In particular we have investigated the metric based on packet delay distribution. Since propagation delays between routers are negligible, most delays occur in queueing and medium access control processing. Contrary to previous common belief there is no need of network synchronization. The objective is to proactively determine the delay in absence of packet data traffic. The estimate of delay distribution is done via analytical method. In order to keep control on quality of service flows we use source routing forwarding options.
6.2.3. Collaborations

- Many contractual collaborations:
  - MoD (QoS, security, interconnection between the OLSR and OSPF routing domains),
  - Hitachi (Vehicular applications, OLSRv2),
- Non contractual:
  - BAE (OLSRv2),
  - Deutsche Telekom Labs/TU-Berlin, Germany,

6.3. Wireless Sensor Networks

Participants: Cédric Adjih, Aline Carneiro Viana, Emmanuel Baccelli, Thomas Clausen, Philippe Jacquet, Pascale Minet, Ichrak Amdouni, Ridha Soua, Erwan Livolant, Paul Mühlethaler, Yasser Toor.

6.3.1. Executive summary

In wireless sensor networks, we focus more particularly on:

- Spatial reuse of the bandwidth,
- Routing according to a cross-layering approach,
- Security,
- Energy efficiency,
- IPv6 support.

6.3.2. Scientific achievements

6.3.2.1. Cryptographic Protocols to Fight Sinkhole Attacks on Tree-based Routing in Wireless Sensor Networks

Wireless Sensor Networks (WSN) are penetrating more and more in our daily life. As a consequence, security has become an important matter for these networks. We introduce two new cryptographic protocols of different complexity and strength in limiting network degradation caused by sinkhole attacks on tree-based routing topologies in Wireless Sensor Networks (WSNs). The main goal of both protocols is to provide continuous operation by improving resilience against, rather than detection of, these attacks. The main benefit of providing resilience is that it allows operating (or graceful degradation) in the presence of attacks. Furthermore, while resilience mechanisms do not dismiss detection mechanisms, detection mechanisms often introduce more complexity and so, more weaknesses to the system, which might not justify their benefits. More specifically our two RESIlient and Simple Topology-based reconfiguration protocols are: RESIST-1 and RESIST-0. RESIST-1 prevents a malicious node from modifying its advertised distance to the sink by more than one hop, while RESIST-0 does not allow such lying at the cost of additional complexity.

6.3.2.2. IPv6 Protocol suite for Sensor Networks

Wireless sensor networking is a key element of the Internet of Things (IoT), a substantial part of the billions of smart objects that are soon to blend into the global IP network, from actuators to home appliances, from smart meters, to smart dust. Sensor nodes are devices used for distributed and automated monitoring of various parameters such as temperature, movement, noise or radioactivity levels etc. Sensors are scattered with minimum planning with respect to their precise physical position (including the central role of the sink, if any), and the set of peers with which a sensor can directly communicate through its wireless interface may change rapidly over time due to asynchronous sleep mode strategies, fluctuations in the radio environment, device failure or mobility. Through its wireless interface, a sensor thus connects to a communication link with undetermined connectivity properties. Sensor networks are a challenge to current IP standards, since on the one hand these protocols were designed to work on wired links and on the other hand these protocols were designed to work on machines that do not have drastic constraints in terms of CPU, power capacities, and memory, as sensor nodes do. In consequence, several key standard protocols (including TCP, UDP, DHCP,
NDP, SLAAC, and OSPF) do not function correctly in this environment. Nevertheless, IPv6-based sensor networking is a viable long term goal because it would enable generic, large scale, seamless integration of millions of sensing devices using heterogeneous radio technologies, at a low cost, and in a future-proof manner. The Internet Engineering Task Force (IETF) is currently engaged into multiple efforts addressing the limitations of existing standards concerning wireless sensor IP networking. Some of the standards under construction aim at fitting IP formats, especially IPv6 formats, to direct wireless communications using low power radio technologies such as IEEE 802.15.4, which require IP format compression. Other standards in development aim at providing multi-hop wireless sensor communication with IPv6, which requires specific routing protocols, efforts in which we actively participate, prompting numerous joint publications with both industrial and academic partners.

6.3.2.3. Coloring in wireless sensor networks

Graph coloring is used in wireless networks to optimize network resources: bandwidth and energy. We focus on grid topologies that constitute regular topologies for large or dense wireless networks. We consider various transmission ranges and identify a color pattern that can be reproduced to color the whole grid with the optimal number of colors. We obtain an optimal periodic coloring of the grid for the considered transmission range. We then evaluate the performance of a 3-hop distributed coloring algorithm, called SERENA. Through simulation results, we highlight the impact of node priority assignment on the number of colors obtained for any network and grids in particular. We then compare these optimal results on grids with those obtained by SERENA and identify directions to improve SERENA.

6.3.2.4. Coloring algorithm optimized for dense wireless networks

In 2011, we also designed OSERENA “Optimized SChEduling RoutEr Node Activity”, a distributed coloring algorithm optimized for dense wireless networks. Network density has an extremely reduced impact on the size of the messages exchanged to color the network. Furthermore, the number of colors used to color the network is not impacted by this optimization. We describe the properties of the algorithm and prove its correctness and termination. Simulation results point out the considerable gains in bandwidth.

6.3.2.5. Multichannel access in wireless sensor networks

In 2011 we started a research activity on multichannel access in wireless sensor networks. A state of the art has been published at the IFIP Wireless Days 2011 Conference.

6.3.3. Collaborations

- Many contractual collaborations:
  - Hitachi (Vehicular applications, OLSRv2),
  - OCARI2 project (industrial wireless sensor network, QoS, cross layer, energy efficiency, routing, node activity scheduling),
  - SAHARA project (wireless sensor network embedded in aircrafts),
  - STIC INRIA-Tunisian Universities: the team of Prof. Leila Saidane at ENSI (Performance improvement in a wireless sensor network),
- Non contractual:
  - BAE (OLSRv2),
  - Freie Universitaet (sensor networks, DHT),
  - Deutsche Telekom Labs/TU-Berlin, Germany,
  - University of Athens, Greece.

6.4. Vehicular and mobile applications

Participants: Cédric Adjih, Emmanuel Baccelli, Thomas Clausen, Philippe Jacquet, Pascale Minet, Paul Mühlethaler, Yasser Toor.
6.4.1. Executive summary

We have the following vision: in the future mobile internet and static internet will have their core deeply intricately. This means that mobile ad hoc networks will be attached to the core network, form extension and even be part of it. For example in disaster area, a wireless network could replace the destroyed infrastructure and help to the emergency operations.

With this perspective items such as Autoconfiguration, Security are of crucial importance. However there is a potential conflict between a large population of fixed nodes based on ancient protocol and a smaller but more dynamic population based on new protocols. In the integration both population must cooperate in an hybrid protocol.

The difficulty is to build protocols that are as dynamic and efficient as MANET protocols but can support the legacy of the old and heavy internet protocols. The challenge is nevertheless achievable, because the dynamic part of the network needs less frequent updates from the fixed part of the network. Moreover the natural abundance of resource in the fixed part of the network allows it to support the more frequent updates from the mobile part.

OLSR has been found to be the natural best candidate for this challenge since it gathers dynamic and optimization with internet legacy.

6.4.2. Scientific achievements

6.4.2.1. Military tactical networks

During year 2011, we conducted several expertises about industrial proposals dealing with OLSR use in military tactical networks.

6.4.2.2. Protocols for vehicular networks

We have achieved numerous studies and design of protocols for vehicular networks and more specifically for V2V (Vehicle-to-Vehicle) network.

First we have studied the channel occupancy induced by the OLSR proactive routing protocol used in a linear Vehicular Ad hoc Network (VANET). Unlike previous studies, which usually use simulations to evaluate the overhead, we have proposed a simple analytical model to carry out this evaluation. Moreover, we did not evaluate the total overhead induced by the routing protocol as is usually proposed, but, for a given node, the channel occupation induced by the routing protocol.

We have studied flooding techniques for safety applications in VANETs. The typical scenario is the diffusion of an alert message after a car crash in a platoon of vehicles. The packet is diffused with the pure flooding, the multipoint relay (MPR) diffusion of OLSR and a geographic aware protocol. For OLSR we have introduced a variant (Robust-MPR) to improve the reliability. Different realistic scenarios were considered and various parameters such as vehicle density, and background traffic load were scrutinized. We have shown that the Robust-MPR and the geographic aware protocol satisfy the requirements of the safety applications while using considerably less overhead than pure flooding.

We have shown that the geographic aware protocols can be improved for the diffusion of an alert message by using opportunistic routing. We have designed OB-VAN (Opportunistic Broadcast for VANets) a new protocol that uses this idea. One of the novelty of this protocol is the use of an active signalling technique in the acknowledgement procedure to select the best relay taking advantage of the reception pattern of each message. We have studied OB-VAN in a linear VANET and have shown that it outperforms the flooding for the delay and the amount of overhead. However the delivery ratio of OB-VAN may be insufficient for safety applications. This remark has led to the design of R-OB-VAN which is a reliable variant of OB-VAN. With extensive simulations, we have shown that R-OB-VAN maintains a high delivery ratio even in the presence of packet loss due to shadowing.
We have studied the performance of the Aloha scheme in linear VANETs. This analysis assumes a SINR (Signal over Interference plus Noise Ratio) based model. In this model, we have derived the probability of a successful transmission between two vehicles at a distance of $R$ meters. We have also computed the mean throughput according to Shannon’s law. In these two models, we have optimized the two quantities directly linked to the achievable network throughput i.e., the mean packet progress and the density of transport.

Finally, we have studied the utilization of opportunistic routing and shown that this technique is also beneficial for point to point traffic. It decreases the delay and increases the throughput compared with shortest path first routing. Moreover, we have also shown that opportunistic routing for point to point traffic eases considerably the optimization of the MAC scheme e.g. the transmission probability for Aloha and the carrier sense threshold for CSMA.

### 6.4.3. Collaboration

We received support from MoD for this activity.
6. New Results

6.1. Behavioral Fingerprinting

Participant: Olivier Festor [contact].

Device fingerprinting aims to automatically determine the types (name and version of software, brand name and series of hardware) of remote devices for a given protocol. Hence, keeping an up-to-date inventory database of devices in use on a network is possible and helpful as for example to check remotely if unauthorized applications have been installed. Some types of devices for which vulnerabilities are known can be easily detected in order to patch them or at least send alerts to the owners. From a security point of view, attackers use specific tools to perform their attack which may also be detected rapidly thanks to fingerprinting. Most current systems rely only on signatures of differences in implementation of a given protocol stack and signatures are often outdated.

We have designed a new fingerprinting scheme that is accurate even on protocol stacks that are completely identical, but which run on hardware having different capabilities (CPU power, memory resources, etc). Our fingerprinting scheme can learn distinctive patterns in the state machine of a particular implementation. We see such a pattern as a restricted tree finite state machine that provides additional time-related information about the transitions performed [15]. The captured identification models were then used to automatically build attack prevention rules [19].

This work was done in cooperation with Jérôme Francois, Radu State and Thomas Engel from the University of Luxembourg.

6.2. Management and monitoring of P2P networks

Participants: Isabelle Chrisment [contact], Olivier Festor, Juan Pablo Timpanaro.

Content pollution is one of the major issues affecting P2P file sharing networks. However, since early studies on FastTrack and Overnet, no recent investigation has reported its impact on current P2P networks. In [21], we presented a method and the supporting architecture to quantify the pollution of contents in the KAD network. We first collected information on many popular files shared in this network. Then, we proposed a new way to detect content pollution by analyzing all filenames linked to a content with a metric based on the Tversky index and which gives very low error rates. By analyzing a large number of popular files, we showed that 2/3 of the contents are polluted, one part by index poisoning but the majority by a new, more dangerous, form of pollution that we call index falsification. This work was done, in collaboration with the University of Technology of Troyes, within the context of the ACDA-P2P3 Project funded by GIS-3SGS4.

BitTorrent is a widely deployed P2P file sharing protocol, extensively used to distribute digital content and software updates, among others. Recent actions against torrent and tracker repositories have fostered the move towards a fully distributed solution based on a distributed hash table to support both torrent search and tracker implementation. We conducted an analysis on one of the BitTorrent’s DHT (Mainline DHT) and developed a monitoring architecture, so as to measure and discover security flaws on the network. In [23] we compared KAD DHT against BitTorrent DHT in terms of security by deploying different attacks on the network. We showed that the lack of security in Mainline DHT allows very efficient attacks that can easily impact the operation of the whole network. We also provided a peer-ID distribution analysis of the network, so as to adapt previous protection schemes to the Mainline DHT. The mechanisms are assessed through large-scale experiments on the real DHT-based BitTorrent tracker.

---

3 Approche Collaborative pour la Détection d’Atttaques dans les réseaux Pair à Pair
4 Groupement d’Intérêt Scientifique - Surveillance, Sureté et Sécurité des grands Systèmes
If BitTorrent’s Mainline DHT is exposed to several identified security issues, in parallel, the KAD DHT has been the core of intense research and was improved over years. We presented a study that motivates the integration of both worlds. We provided a performance comparison of both DHTs in terms of publishing efficiency. We investigated the security threats and showed that the current BitTorrent’s Mainline DHT is more vulnerable to attacks than KAD while the download service of BitTorrent has much better performance. Given the strengths and weaknesses of both DHTs, we designed a hybrid architecture [24], which is based on KAD’s indexation mechanism and BitTorrent download protocol. On the one hand, the client is able to index its files in the well-known KAD DHT, taking advantage of KAD’s security mechanism and its double-indexation scheme. On the other hand, the client uses the BitTorrent download protocol so as to download a given file, which has been proven to surpass KAD’s. We implemented this hybrid architecture, that we called hMule, as a unified KAD-BitTorrent file-sharing application, which is compatible with both P2P file sharing networks and provides the KAD advantages on indexation and the BitTorrent speed for transfer without losing backward compatibility.

We started our research about being anonymous when downloading from BitTorrent. We conducted a set of measurements from High Security Lab aiming to characterize the usage of the I2P network, a low-latency anonymous network based on garlic routing [35]. Our goal was to answer the following questions: what is the network used for? when is it used the most? which kind of applications the network designers should pay more attention to? We designed a distributed monitoring architecture for the I2P network and we showed that, through three one-week long experiments, we were able to identify 32% of all running applications, among web servers and file-sharing clients. Additionally, we identified 37% of published I2P applications, which turned out to be unreachable after their publication on the I2P distributed database.

In parallel, we built-up a model of I2P encryption/decryption approach and using the Avispa tool, we able to find a possible attack on the network. Further work will be focused on probing right and on developing a proof-of-concept of this.

### 6.3. Configuration security automation

**Participants:** Rémi Badonnel [contact], Martin Barrere, Olivier Festor.

The main research challenge addressed in this work has focused on enabling configuration security automation in autonomic networks and services. In particular our objective has been to increase vulnerability awareness in the autonomic management plane in order to prevent configuration vulnerabilities. The continuous growth of networking significantly increases the complexity of management. It requires autonomic networks and services, which are capable of taking in charge their own management by optimizing their parameters, adapting their configurations and ensuring their protection against security attacks. However, the operations and changes they execute during these management activities may generate vulnerable configurations. A first part of our work has therefore consisted in consolidating a security automation strategy for preventing vulnerabilities and maintaining safe configurations in autonomic infrastructures [7]. This solution relies on the integration of configuration vulnerability descriptions into the management plane [8]. The OVAL language, part of the SCAP protocol, has become the de-facto standard for specifying configuration vulnerabilities in a technical viewpoint. We have refined a mathematical modeling for mapping OVAL descriptions into policy rules which can be interpreted by the autonomic Cfengine configuration system. These policies enable the Cfengine system to assess and detect vulnerabilities. We have designed a functional architecture and formalized a translation algorithm for supporting this security automation. We have also prototyped an OVAL-to-Cfengine translation module, called Ovalyzer, and analyzed its interactions with the components of the Cfengine system. Based on vulnerability descriptions extracted from the official OVAL repository, we have performed an extensive set of experiments to quantify the performance and coverage of the Ovalyzer module. A second part of our work has consisted in investigating how our security automation solution can be extended to distributed configuration vulnerabilities. In SCAP-based traditional approaches, a distributed vulnerability is typically understood as the aggregation of individual configuration vulnerabilities which are spread in the network and might allow a multi-step attack. We have shown through the analysis of a case study that this definition does not offer a complete outlook of the problem. In particular, each network device can individually present a secure
configuration, but when combined across the network, a global vulnerable configuration may be produced. In that context, we have introduced in [27] a mathematical definition for distributed vulnerabilities and have specified the DOVAL language (Distributed OVAL), on top of OVAL, as a means for describing these vulnerabilities in a machine readable manner. A case study in the area of VoIP networks and services has been considered for demonstrating the instantiation of DOVAL main constructs. The DOVAL descriptions constitute useful security definitions that in turn can be exploited for security automation. We have built a framework for supporting these distributed configuration vulnerabilities based on the Cfengine system. In particular, we have proposed and evaluated collaborative strategies and optimized algorithms for performing the assessment of DOVAL descriptions.

6.4. Online Risk Management

Participants: Rémi Badonnel [contact], Oussema Dabbebi, Olivier Festor.

Telephony over IP has known a large scale deployment and has been supported by the standardization of dedicated signaling protocols. This service is however exposed to multiple attacks due to a lower confinement in comparison to traditional PSTN networks. While a large variety of methods and techniques has been proposed for protecting VoIP networks, their activation may seriously impact on the quality of such a critical service. Risk management provides new opportunities for addressing this challenge. In particular, our work aims at performing online risk management for VoIP networks and services. The purpose is to adapt the service exposure with respect to the threat potentiality, while maintaining a low security overhead. Based on the classification of VoIP attacks and the analysis of their properties, we have refined in [11] an extended risk modeling for IP telephony infrastructures. This modeling permits to cover a large spectrum of security attacks. It supports our online risk management strategy which is capable of dynamically activating or deactivating security safeguards in the VoIP infrastructure. The mitigation is based on the control of the service exposure using these safeguards. We have compared our solution to other traditional strategies, and have quantified the benefits and limits according to multiple performance criteria. We have also analyzed the impact of the risk model parameters on our mitigation, and showed to what extent the parameterization can be partially automated in [12]. An important part of our efforts has focused in the year 2011 on extending our online risk management strategy to more distributed configurations [32]. While our initial work was centered around Asterisk-based enterprise networks, we have taken a particular interest in P2PSIP networks. They constitute an open decentralized solution where the registration and location servers are implemented by a distributed hash table responsible for storing the bindings between the address-of-record SIP-URI and the contact SIP-URI. We have identified different attack sources and attack scenarios in these P2PSIP networks, considering the functional roles that are played by the SIP peers. The security threats are specific to the P2PSIP protocol or are the result of inheritance from the SIP layer and the peer-to-peer area. In that context, we have analyzed the instantiation of our online risk modeling by taking into account the properties and components of the P2PSIP architecture, and have established a portfolio of dedicated countermeasures, including replication-based an certification-based techniques. We have evaluated the strategy performance and scalability through an extensive set of experiments performed with the OMNET++ simulator. We also have quantified the complementarity of our solution with the RELOAD security framework which relies on a central certificate enrolment server.

6.5. VoIP Security

Participants: Laurent Andrey, Olivier Festor, Abdelkader Lahmadi [contact].

In previous work, we have proposed the prevention system SecSIP [5] for SIP-based networks which uses a rule-based approach to build prevention specifications on SIP protocol activities that stop attacks exploiting an existing vulnerability before reaching their targets. We have pursued our efforts in VoIP security which led to two new contributions:

- Building and maintaining prevention rules using the VeTo language can become a time consuming and error prone task, especially when addressing an important number of vulnerabilities discovered
using a fuzzing tool. The discovered vulnerabilities using such process are usually based on a single exploit message with a malformed field or sequence of vulnerable messages. To reduce this effort, we have designed a generation method to produce VeTo specifications targeting those vulnerable messages. The method mainly characterizes a malformed field within an exploit message or the vulnerable sequence of messages and generates a set of VeTo rules specifications to prevent their exploit. The generated VeTo rules are then deployed and maintained on the SecSIP engine to be applied against the SIP traffic. The solution [19] relies on generating rules using genetic algorithms operating on a set of candidate regular expressions to match a malformed pattern within a SIP message, and evaluate their quality using a well defined fitness function to ensure that their are specific enough to only match exploit messages.

- SecSIP uses a plain text configuration file in which VeTo specifications are authored and managed manually. While extending the deployment of the framework beyond our own lab, support for remote configuration was required. Given the promise of Netconf, we naturally turned our investigations towards this protocol and embraced the YANG data-modeling framework. In [20] we have presented the Yang model built for VeTo policies and the Netconf framework put in place.

We have developed a flexible SIP honeypot. It is flexible in the sense that a behavior can be externally and easily defined. The goal of such a honeypot is to be able to be quickly customized in response to an observation made on a more generic and large scale honeypot. If the initial observation is likely to be an attack the customized honeypot would eventually get deeper and more informative interactions with the attacker. The realization is a module of the Dionaea general framework for honeypot (successor of the well-known nepenthes framework) and we use the SIPP test tool as an engine to animate SIP interactions provided as automata in some XML file. More detail on the implementation can be found in [26].

6.6. VoIP Fraud

Participants: Olivier Festor [Contact], Mohamed Nassar.

In the context of a cooperation with the University of Liege, we have addressed the problem of SPIT from a new perspective [22]. Based on end-user feedback, we have proposed a scheme for generating SPIT signatures from the SIP INVITE messages. Hence it is possible to filter the next SPIT calls before ringing their destinations. The generated SPIT signatures are adaptive to the benign signaling traffic in the sense that they do not conflict with it. The generation of signatures is based on supervised machine learning techniques. We namely investigated decision trees with categorical attributes obtained by parsing the SIP messages. Our system works in two modes: a batch and an online mode. The batch mode consists on training the decision tree over a labeled (spit, normal) data-set and then transforming the tree into an if-else rule-set. In online mode, the successive learnt signatures are aggregated and the possible conflicts are resolved. Experimentation on off-the-shelf SPIT tools showed the efficiency of our approach to find the good signatures. However, experiments show that the J48 decision tree is easily defeated using some obfuscation techniques. We therefore proposed a generalisation approach to translate the tree into an if-else rule-set shows instead good robustness against such attacks. The overall framework provides suitable performance for operational deployment in terms of learning time, required memory, size of the rule-set and the call setup delay. The different parameters of the system (i.e. size of the different buffers and windows) are easily configurable. Different SPIT signatures may imply different SPIT capabilities. For example, a spitter may break a Captcha test by brute-forcing a DTMF guess. Another spitter may start talking by a human-like congratulation in order to bypass a Turing test. One of the goals of our approach is to provide a framework for applying reinforcement learning techniques and hence increasing the efficiency of the filtering process. The reinforcement learning aims at selecting the best challenge to be used when a given SPIT signature is detected. Basically the reinforcement learning maintains a table matching each signature with the best challenge response discovered so far. The table is continuously updated using a trial and error scheme.

We did validate the approach on multiple data-sets obtained from Voice over IP operators members of the SCAMSTOP project.
6.7. Pervasive computing

Participants: Laurent Ciarletta [contact], Tom Leclerc, Julien Siebert, Olivier Festor, André Schaff.

Vincent Chevrier (MAIA Team)

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. Madynes is focusing on the networking aspects of ubiquitous systems. We cooperate with the Maia (and Trio) team(s) to be able to encompass issues and research questions that combine both networking and cognitive aspects.

Pervasive Computing is about interconnected and situated computing resources providing us(ers) with contextual services. 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 as detailed in [10] is going towards both closing the loop between humans and systems and taming the complexity, using multi-modeling (to combine the best of each domain specific model) and co-simulation (to design, develop and evaluate) as part of a global conceptual and practical toolbox.

In 2011 we worked on the following research topics:

• Multi-models of these Pervasive Computing environments (including the users in the modeling and the simulations). We have been focusing on the collaborative simulations of dynamic networks/elements, namely P2P and adhoc networks using agents to drive those simulations. This work is done in collaboration with the MAIA team. The results have been extensively described in the PhD thesis of Julien Siebert [3].
• Study of service discovery protocols, contextual metrics in adhoc networks, and Service Discovery in adhoc networks using an hybrid model between cluster-like (WCPD) and MPR-based (OLSR) broadcasting. The results have been extensively described in the PhD thesis (Contributions for Advanced Service Discovery in Ad hoc Networks) of Tom Leclerc [2]. In this thesis, we consider service discovery in MANETs, that are a collection of devices that communicate with each other over a wireless medium. Such networks are formed spontaneously whenever devices are in transmission range without any preexisting infrastructure. The main characteristic of MANETs is the high dynamics of nodes (induced by the users moving around), the volatile wireless transmissions, the user behavior, the services and their usage. We’ve proposed a complete solution for service discovery in ad hoc networks, from the underlying network up to the service discovery itself. A first contribution, is the Stable Linked Structure Flooding (SLSF) protocol that creates stable based cluster structure and thereby provides scalable and efficient message dissemination. The second contribution is the Stable Linked Structure Routing (SLSR) protocol that uses the SLSF dissemination structure to enable routing capabilities. Using those protocols as basis, we propose to improve service discovery by additionally considering context awareness and adaptation.
• Context awareness and mobility/usage models
  We contributed on improving simulations by coupling simulators and models that, together, can model and simulate the variety and richness of ad hoc related usage scenarios and their human characteristic. A guideline for all of our contributions was to be able to integrate and/or consider context and context awareness in both the proposed protocols and the related research tools and models. On one hand, The proposed protocols all have the capacity to adapt their efforts according to certain metrics, that represent the context. The simulator coupling architecture, on the other hand, permits to model and design scenarios in which the context, such as the service usages or the human behavior, has an impact and matters.
• Energy-constraint geolocalization, addressing, routing and management of wireless devices: a research collaboration with Fireflies RTLS was started in March 2009 and is ongoing. The initial work
has been extended in a joint work with the TRIO Team and leads towards finding a global energy-cost function, and life expectancy of the wireless sensor system.

In the future work, we plan to apply those results to Cyber Physical Systems, within the Aetournos (Airborne Embedded auTonomOUs Robust Network of Objects and Sensors) platform at Loria. We aim at developing cross-layer solutions to robust routing between flying drones.

We are also working inside a CPER project towards management solutions of wireless network sensors (project ECOSUR) used to control Smart Spaces.

6.8. Co-Simulation and multi-modeling

Participants: Laurent Ciarletta [contact], Julien Siebert, Tom Leclerc.

Vincent Chevrier (MAIA team, LORIA) and Tomas Navarette are external collaborators.


Participants: Laurent Ciarletta [contact], Julien Siebert.

Vincent Chevrier (MAIA team, LORIA) is an external collaborator.

this work has been extensively detailed in Julien Siebert’s PhD thesis [3] and partially in Tom Leclerc’s, with an application to ubiquitous adhoc networks and services.

This work has been done between the fields of ubiquitous networks and multi-agent based simulation. The main context is to study mutual influences existing between ubiquitous network performances and their users behaviours. We have highlighted the need for reusing and coupling modelling and simulation softwares together in order to simultaneously integrate several abstraction levels in the study. We target those needs by a multiagent approach and we propose a metamodel: AA4MM. The core idea in AA4MM is to build a society of models, simulators and simulation softwares that solves the core challenges of multimodelling and simulation coupling in an homogeneous perspective. AA4MM major contributions are the possibility to easily reuse, to make interoperable and modular existing heterogeneous models and softwares, to manage scale changes and a simulation algorithm fully decentralized. We apply this metamodel to the field of ubiquitous networks in order to target the question of mutual influences between networks performances and users behaviours.

6.8.2. Adaptive control of a complex system based on its multi-agent model

Participants: Laurent Ciarletta [contact], Julien Siebert.

Vincent Chevrier (MAIA team, LORIA) and Tomas Navarette are external collaborators and main investigators of this theme.

As a starting point, we are exploring how the behavior and other factors such as spatial and temporal dimensions are mutually influencing and the impact of parameters variability of our models in environment where collective behaviors can emerge [6]. We did comparison of five different models. These models are built upon the same individual behavior hypothesis of a collective phenomenon present in peer-to-peer file exchange networks: "free-riding". We studied a global analytical model and four multi-agent models. Multi-agent models include the space and time dimensions rarely seen in the literature discussing aggregated models of the collective phenomenon in question. We have demonstrated that one individual decision algorithm can lead to contradictory information.

Using these results, we want to build a control mechanism for a complex/dynamic system. Specifically, we want to evaluate the effectiveness of creating a control mechanism based on a multi-agent model of the system.

Multi-agent models can be adapted to that purpose since usual approaches using analytical models as a basis can be intractable when dealing with such systems; and if we consider that the available control actions are meant to be applied locally, a multi-agent model is necessary. We are currently working on a case study within the dynamic networks domain, namely the free-riding phenomenon present in peer-to-peer networks.
We propose an architecture that gathers information from the system and uses it to parametrize and tune a set of multi-agent models. The outcome of simulations is used to decide which control actions have to be applied to the system, in order to achieve a predefined control objective. We consider that we do not have complete information to characterize the state of the system.

6.9. Sensor networks management

Participants: Cyril Auburtin, Alexandre Boeglin, Olivier Festor, Abdelkader Lahmadi, Emmanuel Nataf [contact].

6LowPAN networks denotes many embedded devices interconnected by a variety of links ranging from wireless technologies such as 802.15.4, bluetooth, Low Power Wifi to wired technologies such as low power PLC. The common property of such networks is the limited resources of their nodes in terms of power, computing, memory and communication. The network could be described with thousands of devices with very limited internal and external resources and their communication channels are low-bandwidth, high loss rate and volatile links subject to failure over time. These networks rely on the 6LowPAN protocol defined by the IETF as an adaptation layer for the IPv6 protocol to address their low power and lossy properties.

During the year 2011, we have started a research activity around the monitoring and security assessment of 6LowPAN networks. Our contributions are mainly as follows:

- We are developing a novel approach to assign monitoring roles in 6LowPAN networks using available local information provided by the routing layer. The resulting monitoring architecture is adaptive taking benefit from the reactivity of the routing protocol when dynamic changes occur due to connectivity or nodes mobility. Our first simulations results reveal that our assignment approach is more efficient, less aggressive and less resources consuming than its competitors.

- We have also designed and implemented a piggybacking technique to deliver monitoring report into existing packets traveling through 6LowPAN networks. In our solution, we have extended the IPv6 Hop-by-Hop extension header with a new option which contains status data of monitored nodes. This technique can reduce the number of packets and bytes sent across the network since there is no specific monitoring packets competing with existing traffic. Monitoring data shares the routing path of application data packets until it reaches a management node. We have applied our piggybacking technique to discover coap-enabled management agents. Each agent in the deployed wireless sensor network piggybacks its identifier into the RPL routing protocol messages until it reaches a manager node.

- Regarding security management of these networks, we have developed a stateless fuzzing tool for the 6LowPAN protocol [28]. The tool is build upon the Scapy packets manipulation library. It provides different mutation algorithms to be applied on 6LowPAN messages. These messages are defined by interaction scenarios described in an XML format.

- Related also to security, we have modelled an ontology for intrusion detection system in sensor networks [17]. The model exposes family of intrusions depending on their objectives. The service provided by the network, the communication channels and the security mechanisms are the main classes of the model.

6.10. High Security Lab

Participants: Alexandre Boeglin [contact], Olivier Festor, Mohamed Nassar.

The objective of the High Security Lab at INRIA Nancy Grant Est is to provide both the infrastructure and the legal envelope to researchers to perform sensitive security oriented experimentations. We do contribute to this laboratory by (1) designing and operating a large network telescope and (2) performing vulnerability assessment research, network data and malware collection and analysis.
During the year 2011, some maintenance tasks have been carried out on the High Security lab:

- the SDSL line, which previously had a capacity of 1Mbps, has been upgraded to a 2Mbps line, and traffic shaping rules have been added to the router, that allow honeypots to run alongside experiments, without impacting them,
- the storage capacity of our database server, which was starting to get full, has been multiplied by four, and existing data has been migrated to the new equipment.

A set of new experiments have also been deployed:

- a server has been dedicated to a new variant of SGNet, for the VAMPIRE project. This one specifically targets attacks on SIP services, which the other one cannot do,
- in collaboration with the INRIA Nancy Grant Est IT service, we started to log public (thus anonymous) DNS queries and responses made by the research center’s recursive DNS servers, to use the collected data as input set for experiments.

In 2011 we worked also on the automated analysis of malware traces to extract flow-level signatures of malware. We obtained early results regarding network flow-graphs and tested several clustering techniques to separate malware traffic.

6.11. Sensas

**Participants:** Cyril Auburtin, Alexandre Boeglin [contact], Olivier Festor.

The goal of the SensAS ADT, which started in 2011, is to propose applications based on wireless sensor networks, building upon work that has been done through the SensLab and SensTools projects.

The Madynes team is responsible of the SensMGT part of the SensAS project, which focuses on sensor network management and configuration applications.

First, we adapted the existing contiki-snmp implementation to the SensLab WSN430 nodes. We did so by (1) reducing the memory footprint of the code and (2) by implementing several SNMP MIBs.

To reduce the memory footprint, we had to disable some optional features and unused drivers of the Contiki OS.

The MIBs that we chose to implement were:

- the SNMPv2-MIB that provides generic system information,
- the IF-MIB that provides information ans statistics about the network interface of the sensor,
- and the ENTITY-SENSOR-MIB, that provides access to the actual sensors data.

Then, we were facing a problem, as the Contiki versions provided by the SensTools project were only stable releases, and we found it difficult to track the development version of Contiki with them. We then decided to create our own WSN430 drivers and platform definition for Contiki, well integrated with the development repository, and reusing as much as possible of already existing code. Our next step in this direction will be to have our contribution officially integrated in the Contiki OS.

And finally, we devised and implemented a COAP server discovery protocol using the piggybacking technique, which allows every node that offers COAP resources to announce itself to the grounded root of the sensor network, without requiring the transmission of additional packets.
MAESTRO Project-Team

5. New Results

5.1. IP networks

Participants: Eitan Altman, Konstantin Avrachenkov.

5.1.1. Interdisciplinary study of the Internet access and of network neutrality

In our previous research we have identified large inefficiencies that occur when one allows one type of provider (e.g. access provider) to impose costs on another type of provider (e.g. content provider). This part in which E. Altman collaborated with P. Bernhard (INRIA project-team BIOCORE), S. Caron and G. Kesidis (both from Pennsylvania State Univ., USA), J. Rojas-Mora (Univ. of Barcelona, Spain), and S. Wong (Univ. of A Coruña, Spain) has now appeared in [96].

This investigation has been pursued in various directions. In [42], E. Altman, A. Legout (INRIA project-team PLANETE) and Y. Xu (Univ. Avignon/LIA) have studied a hierarchical structure of ISPs, the economic impact of some caching placement policies, and more complex demand functions (the demands of users for content). In [33], E. Altman and the law specialist S. Wong (Univ. of A Coruña, Spain) analyze in cooperation with the economist J. Rojas-Mora (Univ. of Barcelona, Spain) the impact of legislation related to network neutrality on the quality of service for the end users.

5.1.2. Adaptive monitoring system for IP networks

The remarkable growth of the Internet infrastructure and the increasing heterogeneity of applications and users’ behavior make more complex the manageability and monitoring of ISP networks and raises the cost of any new deployment. The main consequence of this trend is an inherent disagreement between existing monitoring solutions and the increasing needs of management applications. In this context, in [62] K. Avrachenkov, I. Lassoued, A. Krifa and C. Barakat (all three from INRIA project-team PLANETE) present the design of an adaptive centralized architecture that provides visibility over the entire network through a network-wide cognitive monitoring system. Practically, given a measurement task and a constraint on the volume of collected information, the proposed architecture drives the sampling rates on the interface of network routers to achieve the maximum possible accuracy, while adapting itself to any change in network traffic conditions. The authors tune the system parameters with the help of FAST sensitivity test.

5.1.3. Size based scheduling

Size-based scheduling is a promising solution to improve the response time of small flows (mice) that have to share bandwidth with large flows (elephants). To do this, one important task is to track the size of the ongoing flows at the router. However, most of the proposed size-based schedulers either employ the trivial way of tracking the size information of all flows, or require changes at end-hosts. Hence, either they are not scalable or they increase complexity. In [55], E. Altman, D. Mon Divakaran (IIT Mandi, India) and P. Vicat-Blanc Primet (Lyatiss, France) have proposed a new way of performing size-based scheduling in a practical and scalable fashion, by identifying and ‘de-prioritizing’ elephants only at times of high load. They exploit TCP’s behavior by using a mechanism that detects a window of packets - called spikes - when the buffer length exceeds a certain threshold. This spike-detection is used to identify elephant flows and thereafter de-prioritize them. Two-level processor-sharing (TLPS) scheduling is employed to schedule flows in two queues, one with the high-priority flows, and the other with the de-prioritized flows. They show that the proposed mechanism not only improves the response time of mice flows in a scalable way, but also gives better response times to other flows by treating them preferentially as long as they do not overload the high-priority queue.
5.1.4. Accuracy of fluid models for bandwidth-sharing networks

Optimal control of stochastic bandwidth-sharing networks is typically difficult. In order to facilitate the analysis, deterministic analogues of stochastic bandwidth-sharing networks, the so-called fluid models, are often chosen for analysis, as their optimal control can be found more easily. The tracking policy translates the fluid optimal control policy back into a control policy for the stochastic model, so that the fluid optimality can be achieved asymptotically when the stochastic model is scaled properly. In [20] K. Avrachenkov, A. Piunovsky and Y. Zhang (both from the University of Liverpool, UK) study the efficiency of the tracking policy, that is, how fast the fluid optimality can be achieved in the stochastic model with respect to the scaling parameter. In particular, the result of [20] shows that, under certain conditions, the tracking policy can be as efficient as feedback policies.

5.1.5. Bootstrap method for simulating bandwidth sharing

In [71], E. Altman, T. Jimenez and J. Rojas-Mora (both from Univ. Avignon/LIA) identify difficulties in evaluating through simulations the expected transfer time of a file when several TCP connections share a common bottleneck buffer. The main difficulties are due to the fact that the file size over the Internet has been reported to have a Pareto distribution with parameter smaller than 1.5. This implies that the number of ongoing connections as well as the sojourn times have infinite variance. This has two implications: one cannot estimate the confidence intervals for simulation based on the CLT (central Limit Theory approach), and the duration of the simulations needed to get to steady state is very long. The authors show how to solve both problems by the use of the bootstrap approach.

5.2. Wireless communications

Participants: Sara Alouf, Eitan Altman, Konstantin Avrachenkov, Nicaise Choungmo Fofack, Mahmoud El Chamie, Majed Haddad, Manjesh Kumar Hanawal, Philippe Nain, Giovanni Neglia, Manoj Panda, Sreenath Ramanath.

5.2.1. Green networking

Green networking is a new trend in network design that is more aware of the impact of technology on the environment and on humans. Reducing energy has been so far the main concern in that approach, and much of the research has been devoted to understanding the tradeoffs between reducing energy and other performance measures such as coverage and delay.

For several years we have been contributing to this research effort, many of which have been summarized in the survey [67] by E. Altman in collaboration with G. S. Paschos (Center for Research and Technology, Hellas, Greece), P. Mannersalo (VTT, Finland), S. Stanczak (HH-Fraunchofer, Berlin, Germany) and L. Tassiulas (Univ. of Thessaly, Greece). In particular, much of the work involving members of MAESTRO that had appeared in previous years in conferences concerning energy saving in WiMax has now appeared in a journal publication [21] by A. P. Azad, S. Alouf, E. Altman, in cooperation with V. Borkar (TATA Inst. of Fundamental Research, Mumbai, India) and G. S. Paschos (Center for Research and Technology, Hellas, Greece).

In 2011 we started investigating policies for switching off base stations using two new tools: multimodularity and stochastic geometry. The latter has been used by E. Altman and M. K. Hanawal (also with Univ. Avignon/LIA), in cooperation with R. El-Azouzi (Univ. Avignon/LIA) and S. Shamai (Technion, Israel) in [98] to study the tradeoffs related to the uplink, and by E. Altman in cooperation with C. Hasan and J. M. Gorce (both from INSA-Lyon and INRIA project-team SWING) in [89] for the downlink. Optimal policies were obtained within the class of policies that switch off base stations with some fixed probability but independently of each other. To relax this restriction of independence and thus obtain even better policies, S. Ramanath and E. Altman, in collaboration with V. Kavitha (Univ. Avignon/LIA), have used in [69] the theory of multimodularity, which is the discrete counterpart of convexity. Among the most striking points in this research has been the observation in [98] that the conventional energy saving approach can have the opposite effect on the humans in the uplink: when the base station closest to a mobile phone is switched off (for energy...
saving) then the mobile phone has to transmit with a larger power so as to reach a more remote base station. It turns out that the main source of radiation to the human brain is indeed the uplink transmission, which implies that switching off base stations could cause more exposure to radiation. This is of particular concern in view of the announcement by the World Health Organization (May 31, 2011) that cell phones cause cancer.

5.2.2. Cellular networks with continuous connectivity

In [65], S. Alouf and V. Mancuso (Institute IMDEA Networks, Madrid, Spain) analyze the power save and its impact on web traffic performance when customers adopt the continuous connectivity paradigm. Considering realistic http traffic, they evaluate the user access delay, the download time and the expected economy of energy in the cell. The model, validated through packet-level simulations, shows that dramatic energy save can be achieved by both mobile users and base stations. In case of Poisson arrivals, the aggregate behavior of a base station’s users is studied by means of a processor-shared queueing system [105]. The model can be used to maximize the base station energy savings under a given set of QoS performance constraints. With the participation of N. Choungmo Fofack, the work in [65] has been complemented with a sensitivity analysis [95]. The impact of model parameters on the performance and cost metrics is thoroughly assessed.

5.2.3. Power allocation in multicell networks

Power allocation to satisfy user demands in the presence of large number of interferers in a multicellular network is a challenging task. Further, power to be allocated depends upon the system architecture, for example, upon components like coding, modulation, transmit precoder, rate allocation algorithms, available knowledge of the interfering channels, etc. This calls for an algorithm via which each base station in the network can simultaneously allocate power to their respective users so as to meet their demands (when they are within the achievable limits), using whatever information regarding the other users is available. In [70], S. Ramanath, V. Kavitha (Univ. Avignon/LIA) and M. Debbah (SUPELEC) devise such an algorithm which is in fact universal: the proposed algorithm works from a fully cooperative setting to almost no cooperation and or for any configuration of modulation, rate allocation, etc. schemes. The algorithm asymptotically satisfies the user demands, running simultaneously and independently within a given total power budget at each base station. Further, it requires minimal information to achieve this: every base station needs to know its own users demands, its total power constraint and the transmission rates allocated to its users in every time slot. The authors formulate the power allocation problem in a system specific game theoretic setting, define system specific capacity region and analyze the proposed algorithm using ordinary differential equation (ODE) framework. Simulations confirm the effectiveness of the proposed algorithm.

5.2.4. Small cell networks

In [28], S. Ramanath and E. Altman, in collaboration with V. Kavitha (Univ. Avignon/LIA), characterize the performance of Picocell networks in the presence of moving users. They model various traffic types between base-stations and mobiles as different types of queues. They derive explicit expressions for the expected waiting time, service time and drop/block probabilities for both fixed and random velocity of mobiles. They obtain (approximate) closed-form expressions for optimal cell size when the velocity variations of the mobiles is small for both non-elastic and elastic traffic. They conclude from the study that, if the expected call duration is long enough, the optimal cell size depends mainly on the velocity profile of the mobiles, its mean and variance. It is independent of the traffic type or duration of the calls. Further, for any fixed power of transmission, there exists a maximum velocity beyond which successful communication is not possible. This maximum possible velocity increases with the power of transmission. Also, for any given power, the optimal cell size increases when either the mean or the variance of the mobile velocity increases.

5.2.5. New concepts in fair resource allocation

Fair resource allocation is usually studied in a static context, in which a fixed amount of resources is to be shared. In dynamic resource allocation one usually tries to assign resources instantaneously so that the average share of each user is split fairly. The exact definition of the average share may depend on the application, as different applications may require averaging over different time periods or time scales. Our main contribution is to introduce new refined definitions of fairness that take into account the time over which one averages the
performance measures. In [39] E. Altman, K. Avrachenkov and S. Ramanath examine how the constraints on the averaging durations impact the amount of resources that each user gets. The authors apply this new concept in [68] to spectrum allocation and indoor-outdoor femtocells.

5.2.6. Self organization in cellular networks

Time-slots and frequencies are contended in cellular networks and their allocation is determined by base stations. For scalability purposes the resource allocation is decentralized, so that base stations do not share their information with each other. Actions are often taken based on partial information on the system. In particular, the statistics of the channels are often not available. Scheduling decisions of a base station concerning mobiles in its cell cause interference in other cells and there is thus a need to dynamically adjust to interference and to converge to a satisfactory operation point. This has motivated a large amount of work on self-organization in cellular networks based on OFDMA. E. Altman, Z. Altman, R. Combes (both from Orange Labs, Issy les Moulineaux) and M. Haddad have written a series of papers on self-organization. In [53] and [51] self-organization in interference coordination is studied. In [50], R. Combes, Z. Altman and E. Altman, further propose and analyze a self-optimization method for coverage-capacity optimization in OFDMA networks with MIMO. Moreover, they study in [52] self-organization when adding relays so as to increase coverage. Static and dynamic resource sharing mechanisms are investigated. In the static case they use a queuing model to calculate the optimal resource sharing strategy and the maximal capacity of the network analytically. The influence of relay planning and number of deployed relays is investigated, and gains resulting from good planning are evaluated analytically. Self-optimizing dynamic resource allocation is tackled using a Markov Decision Process (MDP) model.


Self-organization has also been used in the past to obtain opportunistic scheduling in a way that achieves proportional fair resource sharing. In [23], R. Combes, Z. Altman (both from Orange Labs, Issy les Moulineaux) and E. Altman, have extended this to the general $\alpha$-fair concept. A dynamic choice of the factor $\alpha$ is proposed, which has the interpretation of trading optimality with fairness in a dynamic way.

5.2.7. Dynamic networks

In source routing, a complete path is chosen for a packet to travel from source to destination. While computing the time to traverse such a path may be straightforward in a fixed, static graph, doing so becomes much more challenging in dynamic graphs, in which the state of an edge in one timeslot (i.e., its presence or absence) is random, and may depend on its state in the previous time step. The traversal time is due to both time spent waiting for edges to appear and time spent crossing them once they become available. In [99], P. Nain in collaboration with A. Bar-Noy (City University of New York), P. Basu (Raytheon BBN Technologies), M. P. Johnson (Pennsylvania State University), F. Yu (City University of New York) and D. Towsley (University of Massachusetts at Amherst) computes the expected traversal time (ETT) for a routing path in a number of special cases of stochastic edge dynamics models, and for three edge failure models, culminating in a surprisingly challenging yet realistic setting in which the initial configuration of edge states for the entire path is known. We show that the ETT for this “initial configuration” setting can be computed in quadratic time, by an algorithm based on probability generating functions. The authors also give several linear-time upper and lower bounds on the ETT.

5.2.8. Sensor networks

In many application scenarios sensors need to calculate the average of some local values, e.g. of local measurements. A possible solution is to rely on consensus algorithms. In this case each sensor maintains a local estimate of the global average, and keeps improving it by performing a weighted sum of the estimates of all its neighbors. The number of iterations needed to reach an accurate estimate depends on the weights used at each sensor. K. Avrachenkov, G. Neglia and M. El Chamie have proposed a new average consensus algorithm, where each sensor selects its own weights on the basis of some local information about its neighborhood [45]. In realistic sensor network topologies, the algorithm shows faster convergence than other existing consensus protocols.
5.2.9. Delay and disruption-tolerant networks (DTNs)

5.2.9.1. Applying risk sensitive control to delay tolerant networks

When controlling the propagation of a message in DTNs, the objective is often to maximize the successful delivery probability of a message within a given deadline. It takes often the form of the expectation of the exponent of some integral cost. So far, models involving such costs have been solved by interchanging the order of expectation and the exponential function. While reducing the problem to a standard optimal control problem, this interchange is only tight in the mean-field limit obtained as the population tends to infinity. In [41] E. Altman, V. Kavitha (Univ. Avignon/LIA), F. De Pellegrini (Create-Net, Trento, Italy), V. Kamble (UC Berkeley, CA, USA) and V. Borkar (TATA Inst., Mumbai, India), identify a general framework from optimal control in finance, known as risk sensitive control, which allows handling the original (multiplicative) cost and obtaining solutions to several novel control problems in DTNs. New optimal control problems which consider the effect of wireless propagation path loss factor and the power constraints at the source and or the destination are proposed for DTNs within this framework. Optimal policies of non-threshold type are found.

5.2.9.2. Multiple destinations

In [73], C. Singh, A. Kumar and R. Sundaresan (all three from IISC Bangalore, India) in collaboration with E. Altman, use Markov Decision Processes to study optimal policies for propagation of contents in DTNs to multiple destinations. They obtain structural properties for a discretized system which allows them to derive the structure of optimal policies to the original problem.

5.2.9.3. Reliable unicast and multicast

In case the DTN does not deliver a packet within some time $T$, it has to be retransmitted. In [36] E. Altman and M. Panda, in collaboration with T. Chahed and A. Ali, (both from Telecom SudParis) and L. Sassatelli (Univ. Nice Sophia Antipolis/I3S), propose protocols for unicast and for multicast that render the connection reliable. These protocols include ACKs and retransmissions. The authors compute the value of $T$ that optimizes the throughput and address energy consumptions aspects.

5.2.9.4. Network coding

In [18], E. Altman studies, in cooperation with F. De Pellegrini (Create-Net, Trento, Italy), how to improve the performance of DTNs by adding network coding. The latter has the effect of efficiently adding spatial redundancy to the network. They identify the structure of optimal policies, which are shown not always to be of a threshold type.

G. Neglia, in collaboration with X. Zhang (Fordham University, New York) and J. Kurose (University of Massachusetts at Amherst), has published a survey on the application of network coding to DTNs [78].

5.2.9.5. Ferry based local area networks

Polling systems are used to model the Ferry assisted Wireless LANs and thereby to obtain the stationary workload performance. Not much theory is available for calculating the stationary workload of polling systems with arrivals in a continuum. In [103], V. Kavitha (Univ. Avignon/LIA) and E. Altman propose a discretization approach, by which the so-called “pseudo conservation law” of the discrete polling systems is utilized to derive the stationary performance of continuous polling systems. The continuous polling results are used in deriving optimal ferry routes.

5.2.9.6. Adaptive epidemic routing in DTNs

G. Neglia and R. Masiero (University of Padua, Italy) have explored a recently proposed optimization framework that relies on local sub-gradient methods and consensus algorithms. The research is described in MAESTRO 2010 activity report and has appeared in [66].

5.2.9.7. Routing in quasi-deterministic networks

G. Neglia, U. Acer (Bell labs Antwerp), P. Giaccone and S. Tarapiah (both from Politecnico di Torino, Italy) and D. Hay (Hebrew University of Jerusalem), have investigated routing in DTNs where the underlying node mobility is known in advance but can be modified by random effects. The research is described in MAESTRO 2010 activity report and has appeared in [35] and [94].
5.3. Information systems

Participants: Eitan Altman, Konstantin Avrachenkov, Nicaise Choungmo Fofack, Majed Haddad, Alain Jean-Marie, Dorian Mazauric, Philippe Nain, Marina Sokol.

5.3.1. Web crawler optimization

A typical web search engine consists of three principal parts: crawling engine, indexing engine, and searching engine. The work [19] by K. Avrachenkov and P. Nain, together with A. Dudin, V. Klimenok, and O. Semenova (all three from Belarusian State University, Belarus), aims to optimize the performance of the crawling engine. The crawling engine finds new web pages and updates existing web pages in the database of the web search engine. The crawling engine has several robots collecting information from the Internet. The authors first calculate various performance measures of the system (e.g., probability of arbitrary page loss due to the buffer overflow, probability of starvation of the system, average time waiting in the buffer). Intuitively, one would like to avoid system starvation and at the same time to minimize the information loss. The authors formulate the problem as a multi-criteria optimization problem and solve it in the class of threshold policies. The authors consider a very general web page arrival process modeled by Batch Marked Markov Arrival Process and a very general service time modeled by Phase-type distribution. The model has been applied to the performance evaluation and optimization of the crawler designed by INRIA-MAESTRO team in the framework of the RIAM INRIA-Canon research project (see MAESTRO 2006 and 2007 activity reports).

5.3.2. PageRank node centrality

In [48] K. Avrachenkov and M. Sokol, together with D. Nemirovsky (former MAESTRO team member), E. Smirnova (INRIA project-team AXIS) and N. Litvak (University of Twente, The Netherlands), study a problem of quick detection of top-k Personalized PageRank (PPR) lists. This problem has a number of important applications such as finding local cuts in large graphs, estimation of similarity distance and person name disambiguation. The authors suggest that two observations are important when finding top-k PPR lists. Firstly, it is crucial that one detects fast the top-k most important neighbors of a node, while the exact order in the top-k list and the exact values of PPR are by far not so crucial. Secondly, by allowing a small number of “wrong” elements in top-k lists, one achieves great computational savings, in fact, without degrading the quality of the results. Based on these ideas, the authors propose Monte Carlo methods for quick detection of top-k PPR lists. We demonstrate the effectiveness of these methods on the Web and Wikipedia graphs, provide performance evaluation and supply stopping criteria.

5.3.3. Analysis of YouTube

E. Altman and M. Haddad, in collaboration with S.-E. Elayoubi (Orange Labs, Issy les Moulineaux), R. El-Azouzi, T. Jimenez and Y. Xu (all three from Univ. Avignon/LIA) have been investigating streaming protocols similar to the one used by YouTube. After preparing a survey on the state-of-the-art in [57], they used Ballot theorems in [106] in order to compute the starvation probabilities (these are the probability that the queue empties before completing to send a streaming application).

This work is carried out in the framework of the Grant with Orange Labs (see Section 6.3) on “Quality of Service and Quality of Experience”.

5.3.4. Peer-to-peer networks

5.3.4.1. Real-time control of contents download

In the course of the VOODOO project, the question of assessing the theoretical limits of prefetching information in real-time arose. Given a network bandwidth and a graph of documents, is it possible to download documents in advance, so that the document surfer is never blocked because of missing information? The problem is modeled using a “cops-and-robbers” game and some of its algorithmic properties are derived. This work of A. Jean-Marie and D. Mazauric is joint with F. Fomin (Univ. Bergen) and F. Giroire and N. Nisse (both from INRIA project-team MASCOTTE) [86].
5.3.4.2. P2P traffic classification

P2P downloads still represent a large portion of today’s Internet traffic. More than 100 million users operate BitTorrent and generate more than 30% of the total Internet traffic. According to the Wikipedia article about BitTorrent, the traffic generated by BitTorrent is greater than the traffic generated by Netflix and Hulu combined. Recently, a significant research effort has been done to develop tools for automatic classification of Internet traffic by application. The purpose of the work [47] by K. Avrachenkov and M. Sokol, together with A. Legout (INRIA project-team PLANETE) and P. Gonçalves (INRIA project-team RESO), is to provide a framework for subclassification of P2P traffic generated by the BitTorrent protocol. The general intuition is that users with similar interests download similar contents. This intuition can be rigorously formalized with the help of graph based semi-supervised learning approach. In particular, the authors propose to work with PageRank based semi-supervised learning method, which scales well with very large volumes of data.

5.3.4.3. BitTyrant

The success of BitTorrent has fostered the development of variants to its basic components. Some of the variants adopt greedy approaches aiming at exploiting the intrinsic altruism of the original version of BitTorrent in order to maximize the benefit of participating to a torrent. G. Neglia, D. Carra (University of Verona, Italy), P. Michiardi and F. Albanese (both from INSTITUT EURECOM) have studied BitTyrant, a recently proposed strategic client. The research is described in MAESTRO 2008 activity report. Results have been extended and supported by PlanetLab experiments in [22].

5.3.5. Content-centric networks

In [100] N. Choungmo Fofack, P. Nain and G. Neglia, together with D. Towsley (University of Massachusetts at Amherst), provide building blocks for the performance evaluation of Content Centric-like Networks (CCNs). In CCNs if a cache receives a request for a content it does not store, it forwards the request to a higher-level cache, if any, or to the server. When located, the document is routed on the reverse-path and a copy is placed in each cache along the path. In this work the authors consider a cache replacement policy based on Time-to-Lives (TTLs) like in a DNS network. A local TTL is set when the content is first stored at the cache and is renewed every time the cache can satisfy a request for this content (at each hit). The content is removed when the TTL expires. Under the assumption that requests follow a renewal process and the TTLs are exponential random variables, we determine exact formulas for the performance metrics of interest (average cache occupancy, hit and miss probabilities/rates) for some specific architectures (a linear network and a tree network with one root node and \( N \) leaf nodes). For more general topologies and general TTL distributions, an approximate solution is proposed. Numerical results show the approximations to be accurate, with relative errors smaller than \( 10^{-3} \) and \( 10^{-2} \) respectively for exponentially distributed and constant TTLs.

This work is carried out in the framework of the Grant with Orange Labs on “Content-centric networks” (Section 6.2).

5.4. Game theory applied to networking

Participants: Eitan Altman, Konstantin Avrachenkov, Majed Haddad, Manoj Panda, Giovanni Neglia.

5.4.1. Resource allocation in wireless networks

5.4.1.1. Power control

In [14] E. Altman, K. Avrachenkov and A. Garnaev (St. Petersburg State University, Russia), study power control for Gaussian interference channel in optimization and game frameworks. In the optimization framework there is a single decision maker who assigns network resources and in the game framework users share the network resources according to Nash equilibrium. The authors enhance the water-filling technique with explicit analytic solutions. The authors also provide an alternative simple proof of the convergence of the Iterative Water Filling Algorithm. Finally, the authors compare the non-cooperative approach with the cooperative approach and show that the non-cooperative approach results in a more fair resource distribution.
There has been a debate between those proposing protocols based on a centralized controller and those favoring decentralized protocols based on non-cooperative game theory. In [26] E. Altman and M. Haddad, together with S.-E. Elayoubi and Z. Altman (both from Orange Labs, Issy les Moulineaux), consider a situation where a base station lets mobiles take power control decisions in some system states and imposes actions in other states. The authors study how best to choose what information to make available and how mobiles should react.

5.4.1.2. Joint power and rate allocation

In [63] X. Lei and L. Cottatellucci (both from Institut Eurecom) and K. Avrachenkov consider a block fading interference channels with partial channel state information and address the issue of joint power and rate allocation in a game theoretic framework. Resource allocation algorithms based on Bayesian games are proposed. The existence, uniqueness, and some stability properties of Nash equilibria are analyzed. For some asymptotic setting, closed-form expressions of Nash equilibria are also provided.

5.4.1.3. Jamming in wireless networks

Jamming is a form of a denial of service attack in which an adversary can degrade the quality of the reception by creating interference. One can study jamming both in the purpose of protecting a wireless network against such attack or, on the contrary, in order to efficiently disrupt the communications of some adversary. In both cases jamming is part of a conflict for which game theory is an appropriate tool. In [15] E. Altman, K. Avrachenkov and A. Garnaev (St. Petersburg State University), investigate the effect of partially available information in which the user does not even know whether or not the jammer is indeed present. The problem is formulated as a zero-sum game. The authors find the equilibrium strategies in closed-form and specify the range of sub-carriers where the user can expect the jamming attack.

5.4.1.4. Channel access

In WiFi networks, mobile nodes compete for accessing the shared channel by means of a random access protocol called Distributed Coordination Function (DCF), which is long term fair. Selfish nodes could benefit from violating the protocol and increasing their transmission probability. G. Neglia, I. Tinnirello and L. Giarré (University of Palermo, Italy) have been studying the interaction of selfish nodes in the last two years (the research activity is described in MAESTRO 2009 and MAESTRO 2010 activity reports). [31], [74] further extend the results to a heterogeneous scenario, where nodes have different requirements in terms of uplink/downlink ratios.

5.4.2. Network formation games

The continued growth of computer networks such as the Internet has raised the interest in understanding how networks get formed. The design of such networks is generally carried out by a large number of self-interested actors (users, Internet Service Providers ...), all of whom seek to optimize the quality and cost of their own operation. Previous works have addressed the “Network Formation” problem considering almost exclusively networks designed by selfish users, which can be consistently suboptimal. In [46] K. Avrachenkov and G. Neglia, together with J. Elias (University Paris Descartes), F. Martignon (University Paris-Sud 11), and L. Petroxyan (St. Petersburg State University, Russia), address the network formation issue using cooperative game theory, which permits to study ways to enforce and sustain cooperation among agents. Both the Nash bargaining solution and the Shapley value are investigated. After the comparison of these two approaches, the authors conclude that the Nash bargaining solution is more suitable to enforce cooperation in the network formation game in terms of cost allocation to users and computation time to get the solution.

5.4.2.1. Network design with socially-aware users

In many scenarios network design is not enforced by a central authority, but arises from the interactions of several self-interested agents. This is the case of the Internet itself. K. Avrachenkov and G. Neglia, in collaboration with J. Elias (University Paris Descartes) and F. Martignon (University Paris-Sud 11), have proposed two novel socially-aware network design games. The research has been described in MAESTRO 2010 activity reports. [24] extends the results for the case when users’ utility functions incorporate a socially-aware component.
5.4.2.2. Stochastic games for cooperative network routing

In [64], K. Avrachenkov, L. Maggi and L. Cottatellucci (both from INSTITUT EURECOM) consider a system where several providers share the same network and control the routing in disjoint sets of nodes. They provide connection toward a unique server (destination) to their customers. The objective is to facilitate the design of the available network links and their costs such that all network providers are interested in cooperating and none of them withdraw from the coalition. More specifically, the authors establish the framework of a coalition game by providing an algorithm to compute the transferable coalition values. As by-product, the authors apply the proposed algorithm to two-player games both in networks subject to hacker attacks and in epidemic networks.

5.4.2.3. Association games

Using tools from coalition game theory, E. Altman, in cooperation with C. Singh (IISc Bangalore, India), considers in [72] a wireless framework in which several mobile terminals can receive and decode the same signal of the base station, and where the cost for broadcasting is taken to be the transmission power. They begin by proposing various schemes to share the cost and study their properties. Then, they study the association with partial information: an arriving user knowing its location has to decide without knowledge of the location of the other users and their number whether to join the multicast tree and pay according to a given cost sharing scheme, or to have a unicast connection at a given cost. The unicast alternative that each mobile has, results in a limitation on the coverage (area covered by the multicast session) and on the capacity (number of mobiles connected to the multicast session). The authors derive the expected capacity and coverage as a function of the cost sharing policy. This work is extended in [58] to the case of several base stations by E. Altman in collaboration with C. Hasan and J. M. Gorce (both from INSA Lyon and INRIA project-team SWING).

5.4.3. Routing games

In [40], E. Altman, M. Panda and A. Estanisla (Master student at UPMC) study ring networks extensively used in both road traffic and telecommunications (in local area networks) in which each source with a given origin and destination on the ring, can split its traffic and send some part in one direction of the ring and some other part in the other direction. They compute the equilibria and find out that due to non-cooperation, much traffic is sent at equilibrium along long paths.

In [16], E. Altman, O. Pourtallier (INRIA project-team COPRIN), T. Jimenez (Univ. Avignon/LIA) and H. Kameda (Univ. Tsukuba, Japan) study a load balancing processor sharing problem. The classical framework of routing games turned out not to apply here. Indeed, it had been used to model situations where the flow from each class of users is split among paths without any information on the realization of the sizes of each packet. In contrast, in this paper, each individual knows its size. The authors have succeeded in computing the equilibrium within the new setting.

Collusion is the situation where several players decide to cooperate and to choose their actions as if they were a single player - each player maximizes the sum of utilities of that group instead of only its own utility. In [90], E. Altman in collaboration with Y. Hayel (Univ. Avignon/LIA) and H. Kameda (Univ. Tsukuba, Japan), has proposed various concepts that evaluate the impact of collusions. The authors have further studied collusions in routing games and identified situations where collusions are bad for all players: both those that collide loose in performance as well as those who remain independent.

5.5. Stochastic processes, queueing, control theory and game theory

Participants: Eitan Altman, Julien Gaillard, Majed Haddad, Alain Jean-Marie.

5.5.1. Convergence of rolling horizon control

In collaboration with E. Della Vecchia and S. Di Marco (both from National Univ. Rosario, Argentina), A. Jean-Marie has investigated the performance (convergence and error bounds) of the Rolling Horizon heuristic for optimal stochastic control and stochastic games in different modeling situations.
In the case of the long-term average expected gain, they have shown \[ 85 \] that convergence occurs whenever the value iteration algorithm converges. They have then considered zero-sum semi-Markov games with discounted payoff \[ 54 \], \[ 76 \], for which they have proved geometric convergence under the usual assumptions of the literature.

5.5.2. Impulse control versus continuous control

Impulse control is a modeling framework of optimal control theory, in which the control actions can provoke instantaneous changes in the value of the state. For modelers, it has the features of both continuous-time and discrete-time models, and it can help understand which one to choose in a given optimization situation. A. Jean-Marie has studied the question in conjunction with K. Erdlenbruch (CEMAGREF), M. Tidball (INRA) and M. Moreaux (Univ. Toulouse 1). In a quite generic single-dimensional model, they show that the optimality of impulse policies with respect to “smooth” control policies is strongly related to a submodularity property of the instantaneous cost function \[ 101 \].

5.5.3. Routing games

Several fundamental results have been obtained in routing games that model finite number of sources of traffic (players) who decide how to split the traffic among various paths. When the number of players is large, the Wardrop equilibrium concept is often used, where the problem is modeled as one with a continuum of decision makers where each has a negligible impact (non atomic game) on other’s performance. E. Altman and his co-workers have studied the question of whether Wardrop equilibrium is a good approximation for a problem with finitely many players for which the Nash equilibrium is the solution concept. In \[ 38 \], E. Altman, in collaboration with Z. Altman, R. Combes (both from Orange Labs, Issy les Moulineaux) and S. Sorin (Univ. Pierre and Marie Curie (UPMC)) establishes the convergence under mild convexity assumptions on the link costs (or delays). The proof is based on yet another fundamental result derived in that reference and that was later extended in \[ 44 \] by E. Altman in collaboration with O. Pourtallier (INRIA project-team COPRIN), T. Jimenez (Univ. Avignon/LIA) and H. Kameda (Univ. Tsukuba, Japan), that states that if there is some symmetry in a network then any Nash equilibrium will inherit the symmetric properties (for example, if two users have the same source and destination and the same demand then at equilibrium, they will send the same amount of traffic over each link).

In all the above work there is an assumption that the link cost (or delay) per packet is class independent (it depends on the flows through the link only through their sum). In the case of Wardrop equilibrium this assumption implies that the game has an equivalent global optimization problem whose solution coincides with the equilibrium. The link cost evaluated at some \( x \) in the equivalent problem is the integral of the original link cost (from 0 to \( x \)) and is in fact a potential. In the case of class dependent cost, that is, when the cost depends in other ways on the traffic of each class then the result of the integration may depend on the path and one cannot transform the problem to an equivalent optimization one. H. Kameda and J. Li (both from Univ. Tsukuba, Japan) in collaboration with E. Altman, identify in \[ 27 \] other class-dependent cost that have the property of a field, that is, it can be expressed as the gradient of a potential. They obtain the Wardrop equilibrium and study its properties.

Another difficulty occurs in rouging games when the paths available are not the same for all users. This is the case, in particular, when there are priorities. This problem is addressed in \[ 25 \] by J. Elias (University Paris Descartes), F. Martignon (University Paris-Sud 11), A. Capone (Politecnico di Milano, Italy) and E. Altman within an application to non-cooperative spectrum access in cognitive radio networks.

5.5.4. Bio-inspired paradigms

5.5.4.1. Epidemiology

For several years now, E. Altman has been developing techniques for dynamic optimal control and games in cooperation with with S. Sarkar’s group from the University of Pennsylvania (which used to be part of the DAWN associated team with MAESTRO). This year this collaboration has resulted in three additional publications co-authored by M.H.R. Khouzani and S. Sarkar (both from Univ. of Pennsylvania, PA, USA) and E. Altman \[ 61 \], \[ 60 \], \[ 104 \]. All three papers use the Pontriagin maximal principle to derive the structure
of optimal policies applied to a mean-field approximation of the problem. The first two papers do that in a context of optimal control theory while the third one does it in the context of a dynamic game.

5.5.4.2. Sequential anonymous games (SAG)

Sequential Anonymous Games (SAG) can be viewed as an extension of Markov Decision evolutionary games. In both formalisms there are many players modeled as a continuum number of players. A Markov chain is associated with each player. There are several types of players. The fraction of players in each class is called a global state and the state of the Markov chain of an individual is called the individual state. An individual chooses at some sequential decision opportunities actions. It earns some immediate reward (fitness) at each slot and moves with some probability to another individual state. In SAG, both the transition probabilities and the immediate fitness of an individual depend on its current state and action as well as on the current global system state. The latter evolves according to some function averaged over the fitness of the individuals in each class (the fraction of individuals in a class grows if they do better than those in the other classes). In [75] E. Altman investigates, in collaboration with P. Wiecek (Wroclaw Univ. of Technology, Poland), the case where the objective of an individual is to maximize either its total expected fitness during its life time or its expected average fitness. The authors establish the existence of equilibria and study its properties. Applications to power control have appeared in [32] by E. Altman, in collaboration with P. Wiecek (Wroclaw Univ. of Technology, Poland) and Y. Hayel (Univ. Avignon/LIA).

5.5.4.3. Markov decision evolutionary games (MDEGs)

Since his 2004 Infocom paper, E. Altman has been working on this novel paradigm. The model is similar to that of the previous paragraph (SAG) except that in MDEG both immediate fitness and transition probabilities depend linearly on the global state. This reflects a scenario where the interaction between a player and the rest of the population occurs through pairwise interactions: each player encounters from time to time a randomly chosen other player and it finds itself playing a matrix game with that player. The entries of the matrix corresponding to each player as well as the transition probabilities for each player depend on the individual states of the players. E. Altman has applied this model to the dynamic Hawk and Dove game, in which individuals have to choose the degree of aggressiveness in their behavior as a function of their energy state.

Below are three publications both with biological applications and applications to wireless communications (where depending on one’s remaining battery energy, one has to decide at what power to transmit). The first publication in [30], by E. Altman, H. Tembine (SUPELEC), R. El-Azouzi and Y. Hayel (both from Univ. Avignon/LIA), lays the foundations of MDEGs and presents the application to power control in which the individual state is the battery level of energy. The second publication in [59], by Y. Hayel (Univ. Avignon/LIA), E. V. Belmega (SUPELEC) and E. Altman, studies theoretical aspects that arise in case that the global state cannot represent the fractions of different populations but rather their actual size. The third publication in [97] by E. Altman, J. Gaillard, M. Haddad and P. Wiecek (Wroclaw Univ. of Technology, Poland) again studies MDEGs (as in the first paper), but restricts to policies that use static policies: the same mixed strategy is taken by a player at each state. The authors manage to compute explicitly the equilibrium in this game within this class of policies.

5.5.4.4. Delayed evolutionary games

Evolutionary game theory includes much theory on the description of the global system state as a function of the fitness of individuals. The models are often described through differential equations (e.g. the “replicator dynamics”). In many scenarios it is realistic to consider delays between the moment that one receives a given fitness till this is translated to a change in the population size. For example, if the lifetime of a computer is three years then an application that performs better with one computer may take more than a year till it is adopted by other users who do not have the same computer. In [30], H. Tembine (SUPELEC), E. Altman, R. El-Azouzi and Y. Hayel (both from Univ. Avignon/LIA) investigate instability phenomena that are introduced by the delay and derive necessary stability conditions.
6. New Results

6.1. Network Design and Optimization

*Participants:* Jean-Claude Bermond, Nathann Cohen, David Coudert, Frédéric Giroire, Dorian Mazauric, Joanna Moulierac, Nicolas Nisse, Ronan Pardo Soares, Issam Tahiri.

### 6.1.1. Backbone and Broadband Networks

Network design is a very wide subject that concerns all kinds of networks. We mainly study telecommunications networks which can be either physical networks (backbone, access, wireless, ...) or virtual (logical) ones. The objective is to design a network able to route a (given, estimated, dynamic, ...) traffic under some constraints (e.g. capacity) and with some quality of service (QoS) requirements. Usually the traffic is expressed as a family of requests with parameters attached to them. In order to satisfy these requests, we need to find one (or many) path(s) between their end nodes. The set of paths is chosen according to the technology, the protocol or the QoS constraints. For instance, optical backbones use the WDM technology to take better advantage of the capacity of the optical fibers often already installed. This is achieved through the multiplexing of several wavelength channels onto the same fiber. In that case a resource allocation is an optical channel, also called lightpath, which includes a path and wavelengths assigned to its links, one per link. If wavelength translation is performed in optical switching, then each channel may be assigned different wavelengths on the links of its path; otherwise the wavelength continuity imposes all the links to have the same wavelength. Of course, two lightpaths sharing a link must use different wavelengths on that link. The design can be done at the conception of the network (i.e. when conceiving a virtual network in MPLS where we have to establish virtual paths) or to adapt the network to changes (failures, new link, updates of routers, variation of traffic, ...). Finally there are various optimization criteria which differ according to the point of view: for a network user they are related to his/her satisfaction (minimizing delays, increasing available bandwidth, ...), while for a network operator, economics criteria like minimizing deployment and operating costs are more important.

This very wide topic is addressed by a lot of academic and industrial teams in the world. Our approach is to attack these problems with tools from Discrete Mathematics.

#### 6.1.1.1. Traffic Grooming

In a WDM network, routing a connection request consists in assigning to this request a route in the physical network and a wavelength. When each request uses at most $1/C$ of the bandwidth of the wavelength, we say that the grooming factor is $C$. It means that on a given link of the network we can groom at most $C$ requests on the same wavelength. Under this constraint the objective can be either to minimize the number of wavelengths (related to the transmission cost) or to minimize the number of Add/Drop Multiplexers (ADM) used in the network (related to the cost of the nodes). During the last years, we have addressed this problem in various WDM network topologies with the goal of minimizing the total number of required ADMs.

This year, we considered the minimization of the number of ADMs in optical WDM bidirectional rings, considering symmetric shortest path routing and all-to-all unitary requests [24]. We formulate the problem in terms of graph decompositions, and state a general lower bound for all the values of the grooming factor $C$ and $N$, the size of the ring. We have studied exhaustively the cases $C = 1$, $C = 2$, and $C = 3$, providing improved lower bounds, optimal constructions for several infinite families, as well as asymptotically optimal constructions and approximations. We have also studied the case $C > 3$, focusing specifically on the case $C = k(k + 1)/2$ for some $k \geq 1$. We have also proposed optimal decompositions for several congruence classes of $N$ using the existence of some combinatorial designs.
6.1.1.2. Routing Reconfiguration and its Links with Graph Searching

In production networks, traffic evolution, failures and maintenance operations force to adapt regularly the current configuration of the network (virtual topology, routing of connections). The routing reconfiguration problem in WDM networks is thus to schedule the migration of established lightpaths from current routing to a new pre-computed one while minimizing service disruptions. We have shown in the past the relations between this problem and the graph searching problem (see also Section 6.4.3).

This year, we have continued studying the tradeoffs between the total number and the number of simultaneous interruptions that occurs during the reconfiguration process, proving in particular that the knowledge of one parameter does not help to optimize the other [28], [15]. We have also started investigating the influence of physical layer impairment constraints on the reconfiguration problem [74]. More precisely, using a new wavelength in a fiber of a WDM network forces to tune or recalibrate all already used wavelengths. We thus model the cost of using a new wavelength with a linear function of the number of already used wavelengths. We have then studied the problem of minimizing the cost of the reconfiguration according to this function. We have shown that this optimization problem is already NP-complete in a two-node network. We have also obtained general bounds and characterized instances for which the problem can be solved in polynomial time. We have additionally proposed and evaluated heuristics.

6.1.1.3. Green Networking

The minimization of ICT (Information and Communications Technologies) energy consumption has become a priority with the recent increase of energy cost and the new sensibility of public, governments and corporations towards energy consumption. ICT alone is responsible of 2% to 10% (depending on the estimations) of the world power consumption. For example, it is estimated that switches, hubs, routers account for 6 TWh per year in the US.

Several studies exhibit that the traffic load of the routers only has a small influence on their energy consumption. Hence, the power consumption in networks is strongly related to the number of active network elements, such as interfaces, line cards, base chassis, etc. In [78], [15], we have defined and modeled formally the problem of finding a routing that minimizes the (weighted) number of active network elements. We have proved that this problem is not in APX, that is there is no polynomial-time constant-factor approximation algorithm to solve it. We have obtained general bounds for this problem, and bounds for particular topologies such as trees, grids, and cliques. We have also proposed a heuristic algorithm offering good performance on real topologies. Last, we have analyzed the impact of energy efficient routing on the stretch factor and on fault tolerance.

We have also studied potential energy savings in fixed broadband wireless networks [77], [61]. See Section 6.1.2.1 for more details.

6.1.1.4. Xcast6 Treemap Islands

IP multicast is a protocol that deals with group communications with the aim of reducing traffic redundancy in the network. However, due to difficulty in deployment and poor scalability with a large number of multicast groups, IP multicast is still not widely deployed nor used on the Internet. Recently, Xcast6 and Xcast6 Treemap, the two network layer multicast protocols, have been proposed with complementary scaling properties to IP multicast: they support a very large number of active multicast sessions. However, the key limitation of these protocols is that they only support small multicast groups. To overcome this limitation, we have proposed the Xcast6 Treemap Island [96], a hybrid model of Application Layer Multicast (ALM) and Xcast6 that can work for large multicast groups. Our model has several advantages: ease of deployment, efficiency in bandwidth savings, no control message between end-host and router, zero multicast forwarding state at router and no need for a multicast address allocation protocol. In addition, this model is a potential service from which an ISP (Internet Service Provider) can get new revenue. We have shown the feasibility of our model by simulation and comparison with IP multicast and NICE protocols.

6.1.1.5. Time-Dependent Graphs - Applications to Transport Networks

In [70], we focus on time-dependent graphs which seem to be a good way to model transport networks. In the first part, we remind some notations and techniques related to time-dependent graphs. In the second one, we
introduce new algorithms to take into account the notion of probability related to paths in order to guarantee travelling times with a certain accuracy. We also discuss different probabilistic models and show the links between them.

Other results on multi-interface networks were obtained outside of MASCOTTE [37], [36], [65], [63], [66], [52], [20].

6.1.2. Wireless Networks

MASCOTTE has conducted an intense research effort on wireless access networks. From the technological and architectural point of view, the field is broad, from mesh (or multi-hop cellular) networks to ad-hoc and sensor networks. Nevertheless, many questions and approaches are generic from an algorithmic and structural prospect. In particular, we have considered three of the most prominent performance metrics for radio networks. Using combinatorial optimization and centralized algorithmic with a network design flavor, fast data gathering, call scheduling, transport capacity and energy consumption of the networks have been studied. Our approach is complementary with those developed in other INRIA project-teams such as PLANETE, MAESTRO, SWING, or POPS. The complementarity has been exploited through a joint Ph.D. between MAESTRO and MASCOTTE [15], through an ANR VERSO project in which MAESTRO, MASCOTTE, and SWING are involved, and through regular collaborations with POPS. At the international level, we cooperate with some groups in renowned research centers such as CTI of Patras in Greece, RWTH Aachen in Germany, Universities of Roma or Salerno in Italy, the Technion Institute in Israel, SFU in Vancouver, Canada, UFC Universidade Federal do Ceará, Fortaleza, Brazil, or the University of Sao Paulo in Brazil. We studied a wide range of issues of wireless networks, from the design of efficient cross-layer medium access, call scheduling and routing techniques to energy efficient optimization. We developed theoretical tools for integrating dynamic characteristics of the networks in the optimization models, and analyzing and evaluating dynamic networks. Some graph coloring problems motivated by channel assignment in wireless networks are detailed in Section 6.3.

6.1.2.1. Wireless Backhaul

We have investigated network optimization problems related to the design and configuration of fixed wireless microwave backhaul - the portion of the network infrastructure that provides interconnectivity between the access and the core networks. Unlike wired networks, the capacity of a microwave radio link is prone to variations, either due to external factors (e.g., weather) or by the action of the network operator. This fundamental difference raises a variety of new issues to be addressed appropriately. We concentrated on conceiving reliable fixed broadband wireless networks under outage probability constraints [60], [59]. We have developed a joint optimization of data routing and bandwidth assignment that minimizes the total renewal fees of licenses, while handling all the traffic requirements simultaneously. We have proposed a chance-constrained mathematical program taking into account unreliable channel conditions. This approach remains one of the main challenges of modern stochastic programming and it is still considered as very difficult and widely intractable. We have derived integer linear programming (ILP) counterparts for these chance-constrained programs and propose cutset-based valid inequalities to enhance the performance of ILP solvers. Computational results illustrate the price of reliability and present a comparative study on the performance of the different formulations. Moreover, we have been interested in potential energy savings in fixed broadband wireless networks by selectively turning off idle communication devices in low-demand scenarios [77], [61]. We have proposed a mathematical formulation of the problem relying on a fixed-charge capacitated network design (FCCND) problem, which is very hard to optimize. We have derived from this modeling heuristic algorithms producing feasible solutions in a short time. This work was done in collaboration with the SME 3Roam, and partially developed within the scope of the joint project RAISOM (Réseaux de collecte IP sans fil optimisés).

6.1.2.2. Wireless Mesh Networks

We have addressed the problem of computing the transport capacity of Wireless Mesh Networks (WMNs) dedicated to Internet access [26]. Routing and transmission scheduling have a major impact on the capacity provided to the clients. A cross-layer optimization of these problems allows the routing to take into account
contentions due to radio interference. We have presented a generic Mixed Integer Linear Programming (MILP) addressing gateway placement, routing, and scheduling optimizations in a WMN. We have then derived new optimization models that can take into account a large variety of radio interference models, and QoS requirements on the routing. We also provide efficient resolution methods that deal with realistic size instances. It allows to work around the combinatoric of simultaneously achievable transmissions and point out a critical region in the network bounding the network achievable capacity. Based upon strong duality arguments, it is then possible to restrict the computation to a bounded area. It allows for computing solutions very efficiently on large networks. We have then extended our models to deal with the dynamic characteristics of the network [75]. We have proposed a new robust optimization model that considers traffic demand uncertainty, in order to compute an optimal robust routing and bandwidth allocation in WMNs. We have presented a linear program efficiently solved by column generation, and we have quantified the price of robustness, i.e. the additional cost to pay in order to obtain a feasible solution for the robust scheme.

We have additionally investigated on the feasibility of providing network connectivity to vehicles over a predefined trajectory (trains, metros, urban buses, etc.) [14]. The communication between the vehicle and the infrastructure network is based only on WiFi technology. The contributions of this work are two-fold: 1) the horizontal handover (between WiFi access points) and 2) the design and analysis of an infrastructure network (backbone network plus WiFi access network) deployed along the trajectory of the vehicle.

6.1.2.3. Data Gathering

We have studied algorithmic and complexity issues originating from the problem of data gathering in wireless networks [56]. We give an algorithm to construct minimum makespan transmission schedules for data gathering when the communication graph is a tree network, the interference range is any integer \( m \geq 2 \), and no buffering is allowed at intermediate nodes. In the interesting case in which all nodes have to deliver an arbitrary non-zero number of packets, we provide a closed formula for the makespan of the optimal gathering schedule. Additionally, we consider the problem of determining the computational complexity of data gathering in general graphs and show that the problem is weakly NP-complete. On the positive side, we design a simple \((1 + 2/m)\) factor approximation algorithm for general networks. We have also considered the data gathering process in multi-hop wireless sensor networks [76], [57]. Wireless sensors networks (WSNs) are deployed to collect huge amounts of data from the environment. This produced data has to be delivered through sensor’s wireless interface using multi-hop communications toward a sink. The position of the sink impacts the performance of the wireless sensor network regarding delay and energy consumption especially for relaying sensors. Optimizing the data gathering process in multi-hop wireless sensor networks is, therefore, a key issue. We have addressed the problem of data collection using mobile sinks in a WSN. We provide a multi-objective optimization framework that studies the trade-off between energy consumption and delay of data collection. This framework provides solutions that allow decision makers to optimally design the data gathering plan in wireless sensor networks with mobile sinks.

6.1.3. P2P Networks

6.1.3.1. Performance Analysis of Distributed Storage Systems

Distributed or peer-to-peer storage solutions rely on the introduction of redundant data to be fault-tolerant and to achieve high reliability. To ensure long-term fault tolerance, the storage system must have a self-repair service that continuously reconstructs lost fragments of redundancy. The speed of this reconstruction process is crucial for the data survival. In [93], we propose a new analytical framework, based on queuing models, to estimate the repair time and the probability of data loss. This model takes into account the correlation of concurrent repairs. The models and schemes proposed are validated by mathematical analysis, extensive set of simulations, and experimentation using the Grid’5000 test-bed platform. Recently, the Regenerating Codes were proposed as an improvement over classical replication and erasure codes to introduce redundancy. These codes make a better use of the available bandwidth when reconstructing the missing information. In [50], we propose a new code based on a hybrid approach, Double Coding, and compare it to existing codes from the point of view of availability, durability and storage space.
6.1.3.2. Well Balanced Designs for Data Placement

In collaboration with MAESTRO. The problem we consider in [88] is motivated by data placement, in particular, data replication in video on-demand systems. We are given a set \( V \) of \( n \) servers and \( b \) files (data, documents). Each file is replicated on exactly \( k \) servers. A placement consists in finding a family of \( b \) subsets of \( V \) (representing the files) called blocks each of size \( k \). Each server has some probability to fail and we want to find a placement which minimizes the variance of the number of available files. It was conjectured that there always exists an optimal placement (with variance better than that of any other placement for any value of the probability of failure). We show that the conjecture is true if there exists a well balanced design, that is a family of blocks, such that each \( j \)-element subset of \( V \), \( 1 \leq j \leq k \), belongs to the same or almost the same number of blocks (difference at most one). The existence of well balanced designs is a difficult problem as it contains as subproblem the existence of Steiner systems. We completely solve the case \( k = 2 \) and give bounds and constructions for \( k = 3 \) and some values of \( n \) and \( b \).

6.1.3.3. Peer-Assisted Time-shifted Streaming Systems: Design and Promises

Time-shifted streaming (or catch-up TV) allows viewers to watch their TV programs within an expanded time window. In [71], we emphasize the challenging characteristics of time-shifted TV systems that prevent known delivery systems to be used. We model time-shifted TV as multiple-interval graph, then we present a Peer-Assisted Catch-Up Streaming system, namely PACUS, where a set of end users’ computers assists the server for the content delivery. We show in particular how the PACUS tracker server can be efficiently implemented for catch-up TV. We demonstrate the benefits of PACUS by simulations. We especially highlight that PACUS reduces the traffic at the server side with the advantages of lightweight and self-adaptive unstructured peer-to-peer systems.

6.2. Simulation and Optimization Tools

Participants: Olivier Dalle, Luc Hogie, Aurélien Lancin, Emilio Mancini, Juan-Carlos Maureira, Philippe Mussi, Van Dan Nguyen, Judicaël Ribault, Issam Tahiri.

The works related to simulation and optimization tools address two kinds of issues: issues related to the development of the tools and their associated methodology, and issues related to the use of these tools in order to investigate a particular problem or assess the performances or properties of a particular system.

Since 2005, MASCOTTE has been developing a discrete event simulation architecture, named OSA, whose aim is to investigate how new software engineering techniques, such as component-based frameworks or Aspect Oriented Programming can help improving the simulation methodology, especially in terms of software reuse [16], [46], [47]. After six years of research development, OSA entered in the process of being diffused in 2011. This process is supported by a two-year INRIA “Development Action” (ADT) funding. This first year was devoted to cleaning the code base and produce a public release with a significant effort placed on user documentation and tutorial (cf http://osa.inria.fr/).

Aside our efforts on the OSA project, we are strongly involved in the USS-SimGrid ANR funded project, whose aim is at developing an efficient simulation platform geared at Grid Computing and very large scale distributed computing architectures. In this project, we worked on two tasks:

- Monitoring and characterization of the workload of large scale distributed applications [68];
- Support for modeling Peer-to-peer applications in the SimGrid simulator (originally designed for modeling grid-computing platforms).

We also pursued our involvement in the Discrete Event Systems Specification (DEVS) standardization effort [81], [82]. This formalism has reached a strong agreement amongst the community, but it still lacks implementation standard. Since OSA is aimed at providing better support for methodology, we consider necessary to support DEVS and participate to this effort. Our particular focus was on techniques and Architecture Description Languages (ADLs) for describing very large models of distributed applications [69].
Regarding our works on simulation studies and application-oriented developments, this year was the conclusion of our effort on the Internet on Rails project [14]. In this project, we studied and designed, both by means of simulations and experimentations, a low cost communication architecture based on IEEE 802.11 WiFi to provide high quality Internet access onboard high speed trains.

6.3. Graph Theory

Participants: Julio Araújo, Jean-Claude Bermond, Nathann Cohen, Frédéric Giroire, Frédéric Havet, František Kardoš, Ana Karolinna Maia, Dorian Mazauric, Remigiusz Modrzejewski, Leonardo Sampaio, Michel Syska.

MASCOTTE principally investigates applications in telecommunications via Graph Theory (see other objectives). However it also studies a number of theoretical problems of general interest. Our research mainly focused on graph coloring and some other problems arising from networks problems.

6.3.1. Graph Coloring

Coloring and edge-coloring are two central concepts in Graph Theory. There are many important and long-standing conjectures in these areas. We are trying to make advances towards such conjectures, in particular Steinberg’s conjecture, the List coloring Conjecture and the Acyclic Edge-Coloring Conjecture.

We are also interested in coloring problems arising from some practical problems: improper coloring, \( L(p,q) \)-labeling, directed star arboricity and good edge-labelling. The first two are both motivated by channel assignment and the last two by problems arising in WDM networks. For many practical problems are posed in a dynamic setting, we study on-line coloring and list coloring.

We also study some other variants of coloring like non-repetitive coloring or frugal coloring.

For all the coloring problems, we also consider the associated algorithmic problem, which consists in designing algorithms for finding the minimum number of colors of a coloring of a given graph. Algorithmic results on graph coloring are presented in Section 6.4.

The most classical notion of coloring (of edges or vertices) is the one of proper coloring, in which we insist on two adjacent elements to have distinct colors. However, it is usual to consider additional constraints, as well as relaxed constraints. For each variant of coloring, one can consider, its list version in which every element \( x \) is given a list \( L(x) \) of prescribed colors. A graph is said to be \( L \)-colorable if it has an \( L \)-coloring (fulfilling the constraints) such that \( x \in L(x) \) for all element \( x \). The choosability of a graph \( G \) is the smallest integer \( k \) for which \( G \) has an \( L \)-coloring whenever \( |L(x)| \geq k \) for all elements \( x \).

6.3.1.1. Coloring Graphs with Few Crossings

The famous Four Color Theorem states that every planar graph can be properly colored with 4 colors and Thomassen Five Color Theorem states that the choosability of every planar graph is at most 5. Hence, a natural question is to ask about the chromatic number and choosability of graphs with few crossings. In [38], we disprove a conjecture of Oporowski and Zhao stating that every graph with crossing number at most 5 and clique number at most 5 is 5-colorable. However, we show that every graph with crossing number at most 4 and clique number at most 5 is 5-colorable. We also show some colorability results on graphs that can be made planar by removing few edges. In particular, we show that if there exists three edges whose removal leaves the graph planar then it is 5-colorable. In [90], we show that every graph with two crossings is 5-choosable. We also prove that every graph which can be made planar by removing one edge is 5-choosable.

Another famous theorem on planar graphs is the one of Grötzsch, which says that every planar graph with no cycle of length 3 can be properly 3-colored. Steinberg’s Conjecture (1976) asserts that a graph with no cycles of length 4 or 5 is 3-colorable. Many approaches have been used towards this conjecture. We considered the following one in which, we relax the constraints on the color classes. Instead of insisting on them be independent sets, we allow them to induce a graph with some bounded degree. A graph \( G = (V,E) \) is said to be \((i,j,k)\)-colorable if its vertex set can be partitioned into three sets \( V_1, V_2, V_3 \) such that the graphs \( G[V_1], G[V_2], G[V_3] \) induced by the sets \( V_1, V_2, V_3 \) have maximum degree at most \( i, j, k \) respectively. Under
this terminology, Steinberg’s Conjecture says that every graph with no cycle of length 4 or 5 is $(0,0,0)$-colorable. In [91], we prove that every graph of $\mathcal{F}$ is $(2,1,0)$-colorable and $(4,0,0)$-colorable.

### 6.3.1.2. Acyclic, Linear and Frugal Colorings

A classical constraint added to a proper coloring is that at least three colors appears on each cycle, in which case we speak about acyclic coloring. In other words, the graph induced by the elements of any two color classes is a forest. The **acyclic chromatic index** of a graph $G$, denoted $\chi''_a(G)$ is the minimum $k$ such that $G$ admits an acyclic edge-coloring with $k$ colors. The famous Acyclic Edge-Coloring Conjecture asserts that $\chi''_a(G) = \Delta(G) + 2$, where $\Delta(G)$ is the maximum degree of the graph. In [21], we conjecture that if $G$ is planar and $\Delta(G)$ is large enough then $\chi''_a(G) = \Delta(G)$. We settle this conjecture for planar graphs with girth at least 5. We also show that $\chi''_a(G) \leq \Delta(G) + 12$ for all planar $G$.

Even stronger constraints are the following: a proper coloring of a graph is **2-frugal** (resp. **linear**) if the graph induced by the elements of any two color classes is of maximum degree 2 (resp. is a forest of paths). In [29], we improve some bounds on the 2-frugal choosability and linear choosability of graphs with small maximum average degree.

### 6.3.1.3. Coloring of Plane Graphs with Constraints on the Faces

We studied several variants of vertex and edge colorings of plane graphs insisting some constraints on the faces.

A face of a vertex colored plane graph is called **loose** if the number of colors used on its vertices is at least three. The **looseness** of a plane graph $G$ is the minimum $k$ such that any surjective $k$-coloring involves a loose face. In [35], we prove that the looseness of a connected plane graph $G$ equals the maximum number of vertex disjoint cycles in a dual graph $G^*$ increased by 2. We also show upper and lower bounds on the looseness of graphs based on the number of vertices, the edge connectivity, and the girth of the dual graph. These bounds improve the result of Negami for the looseness of plane triangulations. We also present infinite classes of graphs where the equalities are attained.

A vertex coloring of a 2-connected plane graph $G$ is a **strong parity vertex coloring** if for every face $f$ and each color $c$, the number of vertices incident with $f$ colored by $c$ is either zero or odd. Czap et al. [Discrete Math. 311 (2011) 512–520] proved that every 2-connected plane graph has a proper strong parity vertex coloring with at most 118 colors. In [34], we improve this upper bound for some classes of plane graphs.

A **facial parity edge coloring** of a connected bridgeless plane graph is such an edge coloring in which no two face-adjacent edges (consecutive edges of a facial walk of some face) receive the same color, in addition, for each face $f$ and each color $c$, either no edge or an odd number of edges incident with $f$ colored by $c$. From Vizing’s theorem it follows that every 3-connected plane graph has a such coloring with at most $\Delta^* + 1$ colors, where $\Delta^*$ is the size of the largest face. In [33] we prove that any connected bridgeless plane graph has a facial parity edge coloring with at most 92 colors.

A sequence $r_1, r_2, \ldots, r_{2n}$ such that $r_i = r_{n+i}$ for all $1 \leq i \leq n$, is called a **repetition**. A sequence $S$ is called **non-repetitive** if no block (i.e. subsequence of consecutive terms of $S$) is a repetition. Let $G$ be a graph whose edges are colored. A trail is called **non-repetitive** if the sequence of colors of its edges is non-repetitive. If $G$ is a plane graph, a **facial non-repetitive edge-coloring** of $G$ is an edge-coloring such that any **facial trail** (i.e. trail of consecutive edges on the boundary walk of a face) is non-repetitive. We denote $\pi_f^*(G)$ the minimum number of colors of a facial non-repetitive edge-coloring of $G$. In [41], we show that $\pi_f^*(G) \leq 8$ for any plane graph $G$. We also get better upper bounds for $\pi_f^*(G)$ in the cases when $G$ is a tree, a plane triangulation, a simple 3-connected plane graph, a hamiltonian plane graph, an outerplanar graph or a Halin graph. The bound 4 for trees is tight.

### 6.3.1.4. Improper Coloring

In [85] and [48], we study a coloring problem motivated by a practical frequency assignment problem and up to our best knowledge new. In wireless networks, a node interferes with the other nodes the level of interference depending on numerous parameters: distance between the nodes, geographical topography, obstacles, etc. We model this with a weighted graph $G$ where the weights on the edges represent the noise (interference) between
the two end-nodes. The total interference in a node is then the sum of all the noises of the nodes emitting on the same frequency. A weighted \( t \)-improper \( k \)-coloring of \( G \) is a \( k \)-coloring of the nodes of \( G \) (assignment of \( k \) frequencies) such that the interference at each node does not exceed some threshold \( t \). The Weighted Improper Coloring problem, that we consider here consists in determining the weighted \( t \)-improper chromatic number defined as the minimum integer \( k \) such that \( G \) admits a weighted \( t \)-improper \( k \)-coloring. We also consider the dual problem, denoted the Threshold Improper Coloring problem, where given a number \( k \) of colors (frequencies) we want to determine the minimum real \( t \) such that \( G \) admits a weighted \( t \)-improper \( k \)-coloring. We show that both problems are NP-hard and first present general upper bounds; in particular we show a generalization of Lovász’s Theorem for the weighted \( t \)-improper chromatic number. We then show how to transform an instance of the Threshold Improper Coloring problem into another equivalent one where the weights are either 1 or \( M \), for a sufficient big value \( M \). Motivated by the original application, we study a special interference model on various grids (square, triangular, hexagonal) where a node produces a noise of intensity 1 for its neighbors and a noise of intensity 1/2 for the nodes that are at distance 2. Consequently, the problem consists of determining the weighted \( t \)-improper chromatic number when \( G \) is the square of a grid and the weights of the edges are 1, if their end nodes are adjacent in the grid, and 1/2 otherwise. Finally, we model the problem using linear integer programming, propose and test heuristic and exact Branch-and-Bound algorithms on random cell-like graphs, namely the Poisson-Voronoi tessellations.

6.3.1.5. On-line Coloring

Several on-line algorithms producing colorings have been designed. The most basic and most widespread one is the greedy algorithm. The largest number of colors that can be given by the greedy algorithm on some graph, is called its Grundy number. Determining the Grundy number of a graph is NP-hard even for \( P_4 \)-free graphs, while it is polynomial-time solvable for \( P_4 \)-free graphs. In [19], we define a new class of graphs, namely the fat-extended \( P_4 \)-laden graphs, which intersects the class of \( P_5 \)-free graphs and strictly contains the one of \( P_4 \)-free. We show a polynomial-time algorithm to determine the Grundy number of such graphs. It implies that the Grundy number can be computed in polynomial time for most graph classes defined in terms of containing few \( P_4 \)-s: \( P_4 \)-reducible, extended \( P_4 \)-reducible, \( P_4 \)-sparse, extended \( P_4 \)-sparse, ...

In [94], we study a game version of greedy coloring. Given a graph \( G = (V, E) \), two players, Alice and Bob, alternate their turns in choosing uncolored vertices to be colored. Whenever an uncolored vertex is chosen, it is colored by the least positive integer not used by any of its colored neighbors. Alice’s goal is to minimize the total number of colors used in the game, and Bob’s goal is to maximize it. The game Grundy number of \( G \) is the number of colors used in the game when both players use optimal strategies. It is proved in this paper that the maximum game Grundy number of forests is 3, and the game Grundy number of any partial 2-tree is at most 7. We also gave some complexity results on \( b \)-colorings, which is a manner of improving colorings on-line [43].

6.3.1.6. Other Results on Graph Coloring

In [18], we aim at characterizing the class of graphs that admit a good edge-labelling. Such graphs are interesting, as they correspond to set of requests in UPP-digraphs (those in which there is at most one dipath from a vertex to another) for which the minimum number of wavelengths is equal to the maximum load. This implies that the problem can be solved efficiently. First, we exhibit infinite families of graphs for which no good edge-labelling can be found. We then show that deciding if a graph admits a good edge-labelling is NP-complete. Finally, we give large classes of graphs admitting a good edge-labelling: \( C_3 \)-free outerplanar graphs, planar graphs of girth at least 6, subcubic \( C_3 \), \( K_{2,3} \)-free graphs.

A wheel is a graph formed by a chordless cycle and a vertex that has at least three neighbors in the cycle. We prove in [83] that every 3-connected graph that does not contain a wheel as a subgraph is in fact minimally 3-connected. We prove that every graph that does not contain a wheel as a subgraph is 3-colorable. We were then told that this result was already proved by Thomassen, though with a different proof.

Gallai-Hasse-Roy-Vitaver Theorem states that every \( n \)-chromatic digraph contains a directed path of order \( n \). Let \( f(k) \) be the smallest integer such that every \( f(k) \)-chromatic digraph contains every oriented tree of order \( k \). Burr proved that \( f(k) \leq (k - 1)^2 \) and conjectured \( f(k) = 2n - 2 \). In [84], we give some sufficient conditions.
for an $n$-chromatic digraphs to contains some oriented tree. In particular, we show that every acyclic $n$-chromatic digraph contains every oriented tree of order $n$. We also show that $f(k) \leq k^2/2 - k/2 + 1$. Finally, we consider the existence of antidirected trees in digraphs. We prove that every antidirected tree of order $k$ is contained in every $(5k - 9)$-chromatic digraph. We conjecture that if $|E(D)| > (k - 2)|V(D)|$, then the digraph $D$ contains every antidirected tree of order $k$. This generalizes Burr’s conjecture for antidirected trees and the celebrated Erdős-Sós Conjecture. We give some evidences for our conjecture to be true.

6.3.2. Matchings and Independent Sets

Matchings and independent sets are important substructures which appears in many problems. In particular, color classes of vertex-colorings and edge-colorings are independent sets and matchings, respectively.

In [45], we show that every (sub)cubic $n$-vertex graph with sufficiently large girth has fractional chromatic number at most 2.2978 which implies that it contains an independent set of size at least 0.4352$n$. Our bound on the independence number is valid to random cubic graphs as well as it improves existing lower bounds on the maximum cut in cubic graphs with large girth.

In [39], we show that every cubic bridgeless graph $G$ has at least $2|V(G)|/3656$ perfect matchings. This confirms an old and celebrated conjecture of Lovász and Plummer in the 1970’s. This improves the first superlinear bound given in [40].

6.3.3. Hypergraphs

Hypergraphs, also called set systems, are a natural generalization of graphs. In a graph an edge is set of two vertices, while in a hypergraph an edge is a set of any size. It turns out to be an important notion in database theory. A digraph is a hypergraph contains every oriented tree of order $k$. This generalizes Burr’s conjecture for antidirected trees and the celebrated Erdős-Sós Conjecture. We give some evidences for our conjecture to be true.

6.3.4. Miscellaneous

6.3.4.1. Zagreb Indices

The first and second Zagreb indices of a graph are defined by $M_1 = \sum_{v \in V(G)} d(v)^2$ and $M_2 = \sum_{u,v \in E(G)} d(u)d(v)$, respectively. They are used in chemistry where it represents properties of molecules. In [17], we present some classes of graphs with prescribed degrees that satisfy $M_1/n \leq M_2/m$, where $M_1$ and $M_2$ are the first and second Zagreb indices. We also prove that for any $\Delta \geq 5$, there is an infinite family of graphs of maximum degree $\Delta$ such that the inequality is false. Moreover, we give alternative and slightly shorter proof of this inequality for trees and unicyclic graphs.

6.3.4.2. Induced Decomposition

An induced $H$-decomposition of a graph $G$ is a partition $(E_1, \cdots, E_k)$ of its edge set $E(G)$, such that the graph induced by each $E_i$, $1 \leq i \leq k$, is a copy of $H$. Bondy and Szwarcbiter asked for the maximum number $\text{ex}(n, H)$ of edges on a graph on $n$ vertices which admits an induced $H$-decomposition. In [13], we prove that for every non-empty graph $H$, $\text{ex}(n, H) = n(n - 1)/2 - o(n^2)$.

6.4. Algorithms

Participants: Julio Araújo, Janna Burman, Nathann Cohen, David Coudert, Gianlorenzo d’Angelo, Frédéric Giroire, František Kardoš, Dorian Mazauric, Nicolas Nisse, Stéphane Pérennes, Ronan Pardo Soares, Leonardo Sampaio.
MASCOTTE 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.

### 6.4.1. Coloring Graphs

Almost all graph coloring problems are NP-hard and most of them are even hard to approximate. Hence, to solve them efficiently, we aim at designing general exponential-time algorithms as well as polynomial-time algorithms for special classes. This is exemplified by the following results.

#### 6.4.1.1. $L(p,q)$-labeling

An $L(p,q)$-labeling of $G$ is an integer assignment $f$ to the vertex set $V(G)$ such that $|f(u) - f(v)| \geq p$, if $u$ and $v$ are adjacent, and $|f(u) - f(v)| \geq q$, if $u$ and $v$ have a common neighbor. Such a concept is a modeling of a simple channel assignment, in which the separation between channels depends on the distance. The goal is to find an $L(p,q)$-labeling of $G$ with minimum span (i.e. $\max \{f(u) - f(v), u, v \in V(G)\}$). It is well known that for all $k \geq 4$, deciding if a graph has an $L(p,1)$-labeling with minimum span $k$ is NP-complete. In [42], we present exact exponential time algorithms that are faster than existing ones.

#### 6.4.1.2. Counting and Enumerating Total and Edge Colorings

In [89], we are interested in computing the number of edge colorings and total colorings of a graph. We prove that the maximum number of $k$-edge-colorings of a $k$-regular graph on $n$ vertices is $k \cdot (k-1)^{n/2}$. Our proof is constructive and leads to a branching algorithm enumerating all the $k$-edge-colorings of a $k$-regular graph using a time $O^*(((k-1)^{n/2})$ and polynomial space. In particular, we obtain a polynomial time algorithm with an expected approximation ratio of $5171$. Again, our proof yields a branching algorithm to enumerate all the $4$-total-colorings of a connected cubic graph.

#### 6.4.1.3. Coloring Graphs of Special Classes

For some coloring problems that are known to be NP-hard for general graphs, we give some polynomial-time algorithms for the restriction to some graph classes. These graph classes are defined in terms of forbidden induced subgraphs. In [95], [79], we provide linear algorithms for coloring $P_3$-free graphs. In [58], we obtain polynomial time algorithms to determine the acyclic chromatic number, the star chromatic number and the harmonious chromatic number of $(q,q-4)$-graphs. Such graphs are those such that no set of at most $q$ vertices induces more than $q-4$ distinct $P_4$'s.

### 6.4.2. Complexity and Computation of Graph Parameters

We used graph theory to model various networks’ 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 where the problems are polynomial-time solvable.

#### 6.4.2.1. Path Vertex Cover

A subset $S$ of vertices of a graph $G$ is called a $k$-path vertex cover if every path of order $k$ in $G$ contains at least one vertex from $S$. The $k$-path vertex cover problem consists in finding such a set with minimum cardinality in $G$. In [25], it is shown that this problem is NP-complete for each $k \geq 2$ while it can be solved in linear-time in trees. The particular case of $k = 3$ is studied in [44], where an exact algorithm is given with running time $O^*(1.5171^n)$ in $n$-node graphs. In [44], we also design a polynomial time randomized approximation algorithm with an expected approximation ratio of $\frac{12}{11}$ for the minimum $3$-path vertex cover.
6.4.2. Convexity in Graphs

The geodesic convexity of graphs naturally extends the notion of convexity in euclidean metric spaces. A set $S$ of vertices of a graph $G = (V, E)$ is convex if any vertex on a shortest path between two vertices of $S$ also belongs to $S$. The convex hull of $S \subseteq V$ is the smallest convex set containing $S$. Finally, a hull set of a graph is a set of vertices the convex hull of it is $V$. The hull number of a graph $G$ is the minimum size of a hull set in $G$. In [86], [49], we prove that computing the hull number is NP-complete in bipartite graphs. We also provide bounds and design various polynomial-time algorithms for this problem in different graph classes as co-bipartite graphs, $P_4$-sparse graphs, etc.

6.4.2.3. Induced Subdivision in Digraphs

In [51], we consider the following problem for oriented graphs and digraphs: Given an oriented graph (digraph) $G$, does it contain an induced subdivision of a prescribed digraph $D$? The complexity of this problem depends on $D$ and on whether $H$ must be an oriented graph or is allowed to contain 2-cycles. We give a number of examples of polynomial instances as well as several NP-completeness proofs.

6.4.2.4. Circuits in Grids

A circuit in a simple undirected graph $G = (V, E)$ is a sequence of vertices \{ $v_1, v_2, \ldots, v_{k+1}$ \} such that $v_1 = v_{k+1}$ and \{$v_i, v_{i+1}$\} $\in E$ for $i = 1, \ldots, k$. A circuit $C$ is said to be edge-simple if no edge of $G$ is used twice in $C$. In [30], we study the following problem: which is the largest integer $k$ such that, given any subset of $k$ ordered vertices of an infinite square grid, there exists an edge-simple circuit visiting the $k$ vertices in the prescribed order? We prove that $k = 10$. To this end, we first provide a counterexample implying that $k < 11$. To show that $k \geq 10$, we introduce a methodology, based on the notion of core graph, to reduce drastically the number of possible vertex configurations, and then we test each one of the resulting configurations with an ILP solver.

6.4.3. Graph Searching, Cops and Robber Games

Pursuit-evasion encompasses a wide variety of combinatorial problems related to the capture of a fugitive residing in a network by a team of searchers. The goal consists in minimizing the number of searchers required to capture the fugitive in a network and in computing the corresponding capture strategy. This can also be viewed as cleaning the edges of a contaminated graph. We investigated several variants of these games.

6.4.3.1. Process Number and Routing Reconfiguration in WDM Networks

Graph searching, where the fugitive is arbitrary fast and moves simultaneously to the searchers, has been widely studied for its close relationship with graph decompositions. More recently, a variant of graph searching, namely the graph processing game, has been widely studied as a model for the routing reconfiguration in WDM networks (see Section 6.1.1.2). In [32], we give a linear time (resp., polynomial-time) algorithm to recognize graphs (resp., digraphs) with process number at most 2, along with a characterization in terms of forbidden minors, and a structural description. In [31], we give a polynomial (both in terms of time complexity and in the number of exchanged messages) distributed algorithm to compute the process number of trees. By slightly modifying the intial parameter of the algorithm, it also allows to compute various parameters of trees as pathwidth, search number, etc.

6.4.3.2. Cops and Robber Games

The “Cops and Robber” games are turn-by-turn games where a team of cops purchase a robber in a graph. We investigated two generalizations of the game introduced by Quilliot, Nowakoski and Winkler in 1983. We provided structural characterizations of graphs where one cop is sufficient to capture a fast fugitive able to hide [27]. In particular, one of these characterizations relies on hyperbolicity of the considered graph.

A surprising application of “Cops and Robber”-like games is the problem for a web-browser to download documents in advance while an internaut is surfing on the Web. In [92], we provide a modelling of the prefetching problem in terms of Cops and Robber games. The parameter to be optimized is then the download-speed necessary for the Internaut only accesses to already download webpages. This allows us to provide several complexity results and polynomial-time algorithms in some graph classes.
6.4.4. Distributed Algorithms

We investigated algorithmic problems arising in complex networks like the Internet or social networks. In this kind of networks, problems are becoming harder or impracticable because of the size and the dynamicity of these networks. One way to handle the dynamicity is to provide (distributed) fault tolerant algorithms. Studying the mobile agents paradigm seems to be a promising approach (somewhat related to Cops and Robber in Section 6.4.3) to address some models of distributed computing. We considered self-stabilizing algorithms for the gathering problem, and algorithms for updating routing tables.

Besides, the more an algorithm uses local information, the easier it is to update/correct the behaviour of the algorithm. In this direction, we investigated communication problems through game theory. We also studied the power of a communication model using only localized information, i.e., we study what can be computed using this communication model.

6.4.4.1. Mobile Agents and Self-stabilization

In [64], we consider a recent model of robot-based computing which makes use of identical, memoryless mobile robots placed on nodes of anonymous graphs. The robots operate in Look-Compute-Move cycles that are performed asynchronously for each robot. In particular, we consider the case of gathering robots on an anonymous ring. We provide a new distributed approach which turns out to be very interesting as it neither completely falls into symmetry-breaking nor into symmetry-preserving techniques.

We address dynamic large scale emerging networks, e.g., mobile sensor (agent) networks. The agents are resource limited and prone to failures. They move almost unpredictably and communicate in pairs. Population Protocol model is a communication model suited for such networks. We use a recently proposed version of this model where every agent is associated with a parameter called Cover Time. Cover Times abstract the interaction characteristics of mobile agents and allow the design of fast converging protocols and the evaluation of their convergence times (this is impossible in the original model). We take advantage of this model and perform first analytical analysis of a data collection protocol used in the ZebraNet project for the wild-life tracking of zebras. We propose alternative data collection protocols for ZebraNet and we analysis their time complexities [72], [53], [54]. To achieve fault-tolerance in population protocols, we develop a generic self-stabilizing transformer [22]. This is an automatic technique to convert a protocol to its self-stabilizing version.

In addition, we address important problems of coordination and synchronization. We present and prove correct two self-stabilizing deterministic protocols solving the classical mutual exclusion problem and the group mutual exclusion one [54].

6.4.4.2. Distributed Update of Routing Tables

In [62], we propose a simple and practical distributed algorithm for computing and updating routing tables for shortest path routing. This algorithm can be combined with every distance vector shortest paths routing algorithm, and allows to reduce the total number of messages sent. We give experimental evidence that it leads to an important gain in terms of the number of messages sent at the price of a little increase in terms of space occupancy per node.

Arc-Flags is a data structure used to speed-up the shortest paths computation in a graph. In [67], we introduce a new data structure, named Road-Signs, which allows us to efficiently update the Arc-Flags of a graph in a dynamic scenario. Road-Signs can be used to compute Arc-Flags, can be efficiently updated and do not require large space consumption for many real-world graphs.

6.4.4.3. Models of Distributed Computation

Since, we need to face both locality and dynamicity issues, we are developing new techniques allowing to obtain global structural information from local (partial) views of the network. In [55], [73], we have investigated the question of determining which graph properties can or cannot be computed using only local information. We consider the following model: each of the $n$ nodes of a graph which only knows its own ID and the IDs of its neighbours is allowed to send a message of $O(\log n)$ bits to some central entity, called the referee. We then investigate whether the referee is able to decide some basic structural properties of the
network topology \( G \) or not. We show that simple questions like, "does \( G \) contain a square?", "does \( G \) contain a triangle?" or "Is the diameter of \( G \) at most 3?" cannot be solved in general \cite{55}, \cite{73}. On the other hand, the referee can decode the messages in order to have full knowledge of \( G \) when \( G \) belongs to many graph classes such as planar graphs, bounded treewidth graphs and, more generally, bounded degeneracy graphs \cite{55}, \cite{73}. Following our framework, we are able to simulate asynchronicity of the network. In particular, we have exhibited a hierarchy of problems and distributed models of computation \cite{87}. 
6. New Results

6.1. Towards Data-Centric Networking


- Disruption Tolerant Networking

We designed an efficient message delivery framework, called MeDeHa, which enables communication in an internet connecting heterogeneous networks that is prone to disruptions in connectivity[24]. MeDeHa is complementary to the IRTF’s Bundle Architecture: besides its ability to store messages for unavailable destinations, MeDeHa can bridge the connectivity gap between infrastructure-based and multi-hop infrastructure-less networks. It benefits from network heterogeneity (e.g., nodes supporting more than one network and nodes having diverse resources) to improve message delivery. For example, in IEEE 802.11 networks, participating nodes may use both infrastructure- and ad-hoc modes to deliver data to otherwise unavailable destinations. It also employs opportunistic routing to support nodes with episodic connectivity. One of MeDeHa’s key features is that any MeDeHa node can relay data to any destination and can act as a gateway to make two networks inter-operate or to connect to the backbone network. The network is able to store data destined to temporarily unavailable nodes till the time of their expiry. This time period depends upon current storage availability as well as quality-of-service needs (e.g., delivery delay bounds) imposed by the application. We showcase MeDeHa’s ability to operate in environments consisting of a diverse set of interconnected networks and evaluate its performance through extensive simulations using a variety of scenarios with realistic synthetic and real mobility traces. Our results show significant improvement in average delivery ratio and a significant decrease in average delivery delay in the face of episodic connectivity. We also demonstrate that MeDeHa supports different levels of quality-of-service through traffic differentiation and message prioritization.

Then, we have extended the MeDeHa framework to support multihop mobile ad-hoc networks (or MANETs). Integrating MANETs to infrastructure-based networks (wired or wireless) allows network coverage to be extended to regions where infrastructure deployment is sparse or nonexistent as well as a way to cope with intermittent connectivity. Indeed, to date there are no comprehensive solutions that integrate MANETs to infrastructure-based networks. We have proposed a message delivery framework that is able to bridge together infrastructure-based and infrastructure-less networks. Through extensive simulations, we have demonstrated the benefits of the extended MeDeHa architecture especially in terms of the extended coverage it provides as well as its ability to cope with arbitrarily long-lived connectivity disruptions. Another important contribution of this work is to deploy and evaluate our message delivery framework on a real network testbed as well as conduct experiments in “hybrid” scenarios running partly on simulation and partly on real nodes [32].

Finally, we have proposed a naming scheme for heterogeneous networks composed of infrastructure-based and infrastructure-less networks where nodes may be subject to intermittent connectivity. The proposed scheme, called Henna, aims at decoupling object identification from location and is designed to operate with status-quo Internet routing. We evaluated the proposed naming scheme using the ns-3 network simulator and demonstrated that nodes were able to receive messages in both infrastructure-based and infrastructure-less networks despite frequent disconnections and changing location identifiers (i.e., IP address), while visiting different networks [31].

Another important contribution of this work is to deploy and evaluate our message delivery framework on a real network testbed as well as conduct experiments in “hybrid” scenarios running partly on simulation and partly on real nodes. This was demonstrated at the ACM Sigcomm conference in Toronto on August 2011 [74].
These different works are the result of collaborations with Katia Obraczka and Marc Mendonca from University of California Santa Cruz (UCSC) in the context of the COMMUNITY Associated Team, see URL http://inrg.cse.ucsc.edu/community/.

Another activity in the same domain relates to efficient scheduling and drop policies in DTNs. We remind that Delay Tolerant Networks are wireless networks where disconnections may occur frequently. In order to achieve data delivery in such challenging environments, researchers have proposed the use of store-carry-and-forward protocols: there, a node may store a message in its buffer and carry it along for long periods of time, until an appropriate forwarding opportunity arises. Multiple message replicas are often propagated to increase delivery probability. This combination of long-term storage and replication imposes a high storage and bandwidth overhead. Thus, efficient scheduling and drop policies are necessary to:

(i) decide on the order by which messages should be replicated when contact durations are limited, and
(ii) which messages should be discarded when nodes’ buffers operate close to their capacity.

We worked on an optimal scheduling and drop policy that can optimize different performance metrics, such as the average delivery rate and the average delivery delay. First, we derived an optimal policy using global knowledge about the network, then we introduced a distributed algorithm that collects statistics about network history and uses appropriate estimators for the global knowledge required by the optimal policy, in practice. At the end, we are able to associate to each message inside the network a utility value that can be calculated locally, and that allows to compare it to other messages upon scheduling and buffer congestion. Our solution called HBSD (History Based Scheduling and Drop) integrates methods to reduce the overhead of the history-collection plane and to adapt to network conditions. The first version of HBSD and the theory behind have been published in 2008. A recent paper [27] provides an extension to a heterogeneous mobility scenario in addition to refinements to the history collection algorithm. An implementation is proposed for the DTN2 architecture as an external router and experiments have been carried out by both real trace driven simulations and experiments over the SCORPION testbed at the University of California Santa Cruz. We refer to the web page of HBSD for more details http://planete.inria.fr/HBSD_DTN2/.

HBSD in its current version is for point-to-point communications. Another interesting schema is to consider one-to-many communications, where requesters for content express their interests to the network, which looks for the content on their behalf and delivers it back to them. We are working on this extension within a new framework called MobiTrade, which provides a utility driven trading system for efficient content dissemination on top of a disruption tolerant network. While simple tit-for-tat (TFT) mechanisms can force nodes to give one to get one, dealing with the inherent tendency of peers to take much but give back little, they can quickly lead to deadlocks when some (or most) of interesting content must be somehow fetched across the network. To resolve this, MobiTrade proposes a trading mechanism that allows a node (merchant) to buy, store, and carry content for other nodes (its clients) so that it can later trade it for content it is personally interested in. To exploit this extra degree of freedom, MobiTrade nodes continuously profile the type of content requested and the collaboration level of encountered devices. An appropriate utility function is then used to collect an optimal inventory that maximizes the expected value of stored content for future encounters, matched to the observed mobility patterns, interest patterns, and collaboration levels of encountered nodes. Using ns-3 simulations based on synthetic and real mobility traces, we show that MobiTrade achieves up to 2 times higher query success rates compared to other content dissemination schemes. Furthermore, we show that MobiTrade successfully isolates selfish devices. For further details on MobiTrade, we refer to [41] and to the web page of the project 1 where the code can be downloaded for both the ns-3 simulator and Android devices.

• Naming and Routing in Content Centric Networks

1 http://planete.inria.fr/MobiTrade/
Content distribution prevails in today's Internet and content-oriented networking proposes to access data directly by their content name instead of their location, changing so the way routing must be conceived. We worked a routing mechanism that faces the new challenge of interconnecting content-oriented networks. Our solution relies on a naming resolution infrastructure that provides the binding between the content name and the content networks that can provide it. Content-oriented messages are sent encapsulated in IP packets between the content-oriented networks. In order to allow scalability and policy management, as well as traffic popularity independence, binding requests are always transmitted to the content owner. The content owner can then dynamically learn the caches in the network and adapt its binding to leverage the cache use.

The work done so far is related to routing between content-oriented networks. We are starting an activity on how to provide routing inside a content network. To that aim, we are investigating on the one hand probabilistic routing and, on the other hand, deterministic routing and possible extension to Bellman-Ford techniques. In addition to routing, we are investigating the problem of congestion in content-oriented networks. Indeed, in this new paradigm, congestion must be controlled on a per-hop basis, as opposed to the end-to-end congestion control that prevails today. We think that we can combine routing and congestion control to optimize resource consumption. Finally, we are studying the implications of using CCN from an economical perspective. This activity was started in October 2011 by Damien Saucez.

- **Application-Level Forward Error Correction Codes (AL-FEC) and their Applications to Broadcast/Multicast Systems**

With the advent of broadcast/multicast systems (e.g., DVB-H/SH), large scale content broadcasting is becoming a key technology. This type of data distribution scheme largely relies on the use of Application Level Forward Error Correction codes (AL-FEC), not only to recover from erasures but also to improve the content broadcasting scheme itself (e.g., with FLUTE/ALC).

Our recent activities, in the context of the PhD of F. Mattoussi, included the design, analysis and improvement of GLDPC-Staircase codes, a "Generalized" extension to LDPC-Staircase codes. We have shown in particular that these codes: (1) offer small rate capabilities, i.e. can produce a large number of repair symbols 'on-the-fly', when needed; (2) feature high erasure recovery capabilities, close to that of ideal codes. Therefore they offer a nice opportunity to extend the field of application of existing LDPC-Staircase codes, while keeping backward compatibility (LDPC-Staircase "codewords" can be decoded with a GPLDPC-Staircase codec).

Our LDPC-Staircase codes, that offer a good balance in terms of performance, have been included as the primary AL-FEC solution for ISDB-Tmm (Integrated Services Digital Broadcasting, Terrestrial Mobile Multimedia), a Japanese standard for digital television (DTV) and digital radio. This is the first adoption of these codes in an international standard.

This success has been made possible, on the one hand, by major efforts in terms of standardization within IETF: the RFC 5170 (2008) defines the codes and their use in FLUTE/ALC, a protocol stack for massively scalable and reliable content delivery services, an active Internet-Draft published last year describes the use of these AL-FEC codes in FECFRAME, a framework for robust real-time streaming applications, and a recent Internet-Draft [66] defines the GOE (Generalized Object Encoding) extension of LDPC-Staircase codes for UEP (Unequal Erasure Protection) and file bundle protection services.

This success has also been made possible, on the other hand, by our efforts in terms of design and evaluation of two efficient software codecs of LDPC-Staircase codes. One of them is distributed in open-source, as part of our OpenFEC project (http://openc.com), a unique initiative that aims at promoting open and free AL-FEC solutions. The second one, a highly optimized version with improved decoding speed and reduced memory requirements, will be commercialized in 2012.
through an industrial partner. This codec proves that LDPC-Staircase codes can offer erasure recovery performances close to ideal codes in many circumstances while keeping decoding speeds over 1Gbps.

The fact that LDPC-Staircase codes have been preferred to a major AL-FEC competitor for the ISDB-Tmm standard, is the recognition of their intrinsic qualities and of an appropriate balance between several technical and non technical criteria.

- Unequal Erasure Protection (UEP) and File bundle protection through the GOE (Generalized Object Encoding) scheme

This activity has been initiated with the PostDoc work of Rodrigue IMAD. It focuses on Unequal Erasure Protection capabilities (UEP) (when a subset of an object has more importance than the remaining) and file bundle protection capabilities (e.g. when one want to globally protect a large set of small objects).

After an in-depth understanding of the well-known PET (Priority Encoding Technique) scheme, and the UOD for RaptorQ (Universal Object Delivery) initiative of Qualcomm, which is a realization of the PET approach, we have designed the GOE FEC Scheme (Generalized Object Encoding) alternative. The idea, simple, is to decouple the FEC protection from the natural object boundaries, and to apply an independent FEC encoding to each “generalized object”. The main difficulty is to find an appropriate signaling solution to synchronize the sender and receiver on the exact way FEC encoding is applied. In [65] we show this is feasible, while keeping a backward compatibility with receivers that do not support GOE FEC schemes. Two well known AL-FEC schemes have also been extended to support this new approach, with very minimal modifications, namely Reed-Solomon and LDPC-Staircase codes [66], [65].

During this work, we compared the GOE and UOD/PET schemes, both from an analytical point of view (we use an N-truncated negative binomial distribution to that purpose) and from an experimental, simulation based, point of view [67]. We have shown that the GOE approach, by the flexibility it offers, its simplicity, its backward compatibility and its good recovery capabilities (under finite of infinite length conditions), outperforms UOD/PET for practical realizations of UEP/file bundle protection systems. See also http://www.ietf.org/proceedings/81/slides/rmt-2.pdf.

- Application-Level Forward Error Correction Codes (AL-FEC) and their Applications to Robust Streaming Systems

AL-FEC codes are known to be useful to protect time-constrained flows. The goal of the IETF FECFRAME working group is to design a generic framework to enable various kinds of AL-FEC schemes to be integrated within RTP/UDP (or similar) data flows. Our contributions in the IETF context are three fold. First of all, we have contributed to the design and standardization of the FECFRAME framework, now published as a Standards Track RFC [68].

Secondly, we have proposed the use of Reed-Solomon codes (with and without RTP encapsulation of repair packets) and LDPC-Staircase codes within the FECFRAME framework: [59] [60] [61].

Finally, in parallel, we have started an implementation of the FECFRAME framework in order to gain an in-depth understanding of the system. Previous results showed the benefits of LDPC-Staircase codes when dealing with high bit-rate real-time flows.

A second type of activity, in the context of robust streaming systems, consisted in the analysis of the Tetrys approach, in [29]. Tetrys is a promising technique that features high reliability while being independent from RTT, and performs better than traditional block FEC techniques in a wide range of operational conditions.
- **A new File Delivery Application for Broadcast/Multicast Systems**
  
  FLUTE has long been the one and only official file delivery application on top of the ALC reliable multicast transport protocol. However FLUTE has several limitations (essentially because the object meta-data are transmitted independently of the objects themselves, in spite of their inter-dependency), features an intrinsic complexity, and is only available for ALC.

  Therefore, we started the design of FCAST, a simple, lightweight file transfer application, that works both on top of both ALC and NORM. This work is carried out as part of the IETF RMT Working Group, in collaboration with B. Adamson (NRL). This document has passed WG Last Call and is currently considered by IESG[56, 57, 58].

- **Security of the Broadcast/Multicast Systems**

  We believe that sooner or later, broadcasting systems will require security services. This is all the more true as heterogeneous broadcasting technologies will be used, for instance hybrid satellite-based and terrestrial networks, some of them being by nature open, as wireless networks (e.g., wimax, wifi). Therefore, one of the key security services is the authentication of the packet origin, and the packet integrity check. A key point is the ability for the terminal to perform these checks easily (the terminal often has limited processing and energy capabilities), while being tolerant to packet losses.

  The TESLA (Timed Efficient Stream Loss-tolerant Authentication) scheme fulfills these requirements. We are therefore standardizing the use of TESLA in the context of the ALC and NORM reliable multicast transport protocols, within the IETF MSEC working group. This document has been published as RFC 5776.

  In parallel, we have specified the use of simple authentication and integrity schemes (i.e., group MAC and digital signatures) in the context of the ALC and NORM protocols in [62, 63, 64]. This activity is also carried out within the IETF RMT working group.

- **High Performance Security Gateways for High Assurance Environments**

  This work focuses on very high performance security gateways, compatible with 10Gbps or higher IPsec tunneling throughput, while offering a high assurance thanks in particular to a clear red/black flow separation. In this context we have studied last year the feasibility of high-bandwidth, secure communications on generic machines equipped with the latest CPUs and General-Purpose Graphical Processing Units (GPGPU).

  The work carried out in 2011 has consisted in setting up and evaluating the high performance platform. This platform heavily relies on the Click modular TCP/IP protocol stack implementation, which turned out to be a key enabler both in terms of specialization of the stack and parallel processing. Our activities also consisted in analyzing the PMTU discovery aspect since it is a critical factor in achieving high bandwidths. To that goal we have designed a new approach for qualifying ICMP blackholes in the Internet, since PMTUD heavily relies on ICMP.

# 6.2 Network Security and Privacy

**Participants:** Sana Ben Hamida, Claude Castelluccia, Walid Dabbous, Mohamed Ali Kaafar, Arnaud Legout, Stevens Le Blond, Daniele Perito.

- **Online users tracking and profiling techniques**
Usernames are ubiquitously used for identification and authentication purposes on web services and the Internet at large, ranging from the local-part of email addresses to identifiers in social networks. Usernames are generally alphanumerical strings chosen by the users and, by design, are unique within the scope of a single organization or web service. In this work, we investigate the feasibility of using usernames to trace or link multiple profiles across services that belong to the same individual. The intuition is that the probability that two usernames refer to the same physical person strongly depends on the entropy of the username string itself. Our experiments, based on usernames gathered from real web services, show that a significant portion of the users’ profiles can be linked using their usernames. In collecting the data needed for our study, we also show that users tend to choose a small number of related usernames and use them across many services. This work is the first to consider usernames as a source of information when profiling users on the Internet. It has been published in PETS 2011 [47], one of the most prestigious conference in the area of Computer Privacy, and has been awarded the Andreas Pfitzmann award for the best contribution.

• **Online Privacy measurements and threats identification in online social networks**

In this work, we show how these seemingly harmless interests (e.g., Music Interests) can leak privacy-sensitive information about users. In particular, we infer their undisclosed (private) attributes using the public attributes of other users sharing similar interests. In order to compare user-defined interest names, we extract their semantics using an ontologized version of Wikipedia and measure their similarity by applying a statistical learning method. Besides self-declared interests in Music, our technique does not rely on any further information about users such as friends relationship or group belongings. Our experiments, based on more than 104K public profiles collected from Facebook and more than 2000 private profiles provided by volunteers, show that our inference technique efficiently predicts attributes that are very often hidden by users. To the best of our knowledge, this is the first time that user interests are used for profiling, and more generally, semantics-driven inference of private data is addressed. This work has been published in the prestigious Network & Distributed System Security Symposium (NDSS) 2012 [37].

• **Privacy Enhancing Technologies**

The increasing amount of personal and sensitive information disseminated over the Internet prompts commensurately growing privacy concerns. Digital data often lingers indefinitely and users lose its control. This motivates the desire to restrict content availability to an expiration time set by the data owner. This work presents and formalizes the notion of Ephemeral Publishing (EphPub), to prevent the access to expired content. We propose an efficient and robust protocol that builds on the Domain Name System (DNS) and its caching mechanism. With EphPub, sensitive content is published encrypted and the key material is distributed, in a steganographic manner, to randomly selected and independent resolvers. The availability of content is then limited by the evanescence of DNS cache entries. The EphPub protocol is transparent to existing applications, and does not rely on trusted hardware, centralized servers, or user proactive actions. We analyze its robustness and show that it incurs a negligible overhead on the DNS infrastructure. We also perform a large-scale study of the caching behavior of 900K open DNS resolvers. Finally, we propose an Android application, Firefox and Thunderbird extensions that provide ephemeral publishing capabilities, as well as a command-line tool to create ephemeral files. This work has been published in ICNP 2011 [36].

• **Differentially private smart metering**
Several countries throughout the world are planning to deploy smart meters in households in the very near future. The main motivation, for governments and electricity suppliers, is to be able to match consumption with generation. Traditional electrical meters only measure total consumption on a given period of time (i.e., one month or one year). As such, they do not provide accurate information of when the energy was consumed. Smart meters, instead, monitor and report consumption in intervals of few minutes. They allow the utility provider to monitor, almost in realtime, consumption and possibly adjust generation and prices according to the demand. Although smart metering might help improving energy management, it creates many new privacy problems. Smart meters provide very accurate consumption data to electricity providers. As the interval of data collected by smart meters decreases, the ability to disaggregate low-resolution data increases.

We developed a new privacy-preserving smart metering system. Our scheme is private under the differential privacy model and therefore provides strong and provable guarantees. With our scheme, an (electricity) supplier can periodically collect data from smart meters and derive aggregated statistics while learning only limited information about the activities of individual households. For example, a supplier cannot tell from a user’s trace when he watched TV or turned on heating. Our scheme is simple, efficient and practical. Processing cost is very limited: smart meters only have to add noise to their data and encrypt the results with an efficient stream cipher.

This work was presented at IH’11 (the Information Hiding Conference, 2011) [34].

- **Protecting against Physical Resource Monitoring**

This work considers the problem of resource monitoring. We consider the scenario where an adversary is physically monitoring on the resource access, such as the electricity line or gas pipeline, of a user in order to learn private information about his victim. Recent works, in the context of smart metering, have shown that a motivated adversary can basically profile a user or a family solely from his electricity traces. However, these works only consider the case of a semi-honest-but-non-intrusive adversary that is only trying to learn information from the consumption reports sent by the user. This work, instead, considers the much more challenging case of a intrusive semi-honest adversary, i.e. a semi-honest adversary that is in addition physically monitoring the resource by modifying the distribution network. We aim at answering to the following question: is it possible to design a resource distribution scheme that prevents resource monitoring and provides strong protection? We propose and analyze several possible solutions. The proposed solutions provide different privacy bounds and performance results. This work was presented at WPES’11 (ACM Workshop on Privacy in the Electronic Society) [35].

- **The Failure of Noise-Based Non-Continuous Audio Captchas**

CAPTCHAs, which are automated tests intended to distinguish humans from programs, are used on many web sites to prevent bot-based account creation and spam. To avoid imposing undue user friction, CAPTCHAs must be easy for humans and difficult for machines. However, the scientific basis for successful CAPTCHA design is still emerging. This project examines the widely used class of audio CAPTCHAs based on distorting non-continuous speech with certain classes of noise and demonstrates that virtually all current schemes, including ones from Microsoft, Yahoo, and eBay, are easily broken. More generally, we describe a set of fundamental techniques, packaged together in our Decaptcha system, that effectively defeat a wide class of audio CAPTCHAs based on non-continuous speech. Decaptcha’s performance on actual observed and synthetic CAPTCHAs indicates that such speech CAPTCHAs are inherently weak and, because of the importance of audio for various classes of users, alternative audio CAPTCHAs must be developed.

This work was presented at IEEE Security and Privacy 2011 [33].
BlueBear: Privacy in P2P systems

We have started a new project called bluebear on privacy threats in the Internet. Indeed, the Internet has never been designed with privacy in mind. For instance, the Internet is based on the IP protocol that exposes the IP address of a user to any other users it is communicating with. However, we believe that current users of the Internet do not realize how much they compromise their privacy by using the Internet. Indeed, the common wisdom is that there are so many users in the Internet that it is not feasible for an attacker, apart may be for national agencies, to globally compromise the privacy of a large fraction of users. Therefore, finding a specific user is like looking for a needle in a haystack. The goal of the bluebear project is to raise attention on privacy issues when using the Internet. In particular, we want to show that without any dedicated infrastructure, it is possible to globally compromise the privacy of Internet users. BitTorrent is arguably the most efficient peer-to-peer protocol for content replication. However, BitTorrent has not been designed with privacy in mind and its popularity could threaten the privacy of millions of users.

In a first study we showed that it is possible to continuously monitor from a single machine most BitTorrent users and to identify the content providers (also called initial seeds). We performed a very large monitoring operation continuously “spying” on most BitTorrent users of the Internet from a single machine and for a long period of time. During a period of 103 days, we collected 148 million IP addresses downloading 2 billion copies of contents. We then identified the IP address of the content providers for 70% of the BitTorrent contents we spied on. We showed that a few content providers inject most contents into BitTorrent and that those content providers are located in foreign data centres. We also showed that an adversary could compromise the privacy of any peer in BitTorrent and identify the big downloaders that we define as the peers who subscribe to a large number of contents. This is a major privacy threat as it is possible for anybody in the Internet to reconstruct all the download and upload history of most BitTorrent users. This work was published in LEET 2010.

To circumvent this kind of monitoring, BitTorrent users are increasingly using anonymizing networks such as TOR to hide their IP address from the tracker and, possibly, from other peers. We explored in a second study whose goal was to Exploit P2P Applications to Trace and Profile Tor Users, to which extent a P2P protocol such as BitTorrent, when not designed to protect users information, leak information that may compromise the identity of users. We quantified such an issue with BitTorrent on top of anonymizing networks. We also designed an attack that reveals the identity of Tor users (We showed that it is possible to retrieve the IP address for more than 70% of BitTorrent users on top of TOR). Moreover, once the IP address of a peer is retrieved, it is possible to link to the IP address other applications used by this peer on top of TOR [45].

The fact that it is hard for a person to map an IP address to an identity mitigates the impact of the privacy attacks we described. However, we show that we can exploit a peer-to-peer VoIP system to associate a social identity (name, email address, etc.) to an IP address [46]. This means that anybody can now find this mapping that was only known by ISPs or big companies (like Google and Facebook), but never communicated unless in case of a legal action. The privacy threat is thus very high because this mapping enables blackmail, social attacks, targeted phishing attacks, etc.

As a proof of concept, we show that it is possible to track VoIP users mobility and BitTorrent downloads [46] using Skype, one of the most popular VoIP system with more that 500 millions registered users.

All these works received a very large media coverage (see http://www-sop.inria.fr/members/Arnaud.Legout/Projects/bluebear.html).
6.3. Network measurement, modeling and understanding


The main objective of our work in this domain is a better monitoring of the Internet and a better control of its resources. We work on new measurement techniques that scale with the fast increase in Internet traffic and growth of its size. We propose solutions for a fast and accurate identification of Internet traffic based on packet size statistics and host profiles. Within the ECODE FP7 project, we work on a network-wide monitoring architecture that, given a measurement task to perform, tune the monitors inside the network optimally so as to maximize the accuracy of the measurement results while limiting the overhead resulting from collected traffic. Within the ANR CMON project, we work on monitoring the quality of the Internet access by end-to-end probes, and on the detection and troubleshooting of network problems by collaboration among end users.

Next, is a sketch of our main contributions in this area.

- **Internet traffic classification by means of packet level statistics**
  One of the most important challenges for network administrators is the identification of applications behind the Internet traffic. This identification serves for many purposes as in network security, traffic engineering and monitoring. The classical methods based on standard port numbers or deep packet inspection are unfortunately becoming less and less efficient because of encryption and the utilization of non standard ports. In this activity, we come up with an online iterative probabilistic method that identifies applications quickly and accurately by only using the size of packets. Our method associates a configurable confidence level to the port number carried in the transport header and is able to consider a variable number of packets at the beginning of a flow. By verification on real traces we observe that even in the case of no confidence in the port number, a very high accuracy can be obtained for well known applications after few packets were examined. In another work [39], we make a complete study about the inter-packet time to prove that it is also a valuable information for the classification of Internet traffic. We discuss how to isolate the noise due to the network conditions and extract the time generated by the application. We present a model to preprocess the inter-packet time and use the result as input to the learning process. We discuss an iterative approach for the online identification of the applications and we evaluate our method on two different real traces. The results show that the inter-packet time is an important parameter to classify Internet traffic.

  We pursued this activity further by accounting for the communication profiles of hosts for the purpose of a better traffic classification [39], [38], [40]. We use the packet size and the inter-packet time as the main features for the classification and we benefit from the traffic profile of the host (i.e. which application and how much) to refine the classification and decide in favor of this or that application. The host profile is then updated online based on the result of the classification of previous flows originated by or addressed to the same host. We evaluate our method on real traces using several applications. The results show that leveraging the traffic pattern of the host ameliorates the performance of statistical methods. They also prove the capacity of our solution to derive profiles for the traffic of Internet hosts and to identify the services they provide.

  For a more thorough study of the traffic classification problem by means of packet statistics and host profiles, we refer to the PhD dissertation of Mohamad Jaber who was the main contributor to this activity inside the EPI Planete.

- **Adaptive network-wide traffic monitoring**
  The remarkable growth of the Internet infrastructure and the increasing heterogeneity of applications and users’ behavior make more complex the manageability and monitoring of ISP networks and raises the cost of any new deployment. The main consequence of this trend is an inherent disagreement between existing monitoring solutions and the increasing needs of management applications.
In this context, we work on the design of an adaptive centralized architecture that provides visibility over the entire network through a network-wide cognitive monitoring system. Given a measurement task, the proposed system drives its own configuration, typically the packet and flow sampling rates in routers, in order to address the tradeoff between monitoring constraints (processing and memory cost, collected data) and measurement task requirements (accuracy, flexibility, scalability). We motivate our architecture with an accounting application: estimating the number of packets per flow, where the flow can be defined in different ways to satisfy different objectives (e.g., Domain-to-Domain traffic, all traffic originated from a domain, destined to a domain). The architecture and the algorithms behind it are explained in paper published in 2010 for the case of a proactive control and in [43] for the case of a reactive control. In [44] the architecture and its algorithms are specified to a flow counting application. In all these works, the performances of our architecture are being validated in typical scenarios over an experimental platform we developed for the purpose of the study. Our platform is called MonLab (Monitoring Lab) and is described with more details in the Section on produced softwares. For now, MonLab presents a new approach for the emulation of Internet traffic and for its monitoring across the different routers. It puts at the disposal of users a real traffic emulation service coupled to a set of libraries and tools capable of Cisco NetFlow data export and collection, the overall destined to run advanced applications for network-wide traffic monitoring and optimization. The activities in this direction are funded by the ECODE FP7 STREP project (Sep. 2008 - Dec. 2011). The dissertation of Imed Lassoued [21] provides an introduction to the field in addition to details on our contributions and the MonLab emulation platform.

- **Spectral analysis of packet sampled traffic**

In network measurement systems, packet sampling techniques are usually adopted to reduce the overall amount of data to collect and process. Being based on a subset of packets, they hence introduce estimation errors that have to be properly counteracted by a fine tuning of the sampling strategy and sophisticated inversion methods. This problem has been deeply investigated in the literature with particular attention to the statistical properties of packet sampling and the recovery of the original network measurements. Herein, we propose a novel approach to predict the energy of the sampling error on the real time traffic volume estimation, based on a spectral analysis in the frequency domain. We start by demonstrating that errors due to packet sampling can be modeled as an aliasing effect in the frequency domain. Then, we exploit this theoretical finding to derive closed-form expressions for the Signal-to-Noise Ratio (SNR), able to predict the distortion of traffic volume estimates over time. The accuracy of the proposed SNR metric is validated by means of real packet traces. The analysis and the expressions of the SNR that stemmed from are described in [26]. In [52], we adopt such a model to design a real-time algorithm, that sets the IPFIX counter export timers in order to grant, to each flow, a target estimation accuracy. The work within this direction has been partially supported by the FP7 ECODE project.

- **Monitoring the quality of the Internet access by end-to-end probes**

The detection of anomalous links and traffic is important to manage the state of the network. Existing techniques focus on detecting the anomalies but little attention has been devoted to quantify to which extent network anomaly affects the end user access link experience. We refer to this aspect as the local seriousness of the anomaly. In order to quantify the local seriousness of an anomaly, we consider the percentage of affected destinations, that we call the impact factor. In order to measure it, a host should monitor all possible routes to detect any variation in performance, but this is not practical in reality. In this activity, funded by the ANR CMON project, we work on finding estimates for the impact factor and the local seriousness of network anomalies through a limited set of measurements to random nodes we call landmarks.
We initially study the user access network to understand the typical features of its connectivity tree. Then, we define an unbiased estimator for the local seriousness of the anomaly and a framework to achieve three main results: (i) the computation of the minimum number of paths to monitor, so that the estimator can achieve a given significance level, (ii) the localization of the anomaly in terms of hop distance from the local user, and (iii) the optimal selection of landmarks. We are using real data to evaluate the local seriousness of the anomaly and to determine the sufficient number of landmarks to select randomly without knowing anything on the Internet topology. The localization mechanism leverages the study on the connectivity tree and the relationship between the impact factor and the minimum hop distance of an anomaly. Our first results show that the impact factor is indeed a meaningful metric to evaluate the quality of Internet access. The current work focuses on extending this solution towards a collaborative setting where different end users collaborate together by exchanging the results of their observations. The objective will be a better estimation of the impact factor by each of them and a finer localization of the origin of any network problem.

On the experimental side, we have implemented the solution in a tool called ACQUA, which stands for Application for Collaborative Estimation of QUality of Internet Access. We design an anomaly detection mechanism based on the histogram of delay measurements and the likelihood of observations. Then, we give to ACQUA a pipeline based software architecture, and we go deeply into experimentation inside and outside Planetlab. We show what the properties and usage of the algorithm are, focusing also on how this tool can help us to get information about the network anomalies detected. Later we extend the idea of Impact Factor Estimation (IFE) by using what we call Inverse IFE from Planetlab, where the computer of the user whose connectivity is tested has a completely passive role in the measurements procedure. We study its strong and weak points, and we show conditions under which Inverse IFE from Planetlab gives similar results to traditional IFE.

**Applied Internet Measurements**

The performance of several Internet applications often relies on the measurability of path similarity between different participants. In particular, the performance of content distribution networks mainly relies on the awareness of content sources topology information. It is commonly admitted nowadays that, in order to ensure either path redundancy or efficient content replication, topological similarities between sources is evaluated by exchanging raw traceroute data, and by a hop by hop comparison of the IP topology observed from the sources to the several hundred or thousands of destinations. In this project, based on real data we collected, we advocate that path similarity comparisons between different Internet entities can be much simplified using lossy coding techniques, such as Bloom filters, to exchange compressed topology information. The technique we introduce to evaluate path similarity enforces both scalability and data confidentiality while maintaining a high level of accuracy. In addition, we demonstrate that our technique is scalable as it requires a small amount of active probing and is not targets dependent. This work has been published in [25].

**Reliability of Geolocation Databases**

In this project, we question the reliability of geolocation databases, the most widely used technique for IP geolocation. It consists in building a database to keep the mapping between IP blocks and a geographic location. Several databases are available and are frequently used by many services and web sites in the Internet. Contrary to widespread belief, geolocation databases are far from being as reliable as they claim. We conduct a comparison of several current geolocation databases -both commercial and free- to have an insight of the limitations in their usability. First, the vast majority of entries in the databases refer only to a few popular countries (e.g., U.S.). This creates an imbalance.

2 http://planete.inria.fr/acqua/
in the representation of countries across the IP blocks of the databases. Second, these entries do not reflect the original allocation of IP blocks, nor BGP announcements. In addition, we quantify the accuracy of geolocation databases on a large European ISP based on ground truth information. This is the first study using a ground truth showing that the overly fine granularity of database entries makes their accuracy worse, not better. Geolocation databases can claim country-level accuracy, but certainly not city-level. This study has been published in CCR [28].

- **Impact of Live Streaming Traffic**

  Video streaming is the most popular traffic in the Internet and a strong case for content-centric networks. Therefore, it is fundamental to understand the network traffic characteristics of video streaming. In this work [49], we extensively studied the network traffic characteristics of YouTube and Netflix (the most popular video streaming traffic in the USA). We have shown that the traffic characteristics vastly depend on the type of browser, mobile application, and container (Flash, Silverlight, HTML5) used.

6.4. **Experimental Environment for Future Internet Architecture**

**Participants:** Walid Dabbous, Thierry Parmentelat, Baris Metin, Frédéric Urbani, Daniel Camara, Alina Quereilhac, Shafqat Ur-Rehman, Thierry Turletti, Julien Tribino.

- **SFA Federation of experimental testbeds**

  The OneLab2 project has come to its end in spring 2010. We are now involved in the NOVI (E.U. STREP) project, the F-Lab (French A.N.R.) project, and have the lead of the “Federation” WorkPackage of OpenLab (E.U. IP) project. Within these frameworks, we are codevelopping with Princeton University a reference implementation for the Testbed-Federation architecture known as SFA for Slice-based Federation Architecture. As a sequel of former activities we also keep a low-noise maintenance activity of the PlanetLab software, which has been running in particular on the PlanetLab global testbed since 2004, with an ad-hoc federated model in place between PlanetLab Central (hosted by Princeton University) and PlanetLab Europe (hosted at INRIA) since 2007.

  During 2011 we have focused on the maturation of the SFA codebase, with several objectives in mind. Firstly we have contributed to a major overhaul of the specification as defined essentially within the GENI (N.S.F.) Project, with participations from all over the world. These changes, that affected both the core API and the schema used to expose and manage resource specifications, aimed at reaching a mature level of interoperability between the PlanetLab world and the EmuLab a.k.a. ProtoGeni world that has its own implementation, and are now available in SFA-2.0 issued late 2011.

  Secondly, the SFA codebase has been redesigned to provide a more generic shelter that other testbeds can easily leverage in order to come up with their own SFA-compliant wrapper. This is perceived as a powerful means to foster further adoption of the architecture, and the Planète team has been instrumental in bringing two entirely different testbeds to the federation, namely Senslab - developed in other INRIA Project-teams - and FEDERICA, the outcome of another E.U.-funded Project. Along the same lines we are working, although more remotely, with NICTA in Australia that publishes the O.M.F. testbed for running wireless testbeds, and who are interested in adopting this federation paradigm.

  Finally, as part of the pure PlanetLab development, we have added a feature for running nodes in a 'reservable' mode, which breaks the usual best-effort PlanetLab model, but turned out very helpful both for making experiments possible, that needed a more reproducible behaviour of experiments, and also in a federation perspective, for closing the usage gap with, notably wireless testbeds, that typically have a reservable-only provisioning mechanisms.
- **Content Centric Networks Simulation**

  We worked this year on the extension of the DCE framework for ns-3 in order to run CCN implementation under the ns-3 simulator. DCE stands for Direct Code Execution, its goal is to execute unmodified C/C++ binaries under ns-3 network simulator. With this tool researchers and developers can use the same code to do simulation and real experiments. DCE operation principle is to catch the standard systems calls done by the real application in the experiment and to emulate them within the ns-3 virtual network topology. Concerning CCN we use the PARC implementation named CCNx which is a well working open source software reference implementation of Content Centric Network protocol. As promised by DCE this integration of CCNx requires no modification of its code, it requires ‘only’ working on adding the system calls used by CCN that are not already supported by DCE. The advantage of this approach is that the integration work of CCN advanced DCE and will be useful in others completely different experiments. Another great advantage is that every evolution of the CCNx implementation is very easy to integrate, all what is needed is to compile the new source code. The next steps will be naturally to use DCE/ns-3 to evaluation CCN protocols in specific scenarii, to improve the coverage of systems calls supported by DCE, and to improve the DCE scheduler to be more realistic and to take into account CPU time spent in router queues. This work is done in the context of the ANR CONNECT project.

- **ns-3 Module store**

  Bake is an integration tool which is used by software developers to automate the reproducible build of a number of projects which depend on each other and which might be developed, and hosted by unrelated parties. This software is being developed with the participation of the Planète group and is intended to be the automatic building tool adopted by the ns-3 project.

  The client version of Bake is already working and the Planète group had a significant participation in its development. The contributions were in the context the addition of new functionalities, bug fixing and in the development of the regression tests. We are now starting the development of the ns-3 modules repository, which is a web portal to store the meta-information of the available modules. In the present state we have already designed and implemented the portal data basis and the main interface. It is already possible to register new modules and browse among the already registered ones.

  The web portal has to be finished, notably the part that will create the xml file that will be used to feed the bake’s client. We also need to add new functionalities to the client part, to enable incremental build over partially deployed environments. As it is today, bake does not enable the user to add just one new module to an already deployed version of the ns-3 simulator. This work is done in the context of the ADT MobSim in collaboration with Hipercom and Swing Inria project-teams.

- **The ns-3 consortium**

  We have founded last year a consortium between INRIA and University of Washington. The goals of this consortium are to (1) provide a point of contact between industrial members and the ns-3 project, to enable them to provide suggestions and feedback about technical aspects, (2) guarantee maintenance of ns-3’s core, organize public events in relation to ns-3, such as users’ day and workshops and (3) provide a public face that is not directly a part of INRIA or NSF by managing the http://www.nsnam.org web site. This web site is now finalized. However, activities related to developing the consortium have slowed down during 2011 due to the leave of Mathieu Lacage. We plan to put more resources on this aspect in 2012.
• **Using Independent Simulators, Emulators, and Testbeds for Easy Experimentation**

Evaluating new network protocols, applications, and architectures uses many kinds of experimentation environments: simulators, emulators, testbeds, and sometimes, combinations of these. As the functionality and complexity of these tools increases, mastering and efficiently using each of them is becoming increasingly difficult.

We designed the preliminary prototype of the Network Experiment Programming Interface (NEPI) whose goal is to make easier the use of different experimentation environments, and switch among them easily. NEPI intends to make it possible to write a single script to control every aspect of a potentially mixed experiment, including a hierarchical network topology description, application-level setup, deployment, monitoring, trace setup, and trace collection. We showed how a single object model which encompasses every aspect of a typical experimentation workflow can be used to completely describe experiments to be run within very different experimentation environments.

The development of NEPI started in 2009 with the implementation of the core API, an address allocator, a routing table configurator, but also a prototype ns-3 backend driven by a simple graphical user interface based on QT. Last year, we validated and evolved the core API with the addition of a new backend based on linux network namespace containers and stabilized the existing ns-3 backend. This year we have enhanced the design of NEPI and provided experiment validation, distributed experiment control, and failure recovery functionalities. In particular, we enforced separation between experiment design and execution stages, with off-line experiment validation. We also introduced a hierarchical distributed monitoring scheme to control experiment execution. We implemented a stateless message-based communication scheme, and added failure recovery mechanisms to improve robustness. The NEPI approach has been validated by implementing support for three complementary environments: a physical testbed, a network emulator, and a network simulator. Furthermore, we showed with a concrete experiment use case, available online for reproduction, how easy it is with NEPI to integrate these environments for hybrid-experimentation [48].

• **Guidelines for the accurate design of empirical studies in wireless networks**

Traditionally, wireless protocol proposals have been often tested and validated using only analytical and simulation models [73]. However, as the wireless environment is very complex to model accurately, and since the cost of wireless cards has decreased in an exponential way, today more and more research papers include evaluation of new proposals using experimentation on real devices. Indeed, experimentation is a mandatory step before possible deployment of new network protocols with real users. However, wireless experimentation is much more complex to set up and run than simulation, and it is important to avoid many pitfalls that can occur during experimentation. The objectives of this work are twofold. First, we described typical problems currently encountered in wireless-based experimentation, and we present simple guidelines to avoid them [50]. Second, we proposed an experimental methodology where the detection of anomalies, calibration of the measurement setup, and clear definition of the scenario (among others) make easier the repeatability of results [55]. This work has been done in collaboration with Cristian Tala, Luciano Ahumada and Diego Dujovne from the Universidad Diego Portales of Chili, in the context of the WELCOME STIC AMSud 2011.

• **Multicast Video Streaming over WiFi Networks: Impact of Multipath Fading and Interference**
We conducted an experimental study in order to analyse the impact of interference, multipath fading and path loss on multicast video streaming (i.e., goodput, packet loss and ultimately on the video quality) using off-the-shelf fixed WiFi equipment in a wireless (802.11 b/g) local area network (WLAN) environment. We used the ricean K-factor as a measure of multipath fading, spectrum analyzer to estimate channel interference and received signal strength indicator (RSSI) as indication of signal power and attenuation. In order to realistically measure aforementioned metrics, we conducted extensive wireless experiments against six test cases representing common real-world situations using off-the-shelf wireless equipment.

We showed that interference has more impact on performance than multipath fading. Multipath fading can result in considerable performance degradation in environments where moving objects cause perturbation. On the contrary, channel interference is more frequent and more prominent cause of performance degradation in wireless networks because ISM 2.4 GHz band is increasingly being utilized in homes and work places. Being able to quantify the impact of multipath fading and interference is crucial in planning, troubleshooting, managing as well as benchmarking and optimizing wireless networks. This study has been published in MediaWin 2011 [51].

- **Making easier Experimentation**

Wireless experimentations are challenging to evaluate due to the high variability of the channel characteristics and its sensitivity to interferences.

Merging traces represents a complex problem especially in wireless experimentations, due to packet redundancy in multiple probes. Merging traces solutions need to be efficient in order to process the large amount of generated traces. These solutions should provide an output data structure that allows easy and fast analysis and must be scalable in order to be used in large and various experimental settings. We have designed an algorithm that performs trace synchronization and merging in a scalable way. The algorithm output is stored in a configured MYSQL database allowing for smart packet trace storage. This solution reduces processing time by 400% and storage space by 200% with regard to raw trace file solutions. It has been implemented in an open source software called CrunchXML, available under the GNU General Public License v2 at http://twiki-sop.inria.fr/twiki/bin/view/Projets/Planete/CrunchXML.

- **An Integration Framework for Network Experimentation**

Many different experimentation environments address complementary aspects of network protocol evaluation, but because of their disparities and complexities it is often hard to use them to reproduce the same experiment scenario.

Simulation is often used for the evaluation of new network protocols and architectures. In order to perform more realistic simulations, modern simulators such as ns-3 integrate more detailed models and even support direct execution of real protocol code. However, such complex models require more computational and memory requirements. We have studied the feasibility of a hybrid approach based on distributing a complex simulation scenario on several nodes in a grid network. We showed that by exploiting the real time operation of the ns-3 simulator, it is possible to map such complex scenarios on grid nodes. We also proposed a basic mapping algorithm to distribute a simulation scenario in several nodes [42].
4. New Results

4.1. Algorithms: Bandwidth Allocation in Optical Networks

Participants: Christine Fricker, Philippe Robert, James Roberts.

The development of dynamic optical switching is widely recognized as an essential requirement to meet anticipated growth in Internet traffic. Since September 2009, RAP has begun an investigation into the traffic management and performance evaluation issues that are particular to this technology. A first analysis of passive optical networks used for high speed Internet access has led to the proposal of an original dynamic bandwidth allocation algorithm and to an evaluation of its traffic capacity. Our activity on optical networking is carried out in collaboration with Orange Labs with whom we have had a research contract and are currently finalizing a new one. We have also established contacts with Alcatel-Lucent Bell Labs and had fruitful exchanges with Iraj Saniee and his team on their proposed time-domain wavelength interleaved networking architecture (TWIN).

We have also analyzed the traffic capacity of wavelength division multiplexing (WDM), passive optical networks (PONs) where user stations (optical network units) are equipped with tunable transmitters. For these systems users can use any of the multiple wavelengths to transmit their data but only within the limit determined by the number of transmitters they possess. A mean field approximation is investigated to estimate the capacity of a limited-gated multiserver polling system with a limit on the number of servers a given station can use simultaneously. The approximation provides an expression for the stability limit under very general assumptions about the traffic process and system configuration.

More generally, motivated by these next generation passive optical networks, a multi-server polling system has been studied where the number of servers that can attend to a queue simultaneously is limited. The stability condition is investigated for this model under quite general assumptions. The result is proved for unlimited service policies. The paper [1] presents a conjecture for the case of limited service policies and general service limits. A simulation study shows that the stability conditions may hold.

In 2011, we have worked on bandwidth allocation in meshed networks. A first study applies the TWIN architecture for a metropolitan area network but with an original medium access control (MAC) algorithm. This algorithm is inspired by our prior work on access networks and ensures an efficient and fair allocation of bandwidth to flows between network nodes. The paper [9] describes this network architecture and presents a performance evaluation using analytical models backed up by simulations.

The TWIN architecture is not extensible to a wide area for reasons of scalability and the excessive signalling delay between geographically distant nodes. We have therefore invented a new notion of a multipoint-to-multipoint lightpath that avoids these problems. A patent application relating to this invention has been submitted. This patent is owned by Orange following the terms of our contract with them. The second patent (that simply perfects the first invention) is jointly owned since the research was performed after the end of this contract. The submitted paper [13] describes the invention and its evaluation. A major advantage demonstrated in this paper is the energy saving achieved by the use of the proposed optical technology in place of electronic routers.

Ongoing research seeks to apply this type of networking solution to data centres, on one hand, and to geographically spread tier 1 Internet carrier networks, on the other. This work is performed in collaboration with Orange Labs and will be covered by a contract that is close to being finalized. We have also participated in the preparation of a European CELTIC project proposal that includes a work package dedicated to the development and experimentation of the network proposed in [13].

4.2. Algorithms: Content-Centric Networking

Participants: Mathieu Feuillet, Christine Fricker, Philippe Robert, James Roberts, Nada Sbihi.
RAP is participating in an ANR project named CONNECT which will contribute to the definition and evaluation of a new paradigm for the future Internet: a content-centric network (CCN) where, rather than interconnecting remote hosts like IP, the network directly manages the information objects that users publish, retrieve and exchange. CCN has been proposed by Van Jacobson and colleagues at the Palo Alto Research Center (PARC). In CCN, content is divided into packet-size chunks identified by a unique name with a particular hierarchical structure. The name and content can be cryptographically encoded and signed, providing a range of security levels. Packets in CCN carry names rather than addresses and this has a fundamental impact on the way the network works. Security concerns are addressed at the content level, relaxing requirements on hosts and the network. Users no longer need a universally known address, greatly facilitating management of mobility and intermittent connectivity. Content is supplied under receiver control, limiting scope for denial of service attacks and similar abuse. Since chunks are self-certifying, they can be freely replicated, facilitating caching and bringing significant bandwidth economies. CCN applies to both stored content and to content that is dynamically generated, as in a telephone conversation, for example. RAP is contributing to the design of CCN in two main areas:

- the design and evaluation of traffic controls recognizing that TCP is no longer applicable and queue management will require new, name-based criteria to ensure fairness and to realize service differentiation;
- the design and evaluation of replication and caching strategies that realize an optimal trade-off of expensive bandwidth for cheap memory.

The team will also contribute to the development of efficient forwarding strategies and investigate economic arguments that make CCN a viable replacement for IP.

The ANR project began in January 2011 and several task meetings have taken place. We have also held meetings with PARC establishing close cooperation with them and with some participants in the NSF project “Named Data Networking”. We also participated in the CCN Community meeting in Palo Alto where we presented our work on traffic control. A paper describing the proposed flow-aware approach and results of a performance evaluation has been accepted for the conference Infocom 2012 [15].

Work on the performance of caching in CCN is ongoing. We have investigated popularity distributions for various types of content and evaluated their impact on the memory bandwidth tradeoff to be realized by CCN.

### 4.3. Algorithms: Channel Access algorithms in wireless networks

**Participants:** Mathieu Feuillet, Philippe Robert.

This is a collaboration with Thomas Bonald (Telecom ParisTech) and Alexandre Proutière (Microsoft Research). In wireless networks, to share available bandwidth between users is necessary. The bandwidth can be divided in several channels (frequency division) or the users can share the whole bandwidth by transmitting in different time slots (time division). We are studying different algorithms that allow users of a wireless network to access the channel. Those algorithms must avoid collisions and use the available bandwidth in the most efficient way. More and more wireless networks are decentralized and those algorithms must be distributed. Moreover, in order to use bandwidth in an efficient way, it is necessary to take the network topology into account. Recent studies have shown that it is possible to use the available bandwidth in a distributed and efficient way without message passing.

We studied a simplified version of the 802.11 channel access algorithm: CSMA/CA (Carrier Sense Multiple Access With Collision Avoidance). We proved that this algorithm does not use the bandwidth in the most efficient way and we proposed in 2010 a modification of this algorithm that is efficient. This result has been extended to multi-channels networks in [8] and its extended version [12].

### 4.4. Scaling Methods: Fluid limits in wireless networks

**Participant:** Philippe Robert.
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. This is, in some sense, a generalization of processor-sharing policies. We have investigated the fluid limits of simple examples of star topologies when the function used is $\log$. We have shown that, under this scaling, some new phenomenon occurs, namely that a node may stabilize for some time at some very high level while the number of jobs of other nodes decreases at some fixed rate or remains finite, i.e. lives in the neighborhood of 0. An averaging phenomenon plays an important role for the return to equilibrium.

4.5. Algorithms: Distributed Hash Table

**Participants:** Mathieu Feuillet, Philippe Robert.

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. We modeled this problem thanks to standard queues in order to understand the behavior of a single file and the global dynamic of the system. 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 have been able to study different time scales of the system with an averaging phenomenon. An article is in preparation on this subject. A more sophisticated distributed model with mean field techniques is under investigation.

On the side of this project, we notably studied the distribution of hitting times of the classical Ehrenfest and Engset models by using martingale techniques, furthermore their asymptotic behavior has been analyzed when the size of the system increases to infinity [5].

4.6. Stochastic Modeling of Biological Networks

**Participants:** Emanuele Leoncini, Philippe Robert.

This is a collaboration with Vincent Fromion from INRA Jouy en Josas, which started on October 2010.

The goal is to propose a mathematical model of the production of proteins in prokaryotes. Proteins are biochemical compounds that play a key role in almost all the cell functions and are crucial for cell survival and for life in general. In bacteria the protein production system has to be capable to produce about 2500 different types of proteins in different proportions (from few dozens for the replication machinery to more than 100000 for certain key metabolic enzymes). Bacteria uses more than the 85% of their resources to the protein production, making it the most relevant process in these organisms. Moreover this production system must meet two opposing problems: on one side it must provide the minimal proteins quantities in order to ensure the smooth-running of the cell, on the other side it can not choose a “overproduction policy” for all the proteins, since this would impact the global performance of the system and of the bacterium itself.

Gene expression is intrinsically a stochastic process: gene activation/deactivation occurs by means the encounter of polymerase/repressor with the specific gene, moreover many molecules that take part in the protein production act at extremely low concentrations.

We have restated mathematically the classical model using Poisson point processes. This representation, well-known in the field of queueing networks but, as far as we know, new in the gene expression modeling, allowed us to weaken few hypothesis of the existing models, in particular the Poisson hypothesis, which is well-suited in some cases, but that, in some situations, is far from the biological reality as we consider for instance the protein assemblage.
The theoretical environment of Poisson point processes has led us to propose a new model of gene expression which captures on one side the main mechanisms of the gene expression and on the other side it tries to consider hypotheses that are more significant from a biological viewpoint. In particular we have modeled: gene activation/deactivation, mRNA production and degradation, ribosome attachment on mRNA, protein production and degradation.

We have shown how the probability distribution of the protein production and the protein lifetime may have a significant impact on the fluctuations of the number of proteins. We have obtained analytic formulas when the duration of protein assemblage and degradation follows a general probability distribution, i.e. without the Poisson hypothesis. We have used our model to compare the variances resulting by choosing different hypotheses for the probability distribution of the protein production and degradation, in particular we have hypothesize the protein assembly and degradation to be deterministic. The model has showed how, under the previous hypothesis, the variance on the number of proteins is bigger than the classical model with the Poisson hypothesis.

4.7. Stochastic networks: large bike sharing systems

Participant: Christine Fricker.

This is a collaboration with Nicolas Gast (EPFL) started in December 2010. Bike sharing systems were launched by numerous cities as a serious alternative in urban transportation, for example Velib (20 000 bikes, 1 500 stations). One of the major issues is the availability of the resources: bikes or free slots to return the bikes. These systems have become a hot topic in Operation Research but there are few studies on these stochastic networks. To our knowledge, no theoretical study of such bike sharing systems exists taking into account the limited capacity of the stations.

We modeled this system in a symmetrical case. Mean field limit theorems give the dynamic of a large system and the stationary behavior of a single station. Analytical results are obtained and convergence proved in the standard model via Lyapunov functions. It allows to find the best ratio of bikes per station and to measure the improvement of incentive mechanisms, as choosing among two stations for example, or redistribution of bikes by trucks. It is under investigation. Further results deal with heterogeneous systems. Our goal is to propose via a theoretical study and tests simple algorithms to improve the system behavior.

4.8. Stochastic networks: heterogeneity

Participants: Christine Fricker, Hanène Mohamed.

Mean field techniques applied to non-symmetrical systems are explored. It appears as a promising way to obtain analytical results on systems with clusters.

4.9. Stochastic Networks: Jackson Networks

Participant: Danielle Tibi.

Lyapunov functions and essential spectral radius of Jackson networks, joint work with I. Ignatiouk-Robert (University of Cergy-Pontoise). A family of explicit multiplicative Lyapunov functions is constructed for any stable Jackson network. Optimizing the multiplicative factor over this family provides an upper bound for the essential spectral radius of the associated Markov process. For some particular classes of Jackson networks, this upper bound coincides with a lower bound derived from large deviations arguments, thus providing the exact value of the essential spectral radius. The main example is given by Jackson networks with routing matrix having a tree structure (in the sense that for any node i, at most one other node can route its customers to i). The result also holds for other types of routing matrices (e.g. completely symmetrical), under some conditions over the different arrival and service rates.

4.10. Scaling Methods: Interaction of TCP Flows

Participant: Philippe Robert.
This is a collaboration with Carl Graham (CMAP, École Polytechnique). Mathematical modeling of data transmission in communication networks has been the subject of intense activity for some time now. For data transmission, the Internet can be described as a very large distributed system with self-adaptive capabilities to the different congestion events that regularly occur at its numerous nodes. The coexistence of numerous connections in a network with a general number of nodes has been analyzed in a previous work through a mean-field limit of a Markovian model describing the interaction of several classes of permanent connections. In [6], this line of work has been generalized to the case when connections are not permanent but can be either active (ON) when it is transmitting data along its route, or idle (OFF). This year, the analysis of dynamic arrivals and departures has been investigated. The main technical problem is that mean-field asymptotics are not anymore usable. Instead, fluid limit schemes have to be considered in a quite delicate context, random measures.
6. New Results

6.1. Optimized protocols implementation and networking equipments

6.1.1. Locating Virtual Infrastructures: Users and InP Perspectives

Participants: Paulo Gonçalves, Guilherme Koslovski.

This is a joint work with Pascale Vicat-Blanc (Lyatiss) and Sébastien Soudan (Lyatiss).

The Cloud Computing wave consolidates the on-demand provisioning of configurable virtual machines. Recent projects have proposed the extension of the original IaaS paradigm to provide dynamic virtual networks to interconnect virtual IT resources, composing Virtual Infrastructures (VIs). In this new scenario, users with different objectives and expectations can rent dynamically provisioned virtual infrastructures to execute their applications during a given time slot. VIs can be allocated anywhere on top of a distributed and virtualized substrate. This decoupling from the geographical location introduces concerns such as a latency increase in network communications (user’s perspective), and the fragmentation of physical resources (Infrastructure Provider’s - InP - perspective). This context motivates efforts to investigate and deploy new models and tools which consider the geographical location of virtual infrastructures. Our work concentrates on the allocation of VIs guided by both the user’s and the InP’s constraints. We propose a formulation of the allocation problem considering the user’s expectations as well as the physical-substrate provider’s goals. Our initial experiments demonstrate that it is possible to improve the quality of the virtual-infrastructure allocation (user perspective) while simultaneously decreasing the physical substrate’s fragmentation and the substrate’s cost.

6.1.2. Energy-efficient reservation infrastructure for large-scale distributed systems

Participants: Anne-Cécile Orgerie, Laurent Lefèvre, Guérin-Lassous Isabelle.

Over the past few years, the energy consumption of Information and Communication Technologies (ICT) has become a major issue. Nowadays, ICT accounts for 2% of the global CO2 emissions, an amount similar to that produced by the aviation industry. Large-scale distributed systems (e.g. Grids, Clouds and high-performance networks) are often heavy electricity consumers because – for high-availability requirements – their resources are always powered on even when they are not in use. Reservation-based systems guarantee quality of service, allow for respect of user constraints and enable fine-grained resource management. For these reasons, in the context of Anne-Cecile Orgerie Phd (defended in September 2011), we proposed an energy-efficient reservation framework to reduce the electric consumption of distributed systems and dedicated networks. The framework, called ERIDIS, is adapted to three different systems: data centers and grids, cloud environments and dedicated wired networks. By validating each derived infrastructure, we show that significant amounts of energy can be saved using ERIDIS in current and future large-scale distributed systems [54].

6.1.3. Energy efficiency in exascale infrastructures

Participants: Mehdi Diouri, Olivier Gluck, Laurent Lefèvre.

Joint work with F. Cappello (JLPC, joint laboratory between INRIA and NCSA).

In Diouri’s PhD, we address the issue of energy efficiency for exascale supercomputers. We first proposed a green architecture for exascale systems gathering some new solutions to “consume less” energy and to “consume better”. This architecture involves interactions with the different actors interfering directly or indirectly with the supercomputer: its user, its administrator, its resource manager and the energy supplier. Then we were interested into leaning on this green architecture in order to propose some green services that will be offered for applications that will run on exascale systems. Our approach consists in evaluating the power overhead induced by some existing services, and by proposing a green version of these services that takes into account the constraints imposed by the different actors involved. In 2011, we specifically aimed to apply our approach for fault tolerance protocols in their normal functioning stage and in case of failure. [38]
6.1.4. Energy profiling and green leverages for high performance computing applications

**Participants:** Ghislain Landry Tsafack, Jean-Patrick Gelas, Laurent Lefevre.

Ghislain Landry TSAFACK CHETSA has started his PhD in January 2011, within the framework of INRIA HEMERA project, on: “Energy profiling and green leverages for high performance computing applications” (co-advisement with Jean-Marc PIERSON and Patricia STOLF from IRIT). During the course of this first year, we have investigated the possibility of characterizing distributed applications considering their energy/power profile. We first carried out a set of experiments for a better understanding of the application’s behavior and the impact that this behavior may have on its power consumption. Results led us to the assumption that any individual application run can be represented as a sequence of basic operations including computation, memory accesses, disk and network accesses over a given time period. We next relay on that assumption to define application’s energy profile. Application’s energy profile helps to prevent the fallout of any action that may be taken to reduce its power usage. To guarantee reasonable results, i.e., reduce energy with less performance degradation, we designed an energy prediction model capable of predicting power usage of a wide range of high performance computing (HPC) applications.

6.1.5. Towards virtualized home gateways

**Participants:** Jean-Patrick Gelas, Laurent Lefevre, Anne-Cecile Orgerie.

Joint work with Dino Lopez Pacheco (University of Nice) and Referi Assefa (Addis Ababa University, Ethiopia).

Virtualizing services located on end to end parts of the networks and making them available for a large number of applications and users is now becoming a real challenge. Within the scope of the GreenTouch project, we are exploring models, simulations tools (ECOFEN) and software prototypes able to demonstrate the impact of such approach in terms of energy reduction.

6.2. Quality of service and transport layer for future networks

6.2.1. On the Impact of the Flow-Size Distribution’s Tail Index on Network Performance with TCP Connections

**Participant:** Paulo Gonçalves.

This is a joint work with Oana Goga (UPMC, Lip6) and Patrick Loiseau (Eurecom).

In this work, we studied the impact of the flow-size distribution on network performance in the case of a single bottleneck with finite buffer. To tackle the case where flows are transmitted with the TCP protocol, we use real experiments and ns-2 simulations. Our preliminary results show that the distribution’s tail index impacts the performance in a more complex way than what is reported in existing literature. In particular, we exhibit situations where a heavier tail gives better performance for certain metrics. We argue that a main cause of our observed results is the transient behavior at the beginning of each flow.

6.2.2. Available Bandwidth Estimation for Multihop Wireless Networks

**Participants:** Isabelle Guérin Lassous, Van Nam Nguyen.

Estimating the available bandwidth in IEEE 802.11-based multi-hop wireless networks is a very difficult task due to the medium sharing among contending nodes and collisions between hidden stations. Several methods have been proposed so far for these networks to compute the available bandwidth on wireless links. If some recent solutions such as ABE and IAB now take into account collisions and their impact on the mean backoff, none of them considers the packet retransmissions due to collisions, although these retransmissions have an impact on the available bandwidth. In this work, we have proposed a new available bandwidth estimation for multi-hop wireless networks called RABE (Retransmission-based Available Bandwidth). This method integrates the average number of retransmission attempts in the available bandwidth estimation, in addition to other relevant parameters like the idle periods durations and the collision probability. RABE has been evaluated...
by simulation and the obtained results show that RABE can achieve a mean error ratio of 17% in comparison with the real measurement. Furthermore RABE is at least two times more accurate than ABE and ten times more accurate than IAB.

6.2.3. On The Recovery Performance of Single- and Multipath OLSR in Wireless Multi-Hop Networks

Participants: Inès Doghri, Isabelle Guérin Lassous.

In this work, we study and improve the recovery properties of single- and multipath routing strategies when facing network failure situations. In particular, we focus our study on two MANET routing protocols: OLSR and its multipath extension MP-OLSR. In various wireless multi-hop network environments, especially in multiple chain topologies, we define and evaluate the latency introduced by these protocols to find a new path after a link failure. Theoretical estimations and simulation results show that, under dual chain-topologies, this latency can be too long and incompatible with the needs of loss and delay constrained applications. As the source nodes cannot detect link failures immediately because of the delay incurred by the well-known nature of link state protocols in general, and of OLSR Topology Control (TC) messages in particular, these nodes keep sending packets along broken paths. We thus study the inconsistencies between the actual network topology and the nodes’ own representation. After analyzing the consequences of this long latency, we seek to alleviate these problems with the introduction of adapted mechanisms. We propose three new different schemes and accordingly extend the original OLSR and MP-OLSR protocols in order to decrease the expected latency and improve the protocol performance. Simulation results show a steep decrease of the latency when using these new schemes in dual chain-topologies. We also discuss these results in terms of packet loss, end-to-end delay and overhead.

6.3. High Speed Network’s traffic metrology and statistical analysis

6.3.1. A long-range dependent model for network traffic with flow-scale correlations

Participant: Paulo Gonçalves.

This is a joint work with Patrick Loiseau (Eurecom) and Pascale Vicat-Blanc (Lyatiss).

For more than a decade, it has been observed that network traffic exhibits long-range dependence and many models have been proposed relating this property to heavy-tailed flow durations. However, none of these models consider correlations at flow scale. Such correlations exist and will become more prominent in the future Internet with the emergence of flow-aware control mechanisms correlating a flow’s transmission to its characteristics (size, duration, etc.). In our present work, we study the impact of the correlation between flow rates and durations on the long-range dependence of aggregate traffic. Our results extend those of existing models by showing that two possible regimes of long-range dependence exist at different time scales. The long-range dependence in each regime can be stronger or weaker than standard predictions, depending on the conditional statistics between the flow rates and durations. In the independent case, our proposed model consistently reduces to former approaches. The pertinence of our model is validated on real web traffic traces, and its ability to accurately explain the Hurst parameter is validated on both web traces and numerical simulations.

6.3.2. A recurrent solution of Ph/M/c/N-like and Ph/M/c-like queues

Participant: Thomas Begin.

This work has been accepted for publication by the Journal of Applied Probability [50] and was performed in collaboration with Pr. Brandwajn (UCSC).
We propose an efficient semi-numerical approach to compute the steady-state probability distribution for the number of requests at arbitrary and at arrival time instants in Ph/M/c-like systems in which the inter-arrival time distribution is represented by an acyclic set of memoryless phases. Our method is based on conditional probabilities and results in a simple computationally stable recurrence. It avoids the explicit manipulation of potentially large matrices and involves no iteration. Due to the use of conditional probabilities, it delays the onset of numerical issues related to floating-point underflow as the number of servers and/or phases increases. For generalized Coxian distributions, the computational complexity of the proposed approach grows linearly with the number of phases in the distribution.

6.3.3. A Markovian model based on SIR epidemic classification to reproduce the workload dynamics of a VoD server

**Participants:** Shubhabrata Roy, Thomas Begin, Paulo Gonçalves.

We have devised a Markovian model, based on the SIR epidemic classification, to reproduce the workload dynamics that can be observed on a VoD (Video on Demand) server. This model basically relies on the dynamic between three distinct populations (i.e., current watchers, past watchers and potential watchers). It also embeds events with very low probability but high impact on its overall behavior corresponding to the occurrence of a flash crowd or the the buzz effect on a VoD server. The steady-state solution to this model has shown that it exhibits a behavior qualitatively close to what can be expected from a real-life VoD server. We have also shown that the workload process as delivered this model satisfies a large deviation principle. Our future work aims at taking advantage of this information to devise a new scheme for allocating available resources in a VoD server.

6.3.4. A comparative study of existing MBAC using real network traces

**Participants:** Doreid Ammar, Thomas Begin, Isabelle Guérin-Lassous.

We have evaluated the respective performance of several MBACs (Measurement-based admission control) using a realistic framework in which the pattern of the background traffic follows experimental traces collected on real-life networks. This study has allowed to highlight the respective discrepancies between MBACs in terms of easiness to implement and attained performance. This work will now focus on the design of a new MBAC based on a iteratively learned model.

6.3.5. Graph Based Classification of Content and Users in BitTorrent

**Participants:** Paulo Gonçalves, Marina Sokol.

This is a joint work with Konstantin Avrachenkov (INRIA Maestro) and Arnaud Legout (INRIA Planete). P2P downloads still represent a large portion of today’s Internet traffic. More than 100 million users operate BitTorrent and generate more than 30% of the total Internet traffic. Recently, a significant research effort has been done to develop tools for automatic classification of Internet traffic by application. The purpose of our present work is to provide a framework for sub-classification of P2P traffic generated by the BitTorrent protocol. Unlike previous works, we cannot rely on packet level characteristics and on the standard supervised machine learning methods. The application of the standard supervised machine learning methods is based on the availability of a large set of parameters (packet size, packet inter-arrival time, etc.). Since P2P transfers are based on the same BitTorrent protocol we cannot use this set of parameters to classify P2P content and users. Instead we can make use of the bipartite user-content graph. This is a graph formed by two sets of nodes: the set of users (peers) and the set of contents (downloaded files). From this basic bipartite graph we also construct the user graph, where two users are connected if they download the same content, and the content graph, where two files are connected if they are both downloaded by at least one same user. The general intuition is that the users with similar interests download similar contents. This intuition can be rigorously formalized with the help of graph based semi-supervised learning approach.
6.3.6. Generalized Optimization Framework for Graph-based Semi-supervised Learning  
**Participants:** Paulo Gonçalves, Marina Sokol.

This is a joint work with Konstantin Avrachenkov (INRIA Maestro).  
We develop a generalized optimization framework for graph-based semi-supervised learning. The framework gives as particular cases the Standard Laplacian, Normalized Laplacian and PageRank based methods. We have also provided new probabilistic interpretation based on random walks and characterized the limiting behavior of the methods. The random walk based interpretation allows us to explain differences between the performances of methods with different smoothing kernels. It appears that the PageRank based method is robust with respect to the choice of the regularization parameter and the labelled data. We illustrate our theoretical results with two realistic datasets, characterizing different challenges: *Les Miserables* characters social network and Wikipedia hyper-link graph. The graph-based semi-supervised learning classifies the Wikipedia articles with very good precision and perfect recall employing only the information about the hyper-text links.

6.3.7. On the estimation of the large deviations spectrum  
**Participant:** Paulo Gonçalves.

This is a joint work with Julien Barral (Univ. Paris 13)  
We propose an estimation algorithm for large deviations spectra of measures and functions. The algorithm converges for natural examples of multifractals.

6.3.8. Adaptive Multiscale Complexity Analysis of Fetal Heart Rate  
**Participant:** Paulo Gonçalves.

This is a joint work with Patrice Abry (ENS Lyon, CNRS) and Muriel Doret (Hospice civils de Lyon, Univ. Lyon 1)  
*Per partum* fetal asphyxia is a major cause of neonatal morbidity and mortality. Fetal heart rate monitoring plays an important role in early detection of acidosis, an indicator for asphyxia. This problem is addressed in this paper by introducing a novel complexity analysis of fetal heart rate data, based on producing a collection of piecewise linear approximations of varying dimensions from which a measure of complexity is extracted. This procedure specifically accounts for the highly non-stationary context of labor by being adaptive and multiscale. Using a reference dataset, made of real *per partum* fetal heart rate data, collected *in situ* and carefully constituted by obstetricians, the behavior of the proposed approach is analyzed and illustrated. Its performance is evaluated in terms of the rate of correct acidosis detection versus the rate of false detection, as well as how early the detection is made. Computational cost is also discussed. The results are shown to be extremely promising and further potential uses of the tool are discussed.
SWING Team

5. New Results

5.1. Flexible Radio Node

Participants: Florin Hutu, Tanguy Risset, Jacques Verdier, Guillaume Villemaud, Cédric Levy-Bencheton.

This section summarizes the early results obtained from the research axis flexible radio nodes.

In [41], [75], a candidate architecture for LTE-Advanced receiver is proposed. Based on the combination of MIMO techniques and flexible spectrum access, LTE-Advanced terminals will require the increasing of the analog front-end complexity. To reduce the complexity of the analog front-end, an innovative architecture based on the merge between the double IQ and the code multiplexing structures is proposed. Simulation and measurement results show that, in a Gaussian case, the bit error rate is similar when using the proposed architecture and the state of the art front-end stack-up structure. A complexity evaluation study reveals significantly reduced power consumption of the proposed single front-end architecture.

The current generation of mobile terminals can communicate on multiple modes using several antennas. However, their energy consumption remains a critical parameter. In [58], [74], we explore the combination of multiple communication modes and MIMO as a possible way to reduce the energy consumption of both the terminals and the network. We propose a realistic energy model for the PHY layer of a MIMO and multi-mode terminal, taking into account the MAC layer behaviour. We show that the combination of MIMO and multi-mode provides a solution to reduce global energy consumption.

Software means programmable. Hence software defined radio means that the radio should now be programmable. We know what computer programming means, and we agree, up to a certain level, on how it should be done. But do we know what programming a radio means? Several questions are still open: what will an SDR platform look like in ten years? Will there exist software radio code? What will be the technical challenges and commercial issues behind this code? Programming is more precise than configuring or tuning, it implies a much greater level of freedom for the programmer. But it also means much cheaper implementations in many cases and in particular a re-use of the same hardware for different protocols (i.e. with different programs). This is, to our point of view [76], the main difficulty of software radio programming: reconfiguration and in particular dynamic reconfiguration. Dynamic (i.e. very fast) reconfiguration is now mandatory because some protocols, 3GPP-LTE (Third Generation Partnership Program Long Term Evolution) for instance, propose channel adapting for each frame, requiring a setting of the channel estimation parameter in a few milliseconds.

5.2. Agile radio resource sharing

Participants: Jean-Marie Gorce, Claire Goursaud, Katia Jaffrè-Runser, Nikolaï Lebedev, Guillaume Villemaud, Paul Ferrand, Philippe Mary.

This section presents our recent results concerning the realistic modeling of wireless links to develop realistic models and efficient simulations. This work include theoretical developments like symbol error outage modeling, but also some applications in the context of LTE multi-cells association, or opportunistic relaying in the context of wireless sensor networks. Other contributions about resource sharing are presented in next sections below, in the section ‘network optimization’ and the section ‘network coding’.

In [28], we addressed the problem of finding a tractable expression for the symbol error outage (SEO) in flat Nakagami-m fading and shadowing channels. We deal with M-ary phase shift keying (M-PSK) and quadrature amplitude modulation (M-QAM) which extends our previous results on BPSK signaling. We propose a new tight approximation of the symbol error probability (SEP) holding for M-PSK and M-QAM signals which is accurate over all signal to noise ratios (SNRs) of interest. We derive a new generic expression for the inverse SEP which facilitates the derivation of a tight approximation of the SEO in a lognormal shadowing environment.
In [44], we consider on-body BAN nodes transmitting information towards a common sink, in a star topology (Body Area Networks (BAN) offer amazing perspectives to instrument and support humans in many aspects of their lives). While this setup is usual in wireless networks, the high instability of the BAN radio channel and the proximity of the body make classical communication protocols inefficient. These networks are further constrained by the low transmission power required for both battery life and health concerns. Opportunistic cooperation techniques are of great interest in such environment to ensure reliable communications. In previous works, we studied simple opportunistic relaying schemes under independent BAN links, using a packet error outage criterion. In this paper, we introduce a more realistic case where shadowing variations around the body are now assumed strongly correlated. Generally speaking, there is a lack of definitive measurements and models for the shadowing correlation in multi-hop networks, while it can play a crucial role at the higher layers. Based on the measurement and simulation results of the French BANET project, we use the BAN context as an illustrative example to exhibit how shadowing correlations have a strong impact on relaying approaches performance.

Opportunistic networking aims at exploiting sporadic radio links to improve the connectivity of a multi-hop network and to foster data transmissions. Broadcast nature of the wireless channel is an important feature that can be exploited to improve transmissions by using several potential receivers. Opportunistic relaying is thus the first brick for opportunistic networking. However, the advantage of opportunistic relaying may be balanced by energy increase related to having simultaneous active receivers. In [32], we proposed a thorough analysis of opportunistic relaying efficiency under different realistic radio channel conditions. The study aims at finding the best trade-off between two objectives: energy and latency minimizations, under a hard reliability constraint. We derive an optimal bound, namely, the Pareto front of the related optimization problem, which offers a good insight into the benefits of opportunistic routings compared with classical multi-hop routing schemes. Meanwhile, the lower bound provides a framework to optimize the parameters in physical layer, MAC layer and routing layer from the viewpoint of cross layer during the design or planning phase of a network.

This work has been extended in In [70] for relay channels. The gain induced by using relay channels in a linear network under both a capacity constraint and a realistic energy model is evaluated. We express a general model based on a convex optimization problem, allowing us to use numerical tools to obtain similar results for outer and inner bounds to the capacity of the full and half duplex relay channel. We then further the study with more complex networks based on relay channels, especially networks formed by a linear chain of nodes. We describe the Pareto optimal solutions of the minimization problem for with respect to the consumed energy and latency in such a linear network. From the simple case of the linear multi-hop network, we study the gains when implementing a linear chain of relay channels and compare these results to the simpler multi-hop transmission. This work will be published in 2012 in IEEE WCNC.

In [82] we extended this formalisms derived for a linear network to a more general case: the problem of deriving fundamental trade-off bounds for wireless ad hoc networks when multiple performance criteria are of interest. It proposes a MultiObjective (MO) performance evaluation framework composed of a broadcast and interference-limited network model, a steady state performance metric derivation inspired by a discrete Markov chain formalism and formulates the associated MO optimization problem. Pareto optimal performance bounds between end-to-end delay and energy for a capacity-achieving network are given for the 1-relay and 2-relay networks and assessed through simulations.

5.3. Autonomous wireless networking

Participants: Isabelle Augé-Blum, Bernard Tourancheau, Fabrice Valois, Ibrahim Amadou, Cédric Chauvenet, Quentin Lampin, Alexandre Mouradian, Bilel Romdhani.

Designing protocols for large scale wireless sensors networks is a challenging issue, if realistic environments are considered. Finding a trade-off between energy consumption and delay, or capacity, is difficult. The most promising ideas rely on zero-protocol approaches and on virtual coordinates use. the special case of VANETs is presented in the next section.
In [64], we focus on Wireless Sensor Networks (WSNs) in a more realistic case than classical studies and previous works: we consider wireless sensor nodes having different transmission ranges according to the environment and/or to the wireless chipset. The main consequence of this heterogeneity is the existence of asymmetric links. Such links in a WSN degrade the performance of most protocols which have not been designed to support this heterogeneity and to deal with asymmetric links: so, mainly, these links are pruned. Under this assumption, we propose a routing protocol for data collection from sensors nodes to the sink node in heterogeneous WSNs. Our proposal detects and takes benefit from asymmetric links caused by this heterogeneity. Our proposal, denoted MURA, (1) provides a high delivery ratio, (2) reduces the number of duplicated packets and (3) reduces the number of hop counts by exploiting the asymmetric links.

Due to the efficiency and scalability of greedy routing in WSNs and the financial cost of GPS chips, Virtual Coordinate Systems (VCSs) for WSNs have been proposed. A category of VCSs is based on the hop-count from the sink, this scheme leads to many nodes having the same coordinate. The main advantage of this system is that the hops number of a packet from a source to the sink is known. Nevertheless, it does not allow to differentiate the nodes with the same hop-count. We propose in [87] a novel hop-count-based VCS which aims at classifying the nodes having the same hop-count depending on their connectivity and at differentiating nodes in a 2-hop neighborhood. Those properties make the coordinates, which also can be viewed as a local identifier, a very powerful metric which can be used in WSNs mechanisms.

Duty-cycled medium access protocols allow for long lasting autonomous networks by periodically putting nodes to sleep. However, this life expectancy improvement comes at the cost of a lesser network capacity and a poor adaptability to bursty traffic loads. Indeed, existing contention algorithms do not provide efficient algorithms to dynamically elect multiple senders per wake-up periods. In [84], the medium is divided in several logical channels (e.g., obtained by a time/frequency division of the communication medium) and we propose to allocate them dynamically among senders. For this purpose, we propose a joint contention/scheduling algorithm, named Extended Slot Selection (ESS), that schedules multiple sender/receiver pairs to available logical channels.

Energy-efficient communication protocol is a primary design goal for Wireless Sensor Networks (WSNs). Many efforts have been done to save energy: MAC with duty cycle, energy-aware routing protocols, data aggregation schemes, etc. Recently, beacon-less strategies have emerged as new direction to improve considerably the WSN lifetime. However, the main contributions are not suitable to real radio environments because of hole avoiding strategies based on either planarization or explicit neighbor solicitations. We propose in [34] PFMAC (Pizza-Forwarding Medium Access Control), which combines beacon-less geo-routing and energy efficient MAC protocol via a cross-layer design to save more energy with higher reliability. PFMAC supports radio interferences, asymmetric radio links, etc. PFMAC supports a greedy forwarding strategy and, a reactive and optimized neighborhood discovery at 2-hop to deal with holes. Intensive simulations are proposed to highlight the behavior and the performance of PFMAC compared to BOSS over BMAC.

To provide for reliability in Wireless Sensor Networks (WSNs), Medium Access Control (MAC) protocols must be adapted by mechanisms taking cross-layer approaches into account. In [51], [52], we describe AreaCast which is designed for enhancing reliability in WSNs. AreaCast is a MAC layer mechanism independent of the routing layer, but uses only local topological and routing information to provide a communication by area instead of a traditional, node-to-node communication (i.e., unicast). In AreaCast, a source node addresses a set of nodes: an explicit relay node chosen as the next hop by a given routing protocol, and k other implicit relay nodes. The neighboring nodes select themselves as implicit relays according to their location from the explicit relay node. This mechanism uses overhearing to take advantage of the inherent broadcast nature of wireless communications. Without changing the routing protocol, AreaCast is able to dynamically avoid a byzantine node or an unstable link, allowing to benefit from the inherent topological redundancy of densely deployed sensor networks. Simulation results show that AreaCast significantly improves the packet delivery rate while having a good reliability-energy consumption trade-off.

Improving the network lifetime is an important design criterion for wireless sensor networks especially if we want to use standard solution like IPv6. In [38], we propose a novel approach which applies source-coding on addresses in heterogeneous IPv6 Cluster-based wireless sensor network. We formulate the problem...
of maximizing the network lifetime when Slepian-wolf coding is applied on addresses in network composed of line-powered and battery-powered sensors. The numerical results show that a significant network lifetime improvement can be achieved (about 25% in typical scenario). In [36], we investigate the sinks mobility in IPv6-based wireless sensors networks and specially in the new IETF proposed protocol RPL (Routing Protocol for Low power and Lossy Networks). We also show that even the mobility of sinks is not an explicit design criteria, the use of mobile sinks improves the network lifetime.

5.4. Wireless networking in VANETs
Participants: Marco Fiore, Sandesh Uppoor.

VANETS (Vehicular Ad hoc Networks) represents a challenging context for designing new protocols as it offers new challenges related to the high dynamicity of the network. In cooperation with external researchers, we derived recent results on mobility modeling and data dissemination in VANETS. This work is a part of the work on ‘Autonomous wireless networking’, but dedicated specially for VANETs.

Simulation is the tool of choice for the large-scale performance evaluation of upcoming telecommunication networking paradigms that involve users aboard vehicles, such as next-generation cellular networks for vehicular access, pure vehicular ad hoc networks, and opportunistic disruption-tolerant networks. The single most distinguishing feature of vehicular networks simulation lies in the mobility of users, which is the result of the interaction of complex macroscopic and microscopic dynamics. Notwithstanding the improvements that vehicular mobility modeling has undergone during the past few years, no car traffic trace is available today that captures both macroscopic and microscopic behaviors of drivers over a large urban region, and does so with the level of detail required for networking research. In [66], we present a realistic synthetic dataset of the car traffic over a typical 24 hours in a 400-km² region around the city of Koln, in Germany. We outline how our mobility description improves today’s existing traces and show the potential impact that a comprehensive representation of vehicular mobility can have one the evaluation of networking technologies.

In [21], [30], we investigate data dissemination in vehicular networks. Content downloading in vehicular networks is a topic of increasing interest: services based upon it are expected to be hugely popular and investments are planned for wireless roadside infrastructure to support it. We focus on a content downloading system leveraging both infrastructure-to-vehicle and vehicle-to-vehicle communication. With the goal to maximize the system throughput, we formulate a max-flow problem that accounts for several practical aspects, including channel contention and the data transfer paradigm. Through our study, we identify the factors that have the largest impact on the performance and derive guidelines for the design of the vehicular network and of the roadside infrastructure supporting it.

In [45] We address cooperative caching in wireless networks, where the nodes may be mobile and exchange information in a peer-to-peer fashion. We consider both cases of nodes with large- and small-sized caches. For large-sized caches, we devise a strategy where nodes, independent of each other, decide whether to cache some content and for how long. In the case of small-sized caches, we aim to design a content replacement strategy that allows nodes to successfully store newly received information while maintaining the good performance of the content distribution system. Under both conditions, each node takes decisions according to its perception of what nearby users may store in their caches and with the aim of differentiating its own cache content from the other nodes’. The result is the creation of content diversity within the nodes neighborhood so that a requesting user likely finds the desired information nearby. We simulate our caching algorithms in different ad hoc network scenarios and compare them with other caching schemes, showing that our solution succeeds in creating the desired content diversity, thus leading to a resource-efficient information access.

Performance and reliability of content access in mobile networks is conditioned by the number and location of content replicas deployed at the network nodes. In [27], we design a practical, distributed solution to content replication that is suitable for dynamic environments and achieves load balancing. Simulation results show that our mechanism, which uses local measurements only, approximates well an optimal solution while being robust against network and demand dynamics. Also, our scheme outperforms alternative approaches in terms of both content access delay and access congestion.
5.5. Optimization in wireless networks


In the context of the common lab between Inria and Alcatel Lucent Bell Labs and the ANR Ecoscells project, we work on optimizing wireless networks performance. In one side, we work on distributed algorithms for optimal resource allocation and/or mobile-BS association. On the other side, we work on mesh wireless networks optimization.

Multi-cell processing, also called Coordinated Multiple Point (CoMP), is a promising distributed technique that uses neighbor cells’ antennas [48]. It is expected to be the part of next generation cellular standards such as LTE-A. Small cell networks in dense urban environments are limited by interferences and CoMP can strongly take advantage of this fact to improve cell-edge users’ throughput. The present study introduces a distributed criterion for mobiles to select their optimal set of Base Stations (BS) to perform CoMP, and evaluates the impact of this association on the fairness and the total cell throughput. For that, we use a known theoretical expression for the capacity outage probability of CoMP under Rayleigh fading and evaluate the goodputs of antennas associations. The proposed criterion is used in combination with fair resource allocation to perform a joint double-objective optimization of fairness and efficiency. In [48], [91], we provide the analysis of the downlink Coordinated Multiple Point (CoMP) used in conjunction with the basic MIMO. The CoMP is the joint multi-cell transmission from several BS to mobiles, coupled here to an open-loop MIMO technique that does not require the perfect channel state knowledge. We show by simulation, that even for $4 \times 4$ MIMO transmission, the CoMP can improve the spectral efficiency for some mobiles, depending on capacity outage requirements.

In [33], we considered downlink transmission in cellular networks where we target to reduce the energy consumption by switching off some base stations by such a way that the distribution of SINR remains unchanged. This is a mean of green networking in cellular networks in downlink consideration. This paper analyzes for line and plane cases, the gain in power consumption obtained after switching off base stations. By computations we observe that the more the operational cost the more the gain in power consumption.

In [47], we propose an autonomous radio resource allocation and optimization scheme that chooses the transmit power and precoding vector among codebooks for multiple antennas transmitters to improve spectral and power efficiency and provide user fairness. Network self-optimization is an essential feature for supporting the cell densification in future wireless cellular systems. The proposed self-optimization is inspired by Gibbs sampler. We show that it can be implemented in a distributed manner and nevertheless achieves system-wide optimization which improves network throughput, power utilization efficiency, and overall service fairness. In addition, we extend the work and include power pricing to parametrize and enhance energy efficiency further. Simulation results show that the proposed scheme can outperform today’s default modes of operation in network throughput, energy efficiency, and user fairness.

In [55], we focused on broadband wireless networks based on OFDMA resource management, such as LTE systems. We have investigated two optimization problems, one concerning a backhauling mesh infrastructure while the other is the allocation of modulation and coding, subcarriers and power to users in LTE. Considering a realistic SINR model of the physical layer with a fine tuned power control at each node, a linear programing model using column generation has been developed for computing power efficient schedules with high network capacity for wireless mesh backhauling networks. Correlation between capacity and energy consumption have been analyzed as well as the impact of physical layer parameters - SINR threshold and path-loss exponent. With these models, we highlight that there is no significant tradeoff between capacity and energy when the power consumption of idle nodes is important. We also show that both energy consumption and network capacity are very sensitive to the SINR threshold variation. Finally, simulation results show that compared to classic reuse schemes the proposed approach is able to pack more users into the same bandwidth, decreasing the probability of user outage.

In [62], we focus on broadband wireless mesh networks like 3GPP LTE-Advanced. This technology is a key enabler for next generation cellular networks which are about to increase by an order of magnitude
the capacity provided to users. Such an objective needs a significative densification of cells which requires an efficient backhauling infrastructure. In many urban areas as well as under-developed countries, wireless mesh networking is the only available solution. Besides, economical and environmental concerns require that the energy expenditure of such infrastructure is optimized. We propose a multi-objective analysis of the correlation between capacity and energy consumption of LTE-like wireless mesh networks. We provide a linear programing modeling using column generation for an efficient computation of the Pareto front between these objectives. Based on this model, we observe that there is actually no significant capacity against energy trade-off.

In [63], broadband wireless mesh networks based on OFDMA resource management are studied considering a realistic SINR model of the physical layer with a fine tuned power control at each node. A linear programing model using column generation leads to compute power efficient schedules with high network capacity. Correlation between capacity and energy consumption is analyzed as well as the impact of physical layer parameters - SINR threshold and path-loss exponent. We highlight that there is no significant tradeoff between capacity and energy when the power consumption of idle nodes is important. We also show that both energy consumption and network capacity are very sensitive to the SINR threshold variation.

5.6. Network coding in WSN

Participants: Jean-Marie Gorce, Cédric Lauradoux, Marco Fiore, Claire Goursaud, Marine Minier, Anya Apavatjruit, Yuanyuan Zhang, Wassim Znaidi.

Network coding associated with Fountain codes is a very efficient approach to increase the throughput of multi-hop networks. However severe outcomes are still expected, especially to develop robust and energy efficient approaches for transmitting data over a large scale networks. Network coding is also very promising for security issues as presented below.

Diversity is a powerful means to increase the transmission performance of wireless communications. For the case of fountain codes relaying, it has been shown previously that introducing diversity is also beneficial since it counteracts transmission losses on the channel. Instead of simply hop-by-hop forwarding information, each sensor node diversifies the information flow using XOR combinations of stored packets. This approach has shown to be efficient for random linear fountain codes. However, random linear codes exhibit high decoding complexity. In [19], we propose diversity increased relaying strategies for the more realistic and lower complexity Luby Transform code in a linear network. Results are provided herein for a linear network assuming uniform imperfect channel states.

In [29], the exact probability that a receiver obtains N linearly independent packets among K over N received packets is computed, when the sender/s use/s random linear network coding over a Galois Field of size q. Such condition maps to the receiver’s capability to decode the original information, and its mathematical characterization helps to design the coding so to guarantee the correctness of the transmission. The proposed formulation represents an improvement over the current upper bound for the decoding probability, and provides theoretical grounding to simulative results in the literature.

In [35], we focus on the proper use of fountain codes for the transmission of sporadic data in a wireless sensor network (WSN). Fountain codes offer great perspectives for the self-organization of WSNs: they self adapt to the channel error rate without any control data. When deploying fountain codes on a WSN, two problems arise. First, the size of the data transmitted by a sensor is small in comparison to the size considered traditionally with fountain codes. Second, the communications are done in an hop-by-hop fashion. It implies that the destination of the data can not acknowledge instantaneously its reception to the source. Therefore, the transmissions of useless packets for the destination can not be prevented. The flooding traffic has been evaluated as well through realistic simulations for three different relaying strategies where packets are lost due to both small scale fading and collisions for an unslotted IEEE 802.15.4 medium access layer.

Network coding has attracted the attention of many researchers in security and cryptography. We have investigated several aspects of network coding security. In [20], we propose efficient solutions to thwart pollution attacks in which an adversary injects false information into data flow. This work was further expanded
in [54] to find rational strategy to minimize the energy cost and the impact of the attack. We also came to the conclusion that dealing with pollution attacks was not enough as long as the acknowledgment messages are not also protected. The risk is to suffer from a flooding attack. This goes beyond the capabilities of cryptographic solutions and we investigate the security capabilities of multipath acknowledgment in [67].

5.7. Security


Security is an important issue for wireless networks, especially for wireless sensor networks facing an amazing increase of the number of nodes. We review in this section all contributions related to the security issue, some of them being strongly related with the PHY layer or the networking protocols. As it can be seen below, some results are strongly connected to the models and protocols derived in the other sections.

In [59], we provide the first independent analysis of the (2nd-round tweaked) 256-bit version of the SHA3 candidate SHAvite-3. By leveraging recently introduced cryptanalysis tools such as rebound attack or Super-Box cryptanalysis, we are able to derive chosen-related-salt distinguishing attacks on the compression function on up to 8 rounds (12 rounds in total) and free-start collisions on up to 7 rounds. In particular, our best results are obtained by carefully controlling the differences in the key schedule of the internal cipher. Most of our results have been implemented and verified experimentally.

In [50], we study a class of insider attacks called the terrorist fraud. This is a relay attack against distance bounding protocols where the prover conspires with an adversary to misrepresent the distance between himself and the verifier. In ideal situations, the adversary does not gain any knowledge about the prover’s long-term secret. This makes designing a distance bounding protocol resistant to such fraud tricky: the secrets of an honest prover must be protected, while those of a dishonest one should be disclosed as an incentive not to cheat. We demonstrate that using a secret-sharing scheme, possibly based on threshold cryptography, is well suited for thwarting the terrorist fraud. Although such an idea has been around since the work of Bussard and Bagga, this is the first time that secret-sharing and terrorist fraud have been systematically studied altogether.

In [40], we deal with the problem of radio jamming. Jamming is a major threat against wireless communications. In this paper, we evaluate the effect of jamming on an UWB link employing a PPM non-coherent receiver. We optimize the jammer parameters that are the central frequency and the bandwidth based on the metric of the signal-to-jamming ratio. The optimization depends on different system parameters such as the channel model and the integration time of the receiver.

In [23], we focus on the resiliency of wireless sensor network routing protocols against selective forwarding attacks by compromised nodes. Informally, resiliency should be understood as the capacity of the routing protocol to endure and mitigate the presence of a certain number of compromised nodes seeking to disturb the routing process. To provide for security when nodes may be compromised, cryptographic solutions must be completed by algorithmic solutions considering “beyond cryptography” approaches. After discussing the shortcomings of existing routing protocols against packet-dropping malicious nodes we describe some protocol behaviors enhancing routing resiliency under several combined routing attacks. We propose in this paper the behaviors enhancing the resiliency of routing protocols under several combined routing attacks.

5.8. Network simulation tools

Several works in 2011 have been using simulation results. Nevertheless, Swing members are strongly working on improving network simulation frameworks to provide realistic simulations. Several contributions to the simulation tools wiplan ans wsnets have been proposed.

Some contributions to WSnet concern BAN environments implementation [44] and network coding features [19], [81]. Different protocols have been also implemented for wireless sensor networks [34], [84], specifically in the context of our collaboration with Orange Labs, Grenoble.
The wiplan simulator has been developed at CITI for several years. It is based on a frequency domain ParFlow (MR-FDPF) implementation that represents a unique finite elements based method for estimating the radio propagation in complex environments. In the context of heterogeneous networks, femtocells are very promising. In order to properly simulate their behavior and their impact on the macrocell layer, it is necessary to be able to simulate the radio coverage of femtocells. Hence ParFlow is a possible deterministic model that can be used for such simulation. In [42], two implementations of ParFlow are presented: time domain and frequency domain. The performance are compared and the advantages/drawbacks of each model are investigated.

In [56], we propose to use finite difference propagation methods to evaluate the wide band properties of the fast fading. For this purpose we adapted the MR-FDPF propagation model to simulate large bandwidth by combining numerous narrow band simulations. The results are compared with a channel sounder measurement campaign covering a bandwidth of up to 70 MHz. It is verified that fading characteristics in wireless channels varies with frequency and the MR-FDPF method is capable for simulating this variation of fadings for wide band systems.

In [56], a new approach is proposed allowing extracting the fading statistics for indoor radio channels based on the electric field strength predicted with the MR-FDPF method. The performance of the proposed approach is verified both by simulations and measurements.

In [65], we propose a new hybrid modeling method for indoor-to-outdoor radio coverage prediction. The proposed method is a combination of a ray-optical channel modeling approach and the frequency domain ParFlow method. While the former is widely used for modeling outdoor propagation environments, the latter is computationally efficient and accurate for modeling indoor environments.

In [90], we propose to use finite difference propagation methods to evaluate the wide band properties of the fast fading. For this purpose we adapted the MR-FDPF propagation model to simulate large bandwidth by combining numerous narrow band simulations. The results are compared with a channel sounder measurement campaign covering a bandwidth of up to 70 MHz. It is verified that fading characteristics in wireless channels varies with frequency and the MR-FDPF method is capable for simulating this variation of fading for wide band systems.
6. New Results

6.1. Design and Performance Analysis of Wireless Networks

Participants: François Baccelli, Florence Bénédit, Bartłomiej Blaszczyszyn, Chung Shue Chen, Mir Omid Haji Mirsadeghi, Frédéric Morlot, Tien Viet Nguyen, Van Minh Nguyen.

This axis bears on the analysis and the design of wireless access communication networks. Our contributions are organized in terms of network classes: cellular networks, wireless LANs and MANETs, VANETs. We also have a section on generic results that regard more general wireless networks. We are interested both in macroscopic models, which are particularly important for economic planning and in models allowing the definition and the optimization of protocols. Our approach combines several tools, queueing theory, point processes, stochastic geometry, random graphs, distributed control algorithms, self organization protocols.

6.1.1. Cellular Networks

The activity on cellular networks has several complementary facets ranging from performance evaluation to protocol design. The work is mainly based on strong collaborations with Alcatel-Lucent and Orange Labs.

6.1.1.1. Effect of Opportunistic Scheduling on the Quality of Service Perceived by the Users in OFDMA Cellular Networks

Our objective in [20] is to analyze the impact of fading and opportunistic scheduling on the quality of service perceived by the users in an Orthogonal Frequency Division Multiple Access (OFDMA) cellular network. To this end, assuming Markovian arrivals and departures of customers that transmit some given data volumes, as well as some temporal channel variability (fading), we study the mean throughput that the network offers to users in the long run of the system. Explicit formulas are obtained in the case of allocation policies, which may or may-not take advantage of the fading, called respectively opportunistic and non-opportunistic. The main practical results of the present work are the following. Firstly we evaluate for the non-opportunistic allocation the degradation due to fading compared to Additive White Gaussian Noise (AWGN) (that is, a decrease of at least 13% of the throughput). Secondly, we evaluate the gain induced by the opportunistic allocation. In particular, when the traffic demand per cell exceeds some value (about 2 Mbits/s in our numerical example), the gain induced by opportunism compensates the degradation induced by fading compared to AWGN. Partial results were presented at ComNet in 2009 [62].

6.1.1.2. Impact of Shadowing on QoS in Cellular Networks

Shadowing is believed to degrade the quality of service (QoS) in wireless cellular networks. Assuming log-normal shadowing, and studying mobile’s path-loss with respect to the strongest (serving) base station (BS) and the corresponding interference factor (the ratio of the sum of the path-gains form interfering BS’s to the path-gain from the serving BS), which are two key ingredients of the analysis and design of the cellular networks, in [48] we discovered a more subtle reality. We observe, as commonly expected, that a strong variance of the shadowing increases the mean path-loss with respect to the serving BS, which in consequence, may compromise QoS. However, in some cases, an increase of the variance of the shadowing can significantly reduce the mean interference factor and, in consequence, improve some QoS metrics in interference limited systems, provided the handover policy selects the strongest BS as the serving one. We exemplify this phenomenon, similar to stochastic resonance, studying the blocking probability in regular, hexagonal networks in a semi-analytic manner, using a spatial version of the Erlang’s loss formula combined with Kaufman-Roberts algorithm. More detailed probabilistic analysis explains that increasing variance of the log-normal shadowing amplifies the ratio between the strongest signal and all other signals thus reducing the interference. The above observations might shed new light, in particular on the design of indoor communication scenarios. Partial results were presented at IFIP WMNC’2010 [63].
6.1.1.3. Self-Optimization of Radio Resources in Cellular Networks

In [65], we developed mathematical and algorithmic tools for the self-optimization of mobile cellular networks. Scalable algorithms which are based on local measurements and do not require heavy coordination among the wireless devices were proposed. We focused on the optimization of transmit power and of user association. The method is applicable to both joint and separate optimizations. The global utility minimized is linked to potential delay fairness. The distributed algorithm adaptively updates the system parameters and achieves global optimality by measuring SINR and interference. The algorithms are built on Gibbs’ sampler and offer a unified framework that can be easily used for different purposes.

In [32], we investigated the joint optimization of radio resources in heterogeneous cellular networks made of a juxtaposition of macro and small cells. We showed that within this context, it is essential to use algorithms able to simultaneously solve the problems of channel selection, user association and power control. In such networks, the unpredictability of the cell and user patterns also requires self-optimized schemes. We proposed a generalized solution which is based on Gibbs’ sampler. It can be implemented in a distributed way and nevertheless achieves minimal system-wide potential delay. Results show that it is effective in both throughput and energy efficiency.

In [35], we extended it to an autonomous radio resource allocation and optimization scheme that chooses the transmit power and precoding vector among codebooks for multiple antennas transmitters to improve spectral and power efficiency and provide user fairness. Network self-optimization is an essential feature for supporting the cell densification in future wireless cellular systems. Besides, we included power pricing to parametrize and to enhance the energy efficiency. Simulation results show that the proposed scheme can outperform today’s default modes of operation in network throughput, energy efficiency, and user fairness.

Three patents were filed under the INRIA/Alcatel–Lucent joint laboratory.

6.1.1.4. Best Signal Quality in a Wireless Network

In a wireless network composed of randomly scattered nodes, the characterization of the distribution of the best signal quality received from a group of nodes is of primary importance for many network design problems. The thesis of Van Minh Nguyen [7] developed a framework for analyzing this distributions using shot noise models for the interference field. The joint distribution of the interference and the maximum signal strength was identified. The best signal quality can be represented as a function of these two quantities. Particular practical scenarios were also analyzed where explicit expressions can be obtained.

6.1.1.5. Cellular Network Tomography

The Foschini-Miljanic’s [67] algorithm is used for power control in cellular networks when users require a fixed bit rate. It leads to an optimal choice of power by the users in a distributed way when such a solution exists. If the users are too greedy or too many, the network saturates, and it is not possible to provide the required bit rates. We have been working on the question of residual bandwidth estimation in [61]. The residual bandwidth of a user is defined as the rate that this user should have to saturate the network when all other users stick to their initial rate requirement and all users use power control. The aim is to determine the residual bandwidth of a given user by local measurements. We showed that by simply changing their SINR target slightly and by listening to the evolution of interference, users can locally inverse Foschini-Miljanic’s algorithm and compute their residual bandwidth.

6.1.1.6. Coverage in Cellular Networks

Cellular networks are usually modeled by placing the base stations according to a regular geometry such as a grid, with the mobile users scattered around the network either as a Poisson point process (i.e. uniform distribution) or deterministically. These models have been used extensively for cellular design and analysis but suffer from being both highly idealized and not very tractable. Thus, complex simulations are used to evaluate key metrics such as coverage probability for a specified target rate (equivalently, the outage probability) or average/sum rate. More tractable models have long been desirable. In a joint work with J. Andrews and R. Ganti [UT Austin, USA] [9] and [34], we developed general models for multi-cell signal-to-noise-plus-interference ratio (SINR) based on homogeneous Poisson point processes and derived the coverage probability
and rate. Under very general assumptions, the resulting expressions for the SINR cumulative distribution function involve quickly computable integrals, and in some important special cases of practical interest these integrals can be simplified to common integrals (e.g., the Q-function) or even to exact and quite simple closed-form expressions. We also derived the mean rate, and then the coverage gain (and mean rate loss) from static frequency reuse. We compared the coverage predictions obtained by this approach to the standard grid model and an actual base station deployment. We observed that the proposed model is pessimistic (a lower bound on coverage) whereas the grid model is optimistic. In addition to being more tractable, the proposed model may better capture the increasingly opportunistic and dense placement of base stations in urban cellular networks with highly variable coverage radii.

Cellular networks are in a major transition from a carefully planned set of large tower-mounted base-stations (BSs) to an irregular deployment of heterogeneous infrastructure elements that often additionally includes micro, pico, and femtocells, as well as distributed antennas. In a collaboration with H. Dhillon, J. Andrews and R. Ganti [UT Austin, USA] [66], we extended the approach of we developed a model for a downlink heterogeneous cellular network (HCN) consisting of K tiers of randomly located BSs, where each tier may differ in terms of average transmit power, supported data rate and BS density. Assuming a mobile user connects to the strongest candidate BS, the resulting Signal-to-Interference-plus-Noise-Ratio (SINR) is greater than 1 when in coverage, Rayleigh fading, we derived an expression for the probability of coverage (equivalently outage) over the entire network under both open and closed access. One interesting observation for interference-limited open access networks is that at a given SINR, adding more tiers and/or BSs neither increases nor decreases the probability of coverage or outage when all the tiers have the same SINR threshold.

6.1.2. Mobile Ad Hoc Networks

A MANET is made of mobile nodes which are at the same time terminals and routers, connected by wireless links, the union of which forms an arbitrary topology. The nodes are free to move randomly and organize themselves arbitrarily. Important issues in such a scenario are connectivity, medium access (MAC), routing and stability. This year, we worked on the analysis of MAC and routing protocols in multi-hop MANETS in collaboration with Paul Mühlethaler [INRIA HIPERCOM], and on a game theoretic view of Spatial Aloha in collaboration with E. Altman and M.K. Hanawal [INRIA MAESTRO] [68].

6.1.2.1. Improvement of CSMA/CA’s Spatial Reuse

The most popular medium access mechanism for such ad hoc networks is CSMA/CA with RTS/CTS. In CSMA-like mechanisms, spatial reuse is achieved by implementing energy based guard zones. In a new collaboration with Qualcomm ([26] and [14]), we considered the problem of simultaneously scheduling the maximum number of links that can achieve a given signal to interference ratio (SIR). Using tools from stochastic geometry, we studied and maximized the medium access probability of a typical link. Our contributions are two-fold: (i) We showed that a simple modification to the RTS/CTS mechanism, viz., changing the receiver yield decision from an energy-level guard zone to an SIR guard zone, leads to performance gains; and (ii) We showed that this combined with a simple modification to the transmit power level – setting it to be inversely proportional to the square root of the link gain – leads to significant improvements in network throughput. Further, this simple power-level choice is no worse than a factor of two away from optimal over the class of all "local" power level selection strategies for fading channels, and further is optimal in the non-fading case. The analysis relies on an extension of the Matérn hard core point process which allows us to quantify both these SIR guard zones and this power control mechanism.

6.1.2.2. Opportunistic versions of CSMA/CA

In collaboration with Gustavo de Veciana and Yuchul Kim [UT Austin, ECE] we studied the benefits of channel-aware (opportunistic) scheduling of transmissions in ad-hoc networks using CSMA/CA [36]. The key challenge in optimizing the performance of such systems is finding a good compromise among three interdependent quantities, the density and channel quality of the scheduled transmitters, and the resulting interference at receivers. We propose two new channel-aware slotted CSMA protocols: opportunistic CSMA (O-CSMA) and quantile-based CSMA (QT-CSMA) and develop stochastic geometric models allowing us to quantify their performance in terms of spatial reuse and spatial fairness. When properly optimized these
protocols offer substantial improvements in terms of both of these metrics relative to CSMA - particularly when the density of nodes is moderate to high. Moreover, we show that a simple version of QT-CSMA can achieve robust performance gains without requiring careful parameter optimization. The paper supports the case that the benefits associated with channel-aware scheduling in ad hoc networks, as in centralized base station scenarios, might far outweigh the associated overhead, and this can be done robustly using a QT-CSMA like protocol.

6.1.3. Cognitive Radio Networks

We wrote a survey [22] on the probabilistic framework which can be used to model and analyze cognitive radio networks using various classes of MAC protocols (including carrier sensing based multiple access schemes and Aloha schemes). For each model, analytical results were derived for important performance metrics. This leads to a quantification of the interplay between primary and secondary users in such networks.

6.1.4. Generic Wireless Networks

6.1.4.1. Power Control in Wireless Networks

In [10], in collaboration with N. Bambos, [Stanford] and N. Gast [EPFL], we formulated a delay-power control (DPC) scheme for wireless networking, which balances delay against transmitter power on each wireless link. The DPC scheme is scalable, as each link autonomously updates its power based on the interference observed at its receiver; no cross-link communication is required. It is shown that DPC converges to a unique equilibrium power and several key properties are established, concerning the nature of channel bandwidth sharing achieved by the links. The DPC scheme is contrasted to the well-known Foschini-Miljanic (FM) formulation for transmitter power control in wireless networks, and some key advantages are established. Based on the DPC and FM schemes, two protocols are developed, which leverage adaptive tuning of DPC parameters. One of them is inspired by TCP and exhibits analogous behavior.

In [21], we studied the weighted sum rate maximization problem in wireless networks consisting of multiple source-destination pairs. The optimization problem is to maximize a weighted sum of data rates by adjusting the power of each user. The problem is in general a non-convex optimization problem that will lead to multiple local maxima. A Gauss-Seidel type iterative power control algorithm was presented. We showed by simulation that the proposed algorithm converges to the global maximum with very high probability, if we initialize the initial power allocation uniformly at random. The proposed algorithm also has the favorable properties that only simple operations are needed in each iteration, and the convergence is fast. Performance comparison under different user densities has also indicated its effectiveness. Finally, we discussed some simple and optimal power allocation strategies under special cases of the problem if the network can be represented by a certain approximation.

6.1.4.2. Simultaneous Decoding

In [15], in collaboration with A. El Gamal [Stanford, USA] and D. Tse [UC Berkeley, USA], we analyzed a network made of a collection of transmitter-receiver links where each link is considered to be part of a Multiple Access Channel (MAC) together with a collection of co-transmitters, rather than treating the messages of the latter as noise. This MAC extension is meant to improve the rate of the link and not to decode the messages of the co-transmitters. The necessary and sufficient condition for the feasibility of some rate when using successive interference cancellation and simultaneous decoding were provided. The reasons why simultaneous decoding is preferable to successive interference cancellation were also given. The gain obtained when using this type of simultaneous decoding rather than treating interference as noise was then quantified in a network made of a large random collection of such links. The gains in coverage and in rate were analyzed in terms of ensemble averages, evaluated using stochastic geometry. Closed form or integral expressions were obtained for the outage/coverage probability in networks where nodes are randomly distributed like a Poisson point process on an infinite plane. In the CDMA limit (large bandwidth, low SINR per hertz, high density), the ensemble average of the link rates tends to 0 when interference is treated as noise whereas it tends to a positive constant when simultaneous decoding of infinite order is used. The whole analysis was conducted in the AWGN case.
6.2. Network Dynamics


This traditional research topic of TREC has several new threads like perfect simulation, active probing or Markov decision.

6.2.1. Network Calculus

Network calculus is a theory that aims at computing deterministic performance guarantees in communication networks. This theory is based on the (min,plus) algebra. Flows are modeled by an arrival curve that upper-bounds the amount of data that can arrive during any interval, and network elements are modeled by a service curve that gives a lower bound on the amount of service offered to the flows crossing that element. Worst-case performances are then derived by combining these curves.

6.2.1.1. Performance bounds networks with static priorities

In cooperation with Aurore Junier [INRIA/IRISA], we present in [29] algorithms to compute worst-case performance upper bounds when the service policy is static priorities, using linear programming. Linear programming does not lead to tight bounds, but when combining this method with (min,plus) methods, we obtain bounds that outperform the already known bounds. Also, we prove that in tandem networks, the the worst-case performance bound under arbitrary multiplexing can be obtain by a policy with static priorities, the “shortest-destination first” policy.

6.2.1.2. Feed-forward networks with wormhole routing discipline

In collaboration with Bruno Gaujal [INRIA Rhone Alpes] and Nadir Farhi [IFFSTAR] we are working on a model of performance bound calculus on feed-forward networks where data packets are routed under wormhole routing discipline. We are interested in determining maximum end-to-end delays and backlogs for packets going from a source node to a destination node, through a given virtual path in the network. Our objective is to give a “network calculus” approach to calculate the performance bounds. For this, we propose a new concept of curves that we call packet curves. The curves permit to model constraints on packet lengths for data flows, when the lengths are allowed to be different. We used this new concept to propose an approach for calculating residual services for data flows served under non preemptive service disciplines. This notion also enabled us to differentiate different classes of service policies: those that are based on a packet count (like round-robin and its generalized version), where the packet curve will be useful to tighten the bounds computed, and those that are based on the amount of data served (FIFO, priorities), where it won’t be useful. These results can be found in [44] and have been presented in ILAS 2011.

6.2.1.3. Composition of service curves in Network Calculus

In envelope-based models for worst-case performance evaluation like Network Calculus or Real-Time Calculus, several types of service curves have been introduced to quantify some deterministic service guarantees. We compare those different classes of service curves regarding the composition (servers in tandem) and individual service curves (when several flows share a server, what service curve can be guaranteed to each of the flows?). In short, there are two main classes of service curves, simple and strict service curves. Individual service curve can not always be computed when simple service curves are considered, and strict service curves is not a stable class regarding the two operations described. We show that there can be no equivalence between the two main classes of service curves and that no notion of service curve in-between can be defined, that behaves well for the composition. We complete this study by studying other classes of service curves from this viewpoint. These results have been presented in [28].

6.2.1.4. Residuation in (max,plus) automata

With Éric Badouel, Philippe Darondeau [INRIA/IRISA] and Jan Komenda [Institute of Mathematics, Brnó], we study in [27] the decidability of existence and the rationality of delay controllers for systems with time weights in the tropical and interval semirings. Depending on the (max,+) or (min,+) rationality of the series specifying the controlled system and the control objective, cases are identified where the controller series
defined by residuation is rational, and when it is positive (i.e., when delay control is feasible). When the control objective is specified by a tolerance, i.e. by two bounding rational series, a nice case is identified in which the controller series is of the same rational type as the system specification series.

6.2.2. Queueing Theory and Active Probing

6.2.2.1. Inverse Problems

Active probing began by measuring end-to-end path metrics, such as delay and loss, in a direct measurement process which did not require inference of internal network parameters. The field has since progressed to measuring network metrics, from link capacities to available bandwidth and cross traffic itself, which reach deeper and deeper into the network and require increasingly complex inversion methodologies. The thesis of B. Kauffmann [6] investigates this line of thought as a set of inverse problems in queueing theory. Queueing theory is typically concerned with the solution of direct problems, where the trajectory of the queueing system, and laws thereof, are derived based on a complete specification of the system, its inputs and initial conditions. Inverse problems aim to deduce unknown parameters of the system based on partially observed trajectories. A general definition of the inverse problems in this class was provided and the key variants were mapped out: the analytical methods, the statistical methods and the design of experiments. We also show how this inverse problem viewpoint translates to the design of concrete Internet probing applications.

Inverse problems in bandwidth sharing networks theory were also investigated. A bandwidth sharing networks allocates the bandwidth to each flow in order to maximize a given utility function (typically an $\alpha$-fairness), with the constraints given by the capacity of the different servers. In particular, it has been shown that the equilibrium distribution of the bandwidth allocated by TCP to many competing connections is oscillating around an $\alpha$-fair allocation. As such, the theory of bandwidth sharing network is a high-level viewpoint of networks. The meaning of inverse problems in this theory, and their relation to the active probing paradigm are analyzed. In two simple examples of network, the capacity of the different servers and the flow population can estimated, and an algorithm to perform this estimation was proposed.

6.2.2.2. Internet Tomography

Most active probing techniques suffer of the “Bottleneck” limitation: all characteristics of the path after the bottleneck link are erased and unreachable. We are currently investigating a new tomography technique, based on the measurement of the fluctuations of point-to-point end-to-end delays, and allowing one to get insight on the residual available bandwidth along the whole path. For this, we combined classical queueing theory models with statistical analysis to obtain estimators of residual bandwidth on all links of the path. These estimators were proved to be tractable, consistent and efficient. In [59] we evaluated their performance with simulation and trace-based experiments.

Lately this method has been generalized in [72] to a probing multicast tree instead of a single path. This work deals with the complexity of the combinatorials in trees, and gives an explicit formula for the iteration of the Expectation-Maximization (E-M) algorithm. The E-M algorithm is notoriously slow, and we provided three speed-up techniques which are effective in our case (up to a factor $10^3$ in the computation time). These techniques are general, and can be applied to other instances of E-M, or even several other iterative algorithms.

6.2.3. Perfect Sampling of Queueing Systems

Propp and Wilson introduced in 1996 a perfect sampling algorithm that uses coupling arguments to give an unbiased sample from the stationary distribution of a Markov chain on a finite state space $\mathcal{X}$. In the general case, the algorithm starts trajectories from all $x \in \mathcal{X}$ at some time in the past until time $t = 0$. If the final state is the same for all trajectories, then the chain has coupled and the final state has the stationary distribution of the Markov chain. Otherwise, the simulations are started further in the past. This technique is very efficient if all the events in the system have appropriate monotonicity properties. However, in the general (non-monotone) case, this technique requires that one consider the whole state space, which limits its application only to chains with a state space of small cardinality.
6.2.3.1. Piecewise Homogeneous Events

In collaboration with Bruno Gaujal [INRIA Grenoble - Rhone-Alpes], we proposed in [47] a new approach for the general case that only needs to consider two trajectories. Instead of the original chain, we used two bounding processes (envelopes) and we showed that, whenever they couple, one obtains a sample under the stationary distribution of the original chain. We showed that this new approach is particularly effective when the state space can be partitioned into pieces where envelopes can be easily computed. We further showed that most Markovian queuing networks have this property and we propose efficient algorithms for some of them.

The envelope technique has been implemented in a software tool PSI2 (see Section 5.2).

6.2.3.2. Acceleration of Perfect Sampling by Skipping Events

In collaboration with Bruno Gaujal [INRIA Grenoble - Rhone-Alpes], we proposed a new method to speed up perfect sampling of Markov chains by skipping passive events during the simulation [38]. We showed that this can be done without altering the distribution of the samples. This technique is particularly efficient for the simulation of Markov chains with different time scales such as queueing networks where certain servers are much faster than others. In such cases, the coupling time of the Markov chain can be arbitrarily large while the runtime of the skipping algorithm remains bounded. This was further illustrated by several experiments that also show the role played by the entropy of the system in the performance of our algorithm.

6.2.3.3. Aggregated Envelopes

When the cardinality of the state space is so huge that even storing the state of the Markov chain becomes challenging, we propose to combine the ideas of bounding processes and the aggregation of Markov chains [30]. We illustrate the proposed approach of aggregated envelope bounding chains on queueing models with joint arrivals and joint services, often referred to in the literature as assemble-to-order systems. Due to the finite capacity, and coupling in arrivals and services, the exact solving techniques are inefficient for larger problem instances. For instance, for the service tools model proposed by Vliegen and Van Houtum (2009), the aggregated envelope method reduces exponentially the dimension of the state space and allows effective perfect sampling algorithms. We also provide bounds for the coupling time, under the high service rate assumptions.

6.2.4. Markov Chains and Markov Decision Processes

Solving Markov chains is in general difficult if the state space of the chain is very large (or infinite) and lacking a simple repeating structure. One alternative to solving such chains is to construct models that are simple to analyze and provide bounds for a reward function of interest. The bounds can be established by using different qualitative properties, such as stochastic monotonicity, convexity, submodularity, etc. In the case of Markov decision processes, similar properties can be used to show that the optimal policy has some desired structure (e.g. the critical level policies).

6.2.4.1. Stochastic Monotonicity

In collaboration with Jean-Michel Fourneau [PRiSM, Université de Versailles Saint-Quentin] we consider two different applications of stochastic monotonicity in performance evaluation of networks [18]. In the first one, we assume that a Markov chain of the model depends on a parameter that can be estimated only up to a certain level and we have only an interval that contains the exact value of the parameter. Instead of taking an approximated value for the unknown parameter, we show how we can use the monotonicity properties of the Markov chain to take into account the error bound from the measurements. In the second application, we consider a well known approximation method: the decomposition into submodels. In such an approach, models of complex networks are decomposed into submodels whose results are then used as parameters for the next submodel in an iterative computation. One obtains a fixed point system which is solved numerically. In general, we have neither an existence proof of the solution of the fixed point system nor a convergence proof of the iterative algorithm. Here we show how stochastic monotonicity can be used to answer these questions. Furthermore, monotonicity properties can also help to derive more efficient algorithms to solve fixed point systems.
6.2.4.2. Componentwise Bounds

In collaboration with Jean-Michel Fourneau [PRISM, Université de Versailles Saint-Quentin] we proposed an iterative algorithm to compute component-wise bounds of the steady-state distribution of an irreducible and aperiodic Markov chain [17]. These bounds are based on very simple properties of $(\max, +)$ and $(\min, +)$ sequences. We showed that, under some assumptions on the Markov chain, these bounds converge to the exact solution. In that case we have a clear tradeoff between computation and the tightness of bounds. Furthermore, at each step we know that the exact solution is within an interval, which provides a more effective convergence test than usual iterative methods.

6.2.4.3. Markov Reward Processes and Aggregation

In a joint work with I.M. H. Vliegen [University of Twente, The Netherlands] and A. Scheller-Wolf [Carnegie Mellon University, USA] [19], we presented a new bounding method for Markov chains inspired by Markov reward theory: Our method constructs bounds by redirecting selected sets of transitions, facilitating an intuitive interpretation of the modifications of the original system. We show that our method is compatible with strong aggregation of Markov chains; thus we can obtain bounds for an initial chain by analyzing a much smaller chain. We illustrated our method by using it to prove monotonicity results and bounds for assemble-to-order systems.

6.2.4.4. Critical Level Policies in Controlled Queuing Systems

In a joint work with Emmanuel Hyon [University of Paris Ouest Nanterre La Defense and LIP6] [39], we consider a single-item lost sales inventory model with different classes of customers. Each customer class may have different lost sale penalty costs. We assume that the demands follow a Poisson process and we consider a single replenishment hypoexponential server. We give a Markov decision process associated with this optimal control problem and prove some structural properties of its dynamic programming operator. This allows us to show that the optimal policy is a critical level policy. We also discuss some possible extensions to other replenishment distributions and give some numerical results for the hyperexponential server case.

6.2.5. Dynamic Systems with Local Interactions

Dynamic systems with local interactions can be used to model problems in distributed computing: gathering a global information by exchanging only local information. The challenge is two-fold: first, it is impossible to centralize the information (cells are indistinguishable); second, the cells contain only a limited information (represented by a finite alphabet $\mathcal{A}$: $\mathcal{A} = \{0, 1\}$ in our case). Two natural instantiations of dynamical systems are considered, one with synchronous updates of the cells, and one with asynchronous updates. In the first case, time is discrete, all cells are updated at each time step, and the model is known as a Probabilistic Cellular Automaton (PCA) (e.g. Dobrushin, R., Kryukov, V., Toom, A.: Probabilistic cellular systems: ergodicity, memory, morphogenesis, 1990). In the second case, time is continuous, cells are updated at random instants, at most one cell is updated at any given time, and the model is known as a (finite range) Interacting Particle System (IPS) (e.g. Liggett, T.M.: Interacting particle systems, 2005).

6.2.5.1. Density Classification on Infinite Lattices and Trees

In a joint work with N. Fates [INRIA Nancy – Grand-Est], J. Mairesse and I. Marcovici [LIAFA, CNRS and Université Paris 7] [46] we consider an infinite graph with nodes initially labeled by independent Bernoulli random variables of parameter $p$. We address the density classification problem, that is, we want to design a (probabilistic or deterministic) cellular automaton or a finite-range interacting particle system that evolves on this graph and decides whether $p$ is smaller or larger than $1/2$. Precisely, the trajectories should converge (weakly) to the uniform configuration with only $0$’s if $p < 1/2$, and only $1$’s if $p > 1/2$. We present solutions to that problem on $\mathbb{Z}^d$, for any $d \geq 2$, and on the regular infinite trees. For $\mathbb{Z}$, we propose some candidates that we back up with numerical simulations.

6.2.6. Stochastic Stability

6.2.6.1. Ergodicity of Probabilistic Cellular Automata
In a joint work with J. Mairesse and I. Marcovici [LIAFA, CNRS and Université Paris 7] [31], we considered ergodicity properties of probabilistic cellular automata (PCA). A classical cellular automaton (CA) is a particular case of PCA. For a 1-dimensional CA, we proved that ergodicity is equivalent to nilpotency, and is therefore undecidable. We then proposed an efficient perfect sampling algorithm for the invariant measure of an ergodic PCA. Our algorithm does not assume any monotonicity properties of the local rule. It is based on a bounding process which is shown to be also a PCA. We then focused on the PCA Majority, whose asymptotic behavior is unknown, and performed numerical experiments using the perfect sampling procedure.

6.2.6.2. Spatial Queues

In a joint work with S. Foss [Heriot–Watt University, UK] [13], we considered a queue where the server is the Euclidean space, the customers are random closed sets of the Euclidean space arriving according to a Poisson rain and where the discipline is a hard exclusion rule: no two intersecting random closed sets can be served at the same time. We use the max plus algebra and Lyapunov exponents to show that under first come first serve assumptions, this queue is stable for a sufficiently small arrival intensity. We also discuss the percolation properties of the stationary regime of the random closed sets in the queue.

6.3. Economics of Networks

Participants: François Baccelli, Emilie Coupechoux, Marc Lelarge.

6.3.1. Diffusion and Cascading Behavior in Random Networks

The spread of new ideas, behaviors or technologies has been extensively studied using epidemic models. In [69], we considered a model of diffusion where the individuals’ behavior is the result of a strategic choice. We studied a simple coordination game with binary choice and give a condition for a new action to become widespread in a random network. We also analyze the possible equilibria of this game and identify conditions for the coexistence of both strategies in large connected sets. Finally we look at how can firms use social networks to promote their goals with limited information.

Our results differ strongly from the one derived with epidemic models. In particular, we showed that connectivity plays an ambiguous role: while it allows the diffusion to spread, when the network is highly connected, the diffusion is also limited by high-degree nodes which are very stable. In the case of a sparse random network of interacting agents, we computed the contagion threshold for a general diffusion model and showed the existence of (continuous and discontinuous) phase transitions. We also computed the minimal size of a seed of new adopters in order to trigger a global cascade if these new adopters can only be sampled without any information on the graph. We showed that this minimal size has a non-trivial behavior as a function of the connectivity. Our analysis extends methods developed in the random graphs literature based on the properties of empirical distributions of independent random variables, and leads to simple proofs.

6.3.2. Impact of Clustering on Diffusions and Contagions in Random Networks

In [33] we extend some results of the previous results to a model of random graphs having both a given degree distribution and a tunable clustering coefficient. This work shed new light on the impact of clustering on the spread of new ideas, technologies, viruses or worms. We consider two types of growth processes: the (classical SI) diffusion model, and the contagion model, which is inspired by a simple coordination game played on the network and is characterized by a threshold rule and a random seed. While clustering inhibits the diffusion process (on regular graphs), 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.

6.3.3. Economic Value of User Localization in Wireless Networks

The defining characteristic of wireless and mobile networking is user mobility, and related to it is the ability for the network to capture (at least partial) information on where users are located and how users change location over time. Information about location is becoming critical, and therefore valuable, for an increasingly larger number of location-based or location-aware services. A key open question, however, is how valuable exactly this information is. Our goal in this paper is to help understand and estimate the economics, or the value of location information.
In a joint work with J. Bolot [Sprint ATL, USA], [25], we addressed in particular the value of different granularities of location information, for example how much more valuable is it to know the GPS location of a mobile user compared to only knowing the access point, or the cell tower, that the user is associated with. We made three main contributions. First, we presented novel models, which capture the location-based economic activity of mobile users. Second, we derived closed-form analytic solutions for the economic value generated by those users. Third, we augmented the models to consider uncertainty about the users’ location, and derived expressions for the economic value generated with different granularities of location information.

6.4. Point Processes, Stochastic Geometry and Random Geometric Graphs
Participants: François Baccelli, Bartłomiej Błaszczyszyn, Pierre Brémaud, Yogeshwaran Dhandapani, Kumar Gaurav, Mir Omid Haji Mirsadeghi, Justin Salez.

6.4.1. Comparison of Clustering and Percolation of Point Processes and Random Graphs
Heuristics indicate that point processes exhibiting clustering of points have larger critical radius $r_c$ for the percolation of their continuum percolation models than spatially homogeneous point processes. It has already been shown in [64], [73] that the directionally convex ($dcx$) ordering of point processes is suitable to compare their clustering tendencies. Hence, it was tempting to conjecture that $r_c$ is increasing in $dcx$ order. Some numerical evidences support this conjecture for a special class of point processes, called perturbed lattices, which are "toy models" for determinantal and permanental point processes. However the conjecture is not true in full generality. In 2011 we have prepared three publications on this subject.

6.4.1.1. On comparison of clustering properties of point processes
In [52] we provide a large class of perturbed lattice point processes, monotone in $dcx$ order and comparable to Poisson point processes that is commonly considered as the reference model in the comparative study of clustering phenomena. We also introduce a weaker order based on the comparison of only 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 weaker sense to the Poisson point process of the same mean measure.

6.4.1.2. Clustering and percolation of point processes
In [49] we show that simple, stationary point processes of a given intensity on $\mathbb{R}^d$, having void probabilities and factorial moment measures smaller than those of a homogeneous Poisson point process of the same intensity, admit uniformly non-degenerate lower and upper bounds on the critical radius $r_c$ for the percolation of their continuum percolation models. Examples are negatively associated point processes and, more specifically, determinantal point processes. More generally, we show that point processes $dcx$ smaller than a homogeneous Poisson point processes (for example perturbed lattices) exhibit phase transitions in certain percolation models based on the level-sets of additive shot-noise fields of these point processes. Examples of such models are $k$-percolation and SINR-percolation models. We also construct a Cox point process with degenerate critical radius $r_c = 0$, that is $dcx$ larger than a given homogeneous Poisson point process. This is a counterexample for the aforementioned conjecture in the full generality.

6.4.1.3. Ordering of non-standard critical radii
As explained above, heuristically one expects finiteness of the critical radii for percolation of sub-Poisson point processes. However, in [49] we have show that it is non-zero as well. In a more elaborate paper [50] we present a reasoning as to why this non-triviality is to be expected. Specifically, we defined two (nonstandard) critical radii for percolation of the Boolean model, called the lower and upper critical radii, and related, respectively, to the finiteness of the expected number of void circuits around the origin and asymptotic of the expected number of long occupied paths from the origin in suitable discrete approximations of the continuum model. These radii sandwich the usual critical radius $r_c$ for percolation of the Boolean model. We show that $dcx$ order preserves the upper critical radii and reverses the lower critical radii.
6.4.1.4. Local weak convergence and stochastic comparison

Many random models are parametrized by the size of the model, and the essential properties of the model are the asymptotic ones as the size of the graph tends to infinity. In the master thesis [57] we show that the theory of local weak converge provides a natural setting to investigate stochastic (convex) ordering of such models. We consider both the geometric context of [71] and the discrete one of Galton-Watson branching process and Configuration Model, cf [5]. In this latter case we define and study a convex order in the context of random trees and graphs which converge in the local weak sense. In particular, we’re interested in the effect of ordering on percolation. It turns out that while in the case of Galton-Watson trees, convex ordering leads to the ordering of percolation probabilities, we cannot conclude this in the case of configuration model. In this case, we could only obtain the ordering of percolation thresholds.

6.4.1.5. AB random geometric graphs

We investigated percolation in the AB Poisson-Boolean model in $d$-dimensional Euclidean space, and asymptotic properties of AB random geometric graphs on Poisson points in $[0, 1]^d$. The AB random geometric graph we studied is a generalization to the continuum of a bi-partite graph called the $AB$ percolation model on discrete lattices. Such an extension is motivated by applications to secure communication networks and frequency division duplex networks. The AB Poisson Boolean model is defined as a bi-partite graph on two independent Poisson point processes of intensities $\lambda$ and $\mu$ in the $d$-dimensional Euclidean space in the same manner as the usual Boolean model with a radius $r$. We showed existence of $AB$ percolation for all $d \geq 2$, and derived bounds for a critical intensity. Further, in $d = 2$, we characterize a critical intensity. The set-up for $AB$ random geometric graphs is to construct a bi-partite graph on two independent Poisson point process of intensities $\pi$ and $c\pi$ in the unit cube. We provided almost sure asymptotic bounds for the connectivity threshold for all $c > 0$ and a suitable choice of radius cut-off functions $r_n(c)$. Further for $c < c_0$, we derived a weak law result for the largest nearest neighbor radius. This work, which was a part of the PhD thesis [73] will appear in [23].

6.4.2. Random Packing Models

Random packing models (RPM) are point processes (p.p.s) where points which "contend" with each other cannot be simultaneously present. These p.p.s play an important role in many studies in physics, chemistry, material science, forestry and geology. For example, in microscopic physics, chemistry and material science, RPMs can be used to describe systems with hard-core interactions. Applications of this type range from reactions on polymer chains, chemisorption on a single-crystal surface, to absorption in colloidal systems. In these models, each point (molecule, particle,···) in the system occupies some space, and two points with overlapping occupied space contend with each other. Another example is the study of seismic and forestry data patterns, where RPMs are used as a reference model for the data set under consideration. In wireless communications, RPMs can be used to model the users simultaneously accessing the medium in a wireless network using Carrier Sensing Medium Access (CSMA). In this context, each point (node, user, transmitter,· · ·) does not occupy space but instead generates interference to other points in the network. Two points contend with each other if either of them generates too much interference to the other. Motivated by this kind of application, we studied in [70] the generating functionals of several models of random packing processes: the classical Matérn hard-core model; its extensions, the $k$-Matérn models and the $\infty$-Matérn model, which is an example of random sequential packing process. The main new results are: 1) A sufficient condition for the $\infty$-Matérn model to be well-defined (unlike the other two, the $\infty$-Matérn model may not be well-defined on unbounded space); 2) the generating functional of the resulting point process which is given for each of the three models as the solution of a differential equation; 3) series representation and bounds on the generating functional of the packing models; 4) moment measures and other useful properties of the considered packing models which are derived from their generating functionals.

6.4.3. Extremal and Additive Matérn Point Processes

In the simplest Matérn point processes, one retains certain points of a Poisson point process in such a way that no pairs of points are at distance less than a threshold. This condition can be reinterpreted as a threshold condition on an extremal shot–noise field associated with the Poisson point process. In a joint work with P.
Bermolen (Universidad de la República, Montevideo, Uruguay) [11], we studied extensions of Matérn point processes where one retains points that satisfy a threshold condition based on an additive shot–noise field of the Poisson point process. We provide an analytical characterization of the intensity of this class of point processes and we compare the packing obtained by the extremal and additive schemes and certain combinations thereof.

6.4.4. Spatial Birth and Death Point Processes

In collaboration with F. Mathieu [INRIA GANG] and Ilkka Norros [VTT, Finland], we started studying a new spatial birth and death point process model where the death rate is a shot noise of the point configuration [60]. We showed that the spatial point process describing the steady state exhibits repulsion. We studied two asymptotic regimes: the fluid regime and the hard–core regime. We derived closed form expressions for the mean (and in some cases the law) of the latency of points as well as for the spatial density of points in the steady state of each regime.

6.4.5. Information Theory and Stochastic Geometry

In a joint work with V. Anantharam [UC Berkeley], [58], we studied the Shannon regime for the random displacement of stationary point processes. Let each point of some initial stationary point process in \( n \)-dimensional Euclidean space give rise to one daughter point, the location of which is obtained by adding a random vector to the coordinates of the mother point, with all displacement vectors independently and identically distributed for all points. The decoding problem is then the following one: the whole mother point process is known as well as the coordinates of some daughter point; the displacements are only known through their law; can one find the mother of this daughter point? The Shannon regime is that where the dimension \( n \) tends to infinity and where the logarithm of the intensity of the point process is proportional to \( n \). We showed that this problem exhibits a sharp threshold: if the sum of the proportionality factor and of the differential entropy rate of the noise is positive, then the probability of finding the right mother point tends to 0 with \( n \) for all point processes and decoding strategies. If this sum is negative, there exist mother point processes, for instance Poisson, and decoding strategies, for instance maximum likelihood, for which the probability of finding the right mother tends to 1 with \( n \). We then used large deviations theory to show that in the latter case, if the entropy spectrum of the noise satisfies a large deviation principle, then the error probability goes exponentially fast to 0 with an exponent that is given in closed form in terms of the rate function of the noise entropy spectrum. This was done for two classes of mother point processes: Poisson and Matérn. The practical interest to information theory comes from the explicit connection that we also establish between this problem and the estimation of error exponents in Shannon’s additive noise channel with power constraints on the codewords.

We currently investigate extensions of this approach to network information theoretic channels.

6.4.6. Navigation on Point Processes and Graphs

In [12], we studied optimal navigations in wireless networks in terms of first passage percolation on some space-time SINR graph. We established both “positive” and “negative” results on the associated the percolation delay rate (delay per unit of Euclidean distance, also called time constant in the classical terminology of percolation). The latter determines the asymptotics of the minimum delay required by a packet to progress from a source node to a destination node when the Euclidean distance between the two tends to infinity. The main negative result states that the percolation delay rate is infinite on the random graph associated with a Poisson point process under natural assumptions on the wireless channels. The main positive result states that when adding a periodic node infrastructure of arbitrarily small intensity to the Poisson point process, the percolation delay rate is positive and finite.

A new direction of research was initiated aiming at defining a new class of measures on a point process which are invariant under the action of a navigation on this point process. This class of measures has properties similar to Palm measures of stationary point processes; but they cannot be defined in the classical framework of Palm measures.

6.5. Random Graphs and Combinatorial Optimization

Participants: Hamed Amini, Emilie Coupechoux, Mathieu Leconte, Marc Lelarge, Justin Salez.
6.5.1. Rank of Large Random Graphs

In [16], with Charles Bordenave (CNRS-Université de Toulouse), we investigated the rank of the adjacency matrix of large diluted random graphs: for a sequence of graphs converging locally to a Galton-Watson tree, we provided an explicit formula for the asymptotic multiplicity of the eigenvalue 0 in terms of the degree generating function. In the first part, we showed that the adjacency operator associated with a Galton-Watson tree is self-adjoint with probability one; we analyzed the associated spectral measure at the root and characterize the distribution of its atomic mass at 0. In the second part, we established a sufficient condition for the expectation of this atomic mass to be precisely the normalized limit of the dimension of the kernel of the adjacency matrices of the sequence of graphs. Our proofs borrow ideas from analysis of algorithms, functional analysis, random matrix theory, and statistical physics.

6.5.2. Matchings in infinite graphs

In [43], we proved that for any sequence of (deterministic or random) graphs converging locally, the corresponding sequence of normalized matching numbers converges, and this limit depends only on the limit of the graph sequence. In the particular case where this limit is a unimodular Galton Watson tree, we were able to compute explicitly the value for the limit of the sequence of (normalized) matching numbers. This leads to an explicit formula that considerably extends the well-known one by Karp and Sipser for Erdős-Rényi random graphs.

We considered a natural family of Gibbs distributions over matchings on a finite graph, parameterized by a single positive number called the temperature. The correlation decay technique can be applied for the analysis of matchings at positive temperature and allowed us to establish the weak convergence of the Gibbs marginal as the underlying graph converges locally. However for the zero temperature problem (i.e. maximum matchings), we showed that there is no correlation decay even in very simple cases. By using a complex temperature and a half-plane property due to Heilmann and Lieb, we were able to let the temperature tend to zero and obtained a limit theorem for the asymptotic size of a maximum matching in the graph sequence.

6.5.3. Counting spanning subgraphs subject to local constraints

In [53], we use negative association and local weak convergence to establish the validity of the cavity method for counting spanning subgraphs subject to local constraints. Specifically, the normalized logarithm of the associated generating polynomial (or partition function) is shown to converge along any sequence of graphs whose random weak limit is a tree, and the limit is directly expressed in terms of the unique solution to a limiting cavity equation. On a Galton-Watson tree, the latter simplifies into a recursive distributional equation which can be solved explicitly. As an illustration, we provide an asymptotic formula for the maximal size of a spanning subgraph with maximal degree $b$ in the Erdős-Rényi model with fixed average degree and diverging size, for any $b \in \mathbb{N}$.

6.5.4. Bipartite graph structures for efficient balancing of heterogeneous loads

With Laurent Massoulié (Technicolor), we extend the results obtained previously on the asymptotic size of maximum matchings in random graphs converging locally to Galton-Watson trees to so-called b-matchings (with non-unitary capacity at vertices as well as constraints on individual edges). Compared to the matching case, this involves studying the convergence of a message passing algorithms which transmits vectors instead of single real numbers. We also look further into an application of these results to large scale distributed content service platforms, such as peer-to-peer video-on-demand systems. In this context, the density of maximum b-matchings corresponds to the maximum fraction of simultaneously satisfiable requests, when the service resources are limited and each server can only handle requests for a predetermined subset of the contents which it has stored in memory. An important design aspect of such systems is the content placement strategy onto the servers depending on the estimated content popularities; the results obtained allow to characterize the efficiency of such placement strategies and the optimal strategies in the limit of large storage capacity at servers are determined.
6.5.5. Flooding in Weighted Random Graphs

In a joint work [24] with Moez Draief [Imperial College London], we studied the impact of the edge weights on distances in diluted random graphs. We interpret these weights as delays, and take them as i.i.d exponential random variables. We analyzed the edge flooding time defined as the minimum time needed to reach all nodes from one uniformly chosen node, and the edge diameter corresponding to the worst case edge flooding time. Under some regularity conditions on the degree sequence of the random graph, we showed that these quantities grow as the logarithm of $n$, when the size of the graph $n$ tends to infinity. We also derived the exact value for the prefactors.

These allowed us to analyze an asynchronous randomized broadcast algorithm for random regular graphs. Our results show that the asynchronous version of the algorithm performs better than its synchronized version: in the large size limit of the graph, it will reach the whole network faster even if the local dynamics are similar on average.