RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2012

Section New Results

Edition: 2013-04-24

ACES Project-Team

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. Traditionnal 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.

Our research 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 [3], [4]. Several innovative aspected are studied in this project. First, the design of an autonomous computing architecture for the waste items and smart containers, enabling early processing in the waste management: for example waste bags can be accepted or rejected accordingly to their content and its conformance with the recipient container. Hazard prevention and human operator safety can also be improved with the knowledge of the nature of the waste.

Autonomy is important as it would be possible to depend on a remote information system for each waste insertion, due to obvious scalability, energy and network costs. An ontology based system has been proposed to determine the possible interactions of tagged products based on their properties and the external conditions [6]. This ontological model is simple enough to be supported entirely by a low power embedded computer at the container level, but can still support the waste application requirements. An unconventional aspect in this architecture is that semantic properties are directly written in the RFID tags, instead of semantic-less identifiers typically used in most RFID applications.

A second innovative aspect of the research is to consider the set of containers in a city as a particular case of sensor network, and developing energy efficient protocol to enable information reporting to a supervising infrastructure.

In the context of this research, some limitations of existing RFID technology become challenging: unlike standard RFID application scenarios, pervasive computing often involves uncontrolled environment for RFID, where tags and reader have to operate in much more difficult situations that those usually encountered or expected for classical RFID systems. In a near future, we seek to work with a team who has a strong expertise in antenna design and radio signal behaviour.

4.2. Integrity checking with coupled objects

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

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. In 2011, we have 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. But the possibility of defining hierarchical couplings was lacking.

This is why we have studied and and designed the Ubi-Tree environment in 2012, which strives to deal with those new requirements. Ubi-Tree 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.

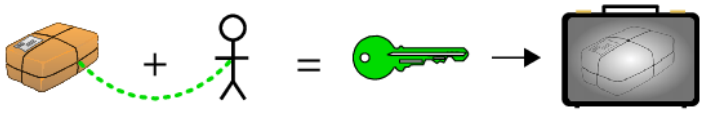

Figure 2. Key to a Ubi-Post briefcase

The Ubi-Tree environment has been experimented through a content-oriented security solution for high value shipping: the Ubi-Post briefcase. Sending sensitive documents or parcels over a delivery service can be a hazardous operation. Goods can be picked up by a fake courier, genuine items can be swapped with copies, the parcel may be received or opened by someone else than the supposed recipient and some items can be missing at the delivery time. As some very high value items are sent over such services, security is critical. We proposed the Ubi-Post briefcase system, a pervasive content-oriented security solution for high value shipping based on the Ubi-Tree physical object coupling software and RFID equipment. The aim of a shipping service is to provide transportation of goods from a sender to the recipient, so the system must ensure that the coupling would be handed over to the recipient. For that purpose, coupled tags will carry an identifier corresponding to the recipient as additional data. Then, the only way to unlock a Ubi-Post briefcase is to insert a recipient card which tag ID is the one expected by the coupling (see figure 2). The Ubi-Post briefcase embeds the same equipment as the coupling station, plus a battery, an HF near field card reader, and a locking mechanism (see Figure 3).

We have produced an interface for users to be sure that the association between RFID tag and physical object is the one that is perceived by our coupling software. The key idea was to be able to identify in the right way the RFID tag associated to a physical object when we place one physical object onto the support of the antenna linked to the RFID reader. The position of this object, and the tag associated to this object, 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, 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 now fully functional.

4.3. Pervasive support for Smart Homes

Participants: Michele Dominici, Bastien Pietropaoli, Sylvain Roche, Frédéric Weis [contact].

A smart home is a residence equipped with information-and-communication-technology (ICT) 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.

The idea of using the Ubiquitous Computing paradigm in the smart home domain is not new. However, the state-of-the-art solutions only partially adhere to its principles. Often the adopted approach consists in a heavy deployment of sensor nodes, which continuously send a lot of data to a central elaboration unit, in charge of the difficult task of extrapolating meaningful information using complex techniques. This is a

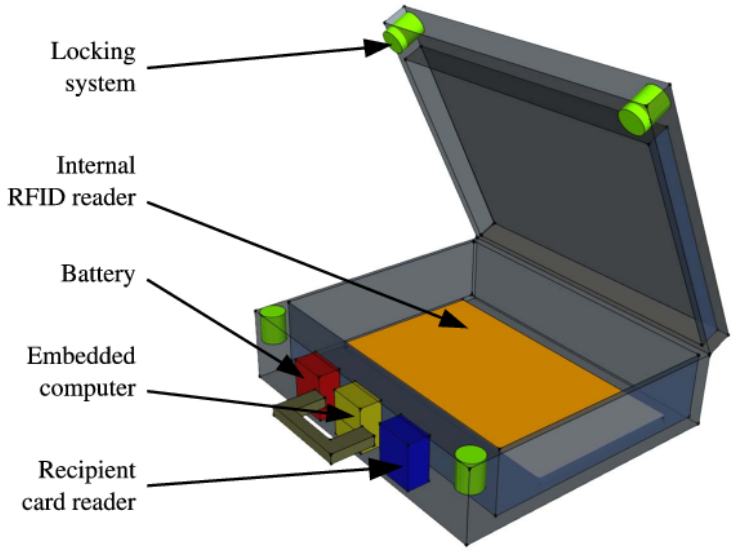

Figure 3. 3D view from the internal components of the Ubi-Post briefcase

logical approach. ACES proposed instead the adoption of a *physical approach*, in which the information is spread in the environment, carried by the entities themselves, and the elaboration is directly executed by these entities "inside" the physical space. This allows performing meaningful exchanges of data that will thereafter need a less complicate processing compared to the current solutions. The result is a smart home that can, in an easier and better way, integrate the context in its functioning and thus seamlessly deliver more useful and effective user services. Our contribution aims at implementing the physical approach in a domestic environment, showing a solution for improving both comfort and energy savings.

Most existing smart home solutions were designed with a technology-driven approach. That is, the designers explored which services, functionalities, actions and controls could be performed exploiting available technologies. This led to solutions for human activity recognition relying on wearable sensors, microphones or video cameras. Those technologies may be difficult to deploy and get accepted in real-world households, because of convenience and privacy concerns. Many people have concerns on carrying equipments or feeling observed or recorded while living their private life. This could seriously impact the acceptability of the smart home system or reduce its diffusion in real households. To avoid such kind of issues, we designed our system with an acceptability-driven approach. That is, we selected technologies that respond to the constraints of a real-world deployment of the future smart home system, namely, convenience and privacy concerns. We decided to take a very conservative approach, choosing technologies that are as unobtrusive as possible, in order to explore the frontiers of what can be done in a smart home with a very limited instrumentation. Following the same considerations, the adopted technologies and techniques had to guarantee a fast and easy configuration, ultimately allowing a plug-and-play deployment.

4.3.1. Design and implementation of a system architecture

In 2012, we have designed and experimented a system architecture of a smart home prototype currently under development. It is the demonstrator of an interdisciplinary project that brings together industrials and researchers, from the fields of ubiquitous computing and cognitive ergonomics. The aim is to develop a smart home system that is able to prevent energy waste and preserve inhabitants' comfort. The key requirement is to

provide functionalities that are seamlessly adapted to ongoing situations and activities of inhabitants, avoiding bothering them with inappropriate interventions. The architecture of such a system has been designed so as to respect the principles and constraints illustrated in the introduction of this section. Namely, we have chosen the necessary equipments among those that should guarantee privacy preservation and high acceptability. When designing the algorithms for context and situation recognition and the human-computer interaction aspects of the system, we have kept in mind the model of human activity described in the previous section. Finally, we have designed the architecture of the system so as to realize successive abstraction of contextual information and to allow uncertainty, imprecision and ignorance to flow between the layers [2].

4.3.2. Layered architecture

The system architecture relies on the principles of the ubiquitous computing paradigm. It also draws its inspiration from the work of Coutaz, who suggest a four-layer model to build context-aware applications. The first layer, "sensing", is in charge of sensing the environment. It is realized by augmented appliances and physical sensors. The augmented household appliances provide information about their state, while the sensors measure physical phenomena (sound level, motion, vibration, etc.). The second layer, called "perception", realizes the abstraction from the raw data. These are processed to obtain more abstract information about the context (e.g. the detection of presence in a room can be obtained combining motion, sound and vibration measures). "Situation and context identification", the third layer, identifies the occurring situations and the activities of inhabitants. For instance, the fact that a given moment a person is ironing can be modeled combining the information that a person is present in a room with the fact that the iron is on and that it is being moved. The top layer, called "exploitation", provides contextual information to applications. More specifically, the contextual information is used to adapt the behavior of the augmented appliances in a semi-automatic way and to allow lowly interruptive takeover by inhabitants.

4.3.3. Design and experimentation of the "perception" layer

In the second layer called "perception", raw sensor data are processed to obtain more abstract information about context called Context Attributes. These are small pieces of context easily understandable by humans and that can be provided to the upper layer. Examples of Context Attributes are the presence, the number of people in a room or the posture of someone. Some raw data are immediately exploitable, like temperature or light level. Others require data fusion in order to obtain more abstract contextual information, such as inhabitants' presence or movement. A certain number of sensors is necessary to obtain sufficient certainty when fusing data, as redundancy can significantly increase the reliability of the sources. Furthermore, heterogeneous sensors allow collecting different physical measurements that can enrich the data fusion process.

Data fusion is a large problem. Many theories offer tools to handle it. In our approach, the main aim of the perception layer is to abstract imperfect raw data to make it computable by higher level reasoning algorithms. Data may be imperfect for different reasons:

- Randomness, due to physical systems (in our case, sensors).
- Inconsistency, due to overload of data or conflicting sources.
- Incompleteness, due to loss of data which may easily happen with wireless communication.
- Ambiguity (or fuzziness), due to models or to natural language imprecision.
- Uncertainty, due to not fully reliable sources.
- Bias, due to systematic errors.
- Redundancy, due to multiple sources measuring the same parameter.

In order to manage many of those imperfections and respect the theoretical constraints, we decided to use as a first layer of abstraction the belief functions theory (BFT). The BFT can be seen as a generalization of the Bayesian theory of subjective probability. It can be used to model probabilities if only atomic focal sets are used in mass functions. Thus, it is totally possible to mix probabilities with real belief functions.

In our approach, we considered that sensors should duce belief for a certain amount of time after the measures because of the continuity of studied context. For instance, a motion sensor in a room could be able to induce a belief on the presence of someone for a longer time than the exact moment at which the measure has been obtained. It is a matter of physical system with inertia. In this example, it is easy to take into account that physical persons cannot move too fast and thus will certainly be there for some seconds before they can exit the room. Thus, this little example brings two questions: how to build evidence from raw data and how to take into account evidence over time? We proposed a simple method already existing to build belief functions from raw data and propose an improvement to take into account timed evidence [5].

4.3.4. Design and experimentation of the "situation and context identification" layer

"Situation and context identification", the third layer, identifies the occurring situations and the activities of inhabitants. For instance, the fact that a given moment a person is ironing can be modeled combining the information that a person is present in a room with the fact that the iron is on and that it is being moved. Having obtained the Context Attributes through abstraction from the raw sensor data, the system has to reason about context, in order to infer higher-level context information, needed to make decisions concerning the functionalities to offer to inhabitants. We needed a unified theory for modeling contextual information, also offering a generic framework for applying different reasoning techniques to infer higher-level context.

We adopted a situation-centric modeling and reasoning approach called *Context Spaces*, based on a unified context modeling and reasoning theory. Using this theory, interesting situations can be modeled as combinations of basic contextual information provided both by a sensor-data-fusion technique and by augmented appliances. Adapted functionalities can be provided when the interesting situations are triggered. The recognition of ongoing situations is made possible by reasoning about available context information. The Context Spaces theory allows managing and propagating uncertainty and ignorance, reasoning on ambiguous contexts and assessing the degree of uncertainty of the resulting inference. It also provides tools to reason on complex logical expressions that combine elementary situations. The use and the extension of the Context Spaces is the core of a PhD thesis that has been finished at the end of 2012 by Michele Dominici (to be defended in March 2013).

4.3.5. Uncertainty and ignorance management

Given the gap between contextual capture capabilities of our architecture and actual complexity of real-world human activities and context, an important issue arises: the management of uncertainty and ignorance. If contextual information has to be abstracted in successive steps, sources are not always reliable. In particular, uncertainty is intrinsic to the physical sensors that are used in the capture. Thus, the uncertainty of lower abstraction layers will negatively impact the inference and decisions of the upper layers. Furthermore, due to the contextual gap illustrated above, any computing model that tries to represent the complexity of real activity will be affected by a certain degree of uncertainty. This reflects on the recognition of the activity itself and can lead to wrong conclusions, which in turn negatively impact the provision of adapted functionalities to inhabitants. As a consequence, we considered that information about uncertainty and ignorance has to be propagated, cumulated and considered at every layer of our pervasive architecture. Whenever the level of uncertainty becomes excessively high, the system tried to evaluate the tradeoff between the potential benefit of providing the right functionality and the risk associated with an unsuitable functionality, which would be provided in case the situation has not been correctly recognized.

ADAM Project-Team

6. New Results

6.1. Software Product Lines

In terms of Software Product Lines [92], we work in four different directions. First, we define a SPL framework for Cloud Computing called SALOON [62] to face challenges in terms of application configuration, cloud platform configuration [59] and deployment automation [58]. Second, we use Dynamic Software Product Lines (DSLP) for mobile applications [21], in order to support self-adaptation of context-aware applications in ubiquitous environments [56] and Wireless Sensor Networks (WSNs) [36]. In both cases, Constraint Satisfaction Problem (CSP) techniques are used in order to find a suitable configuration for the current environment state and to deal with contradictory dimensions (e.g., accuracy and energy saving) in the decision making process. Third, in the YourCast project [76], we work in a Composite SPL for Broadcasting System by identifying the main issues that we need to deal with when defining such kind of SPL. Finally, we define an operator to compute syntactic and semantic differences between feature models [24].

6.2. Software Evolution

The adaptive software paradigm supports the definition of software systems that are continuously adapted at run-time. An adaptation activates multiple features in the system, according to the current execution context (*e.g.*, CPU consumption, available bandwidth). However, the underlying approaches used to implement adaptation are ordered, *i.e.*, the order in which a set of features are turned on or off matters. Assuming feature definition as etched in stone, the identification of the right sequence is a difficult and time-consuming problem. We propose here a composition operator that intrinsically supports the commutativity of adaptations [50]. Using this operator, one can minimize the number of ordered compositions in a system. It relies on an action-based approach, as this representation can support preexisting composition operators as well as our contribution in an uniform way. This approach is validated on the Service-Oriented Architecture domain, and is implemented using first-order logic.

6.3. Green Middleware

The energy consumption of ICT is widely acknowledged as continuously growing over years and its carbon footprint can now be compared to the avionics domain. While green computing has emerged as a new research area concerned with the optimization of the energy consumption of large-scale systems, such as datacenters, our project-team investigates the analysis of the energy consumption from a software engineering point of view. In particular, we developed e-Surgeon, a middleware framework to estimate the power consumption of legacy software at various levels of granularity. With respect to this objective, the first result we obtained in [52] relates to an evaluation of the impact of programming languages and programming styles on the energy consumption of applications. While the current trend in application servers is to adopt interpreted languages (*e.g.*, JavaScript, Python) on the server side, our preliminary results highlight that these languages impose a large overhead to the energy consumption of the resulting system. In [53], this preliminary result is further investigated by identifying energy hotspots within legacy application servers. To do so, we automatically instrument the code of the application server to analyze how the energy is consumed by the application server under various stress scenarios. Our results show that the energy is mostly consumed by a restricted number of classes and methods of these application servers, thus giving hints to software developers on candidate snippets for optimization.

6.4. Human-Competitive Software Engineering

Frequently asked questions (FAQs) are a popular way to document software development knowledge. As creating such documents is expensive, we have invented an approach for automatically extracting FAQs from sources of software development discussion, such as mailing lists and Internet forums, by combining techniques of text mining and natural language processing. We applied the approach to popular mailing lists and carried out a survey among software developers to show that it is able to extract high-quality FAQs that may be further improved by experts. This research has been published at the International Conference on Software Engineering (ICSE'2012 [40]), the flagship conference in the domain. This work takes place in our line of research on "human competitive software engineering", where we try to replace manual tasks requiring costly human skills (such as documentation writing or bug fixing) by automated or semi-automated approaches.

6.5. Reconfigurable Middleware

In the context of our collaboration with the Thales company, especially via the PhD of Jonathan Labéjof defended on 20 December 2012, we obtained some results in the domain of reconfigurable message-oriented middleware (MOM). MOM are a particular class of middleware systems that promote asynchronous communications and weak coupling between communicating entities. They are of particular interest for the design of Systems of Systems (SoS). In this context, we worked on a method for reconfiguring quality of service properties in MOM. The idea is to be able change the properties of communication channels without stopping these channels. We obtained this by defining a bijection between the characteristics of these channels and a component-based software architecture for which we already have means of reconfiguration with our previous results on the FRASCATI platform (see Section 5.5). By this way, reconfiguring the quality of service of a channel is akin to reconfiguring its associated component-based software architecture. This result has been applied to MOM platforms that conform to the OMG DDS standard.

This result has been the topic of a patent application [106] that was filled in Europe in July 2011 and in the US in July 2012. The results were also presented in the SCDI workshop at the EDOC 2012 conference [44].

ARLES Project-Team

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 2012 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
- The use of Requirement Engineering techniques for enabling systems to be self-adaptive under uncertainty (§ 6.6).

6.2. Emergent Middleware Supporting Interoperability in Extreme Distributed Systems

Participants: Emil Andriescu, Amel Bennaceur, Luca Cavallaro, Valérie Issarny, 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 thus been proposed to address the interoperability challenge. However, 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 investigated. 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 (§ 7.2.1.1). Our work during the year has more specifically focused on:

- **Architecture enabling emergent middleware.** We have been finalizing, together with our partners in the CONNECT project, the definition of an overall distributed system 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 [31].

- **Affordance inference.** We have proposed an ontology-based formal model of networked systems based on their affordances (high-level functionalities), interfaces, behavior, and non-functional properties, each of which describes a different facet of the system in a way similar to the service description promoted for semantic Web services. However, legacy systems do not necessarily specify all of the aforementioned facets. Therefore, we have explored 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. In a complementary way, CONNECT partners investigate protocol-learning algorithms to learn the behavior of networked systems on the fly [17].
- **Mediator synthesis for emergent middleware.** We focus on systems that have compatible functionalities, i.e., semantically matching affordances, but are unable to interact successfully due to mismatching interfaces or behaviors. To solve such mismatches, we propose a *mapping-based* approach, whose goal is to automatically synthesize a mediator model that ensures the *safe* interaction of functionnally compatible systems, 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 and for which we investigated a solution using ontology-based model checking [16], the proposed mapping-based approach handles the more general cases of one-to-many and many-to-many mappings. This work has resulted in a supporting software prototype that allows validating the approach; related publication is under writing. A further key research issue we are addressing in emergent middleware is the study of cross-paradigm interaction so as to enable interoperability among highly heterogeneous services (e.g., an IT-based service will likely interact using the client-service scheme while thing-based services rather adopt asynchronous protocols). Toward that goal, we are studying abstract models associated with popular interaction paradigms and higher level, generic interaction paradigms to define cross-paradigm mappings that respect the behavioral semantics of the interacting systems.
- **Automated mediation for cross-layer protocol interoperability.** While existing approaches to interoperability consider either application or middleware heterogeneity separately, we believe that in real world scenarios this does not suffice: application and middleware boundaries are ill-defined and solutions to interoperability must consider them in conjunction. As part of our recent work, we have proposed such a solution, which solves cross-layer interoperability by automatically generating parsers and composers that abstract physical message encapsulation layers into logical protocol layers, thus supporting application layer mediation. Specifically, we support the automated synthesis of mediators at the application layer using the mapping-based approach discussed above, while we introduce *Composite Cross- Layer (CCL) parsers and composers* to handle cross-layer heterogeneity and to provide an abstract representation of the application data exchanged by the interoperating components. In particular, we associate the data embedded in messages with annotations that refers to concepts in a domain ontology. As a result, we are able to reason about the semantics of messages in terms of the operations and the data they require from or provide to one another and automatically synthesize, whenever possible, the appropriate mediators. We have demonstrated the validity of our approach by using the framework to solve cross-layer interoperability between existing conference management systems.
- **Models@run.time.** We have recently integrated the notion of *Models@run.time* 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 way of using runtime models captures syntax and also semantics of behaviour and supports runtime reasoning. Prior models@run.time approaches have generally concentrated on architectural-based runtime models and self-adaptation of existing software artifacts. However, such artefacts cannot always be produced in advance, and we believe that models@runtime have a fundamental role to play in the production of dynamic, adaptive, and on-the-fly software as investigated in the context of emergent middleware [8]. Specifically, two important methods underpin our approach: *i*) automatic inference of the required runtime models during execution and their refinement by exploiting learning and synthesis techniques; and *ii*) using these models for a dynamic software synthesis approach, where mediators are formally characterized (using LTS) to allow the runtime synthesis of software.

In order to enable emergent middleware, we have shown how systems can infer information to build runtime models during execution. Importantly, ontologies were exploited to enrich the runtime models and facilitated the mutual understanding required to perform the matching and mapping between the networked heterogeneous systems. Such reasoning about information that was not necessarily known before execution, is in contrast to the traditional use of models@run.time.

6.3. Revisiting the Abstractions of Service Oriented Computing for the Future Internet

Participants: Dionysis Athanasopoulos, Sandrine Beauche, George Bouloukakis, Oleg Davidyuk, Nikolaos Georgantas, Valérie Issarny, Ajay Kattepur.

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 *ities* 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.

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 middleware instances.
- 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 two realizations of XSB for the CHOReOS project [30], the first on PEtALS ESB and the second on EasyESB. The genericity and modularity of our solution allowed for easily porting from the first implementation to the second one. We support interoperable peer-to-peer interaction among the CS, PS, and TS paradigms and provide templates for systematic and highly facilitated building of binding components for middleware platforms that follow any one of the three paradigms.

6.4. Service Oriented Middleware facing the Challenges of the Internet of Things

Participants: Benjamin Billet, Nikolaos Georgantas, Sara Hachem, Valérie Issarny, Yesid Jarma Alvis, Cong Kinh Nguyen, George Mathioudakis.

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 [6]. 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 this domain has been focused on two related directions:

- **Architecture of a Service Oriented Middleware for the Mobile Internet of Things**: Adopting the service-oriented architecture (SOA) approach towards middleware (see § 3.3), is an adequate solution towards addressing the heterogeneity and the unknown network topology issues in the IoT. SOA is commonly used in IoT solutions to abstract *things* or their measurements as services. The service-oriented paradigm decouples the functionalities of things from their hardware information or other technical details, and supports three core functionalities: *Discovery* and *Composition* of, and *Access* to services. Typically, in traditional uses of SOA, even if millions of services are registered, there is no need to select and access them all simultaneously. However, in the IoT, discovery, composition and access are undoubtedly more complicated. In fact, it is unlikely for a single or even a few services to be sufficient when providing real world measurements. In most cases, to accurately represent a real-world feature, a large number of services are selected to provide their measurements, and subsequently, all acquired values should be properly aggregated. As a consequence, discovery will return a large set of accessible services, redundant as they may be. Consumers are then expected to access the numerous providers to acquire their measurements, over which they should know the exact aggregation/fusion logic to apply. Furthermore, such logic requires precise knowledge

and understanding of the real world and its governing physics and mathematics laws. Clearly, performing discovery, composition and access tasks as presented above incurs high communication and computation costs and is thus not realistic within the large scale IoT. In light of the above issues, we have been *revisiting the SOA and its interaction patterns* to support better scalability and exempt consumers from directly interacting with providers. Specifically, we introduce a **thing-based SOA** to wrap access and computation activities in a middleware that, unlike traditional SOA middleware, is aware of the real world, its physics and its mathematics rules; this has further led to our initial work on the components of such a middleware.

- **Probabilistic Registration for Large Scale Mobile Participatory Sensing**: An increasingly important component of the Internet of Things are modern smart phones, whose constituent sensors and wireless connectivity make them ideal candidates for *mobile participatory sensing*, which aids in providing increased knowledge about the real world while relying on a large number of mobile devices. Those devices can host different types of sensors incorporated in every aspect of our lives. However, given the increasing number of capable mobile devices, any participatory sensing approach should be, first and foremost, *scalable*. To address this challenge, we present an approach to decrease the participation of (sensing) devices in a manner that does not compromise the accuracy of the real-world information while increasing the efficiency of the overall system. To reduce the number of the devices involved, we present a probabilistic registration approach [20], based on a realistic human mobility model, that allows devices to decide whether or not to register their sensing services depending on the probability of other, equivalent devices being present at the locations of their expected path. We used our techniques as the basis of the design and implementation of a registration middleware, using which mobile devices can base their registration decision. Through experiments performed on real and simulated datasets, we show that our approach scales, while not sacrificing significant amounts of sensing coverage.

Our IoT middleware is currently being used by the industrial partners in the FP7 IP CHOReOS project. Complementary to our research on this service oriented middleware for the Internet of Things, we have also been working on suitable abstractions for enabling easy application development for the IoT, discussed next.

6.5. Composing Applications in the Internet of Things

Participants: Peter Sawyer, Pankesh Patel, Animesh Pathak.

As introduced above, the Internet of Things 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 *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 [19].

The larger goal of our research is to propose a suitable application development framework which addresses the challenges introduced above. To that end, our work this year covered the following related areas:

- **Multi-stage Model-driven approach for IoT Application Development**: We have proposed a multi-stage model-driven approach for IoT application development based on a precise definition of the role to be played by each stakeholder involved in the process – domain expert, application designer, application developer, device developer, and network manager. The metamodels/abstractions available to each stakeholder are further customized using the inputs provided in the earlier stages by other stakeholders. We have also implemented code-generation and task-mapping techniques to support our approach. Our initial evaluation based on two realistic scenarios shows that the use of our techniques/framework succeeds in improving productivity in the IoT application development process.
- **Revisiting Requirements Engineering (RE) Practices for IoT**: Requirements engineering (RE) has evolved to discover, model, specify and manage the required and desired properties of software systems. Conventional RE makes an assumption that the knowledge from which the requirements will be formulated exists a-priori, even though the knowledge may be fragmentary, distributed and tacit. Thus, although their discovery may take significant effort, the requirements are discoverable using the appropriate RE practices.

  However, the last decade or so has seen the emergence of new types of systems where this assumption does not hold, including the IoT. Conventional RE is ill-equipped to discover, model, specify and manage these systems' requirements because incomplete knowledge of the context under which they must operate is available at design time. While some progress has been made, by (e.g.) maintaining requirements models that support reasoning over context at runtime, the IoT has now emerged to compound the challenge for RE. Drawing on experiences from ubiquitous computing and WSAN domains, in [22] we provided initial insights into how the field of RE needs to evolve in order to address the challenges brought forth by IoT.

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. Requirements-aware Systems for Self-adaptation under Uncertainty

Participants: Romina Torres, Nelly Bencomo, Valérie Issarny, Peter Sawyer.

The development of software-intensive systems is driven by their requirements. Traditional requirements engineering (RE) methods focus on resolving ambiguities in requirements and advocate specifying requirements in sufficient detail so that the implementation can be checked against them for conformance. In an ideal world, this way of thinking can be very effective. Requirements can be specified clearly, updated as necessary, and evolutions of the software design can be made with the requirements in mind.

Increasingly, however, it is not sufficient to fix requirements statically because they will change at runtime as the operating environment changes. Furthermore, as software systems become more pervasive, there is growing uncertainty about the environment and so requirements changes cannot be predicted at design-time. It is considerations such as these that have led to the development of self-adaptive systems (SASs), which have the ability to dynamically and autonomously reconfigure their behavior to respond to changing external conditions.

The key argument of our research is that current software engineering (SE) methods do not support well the kind of dynamic appraisal of requirements needed by a SAS. definition and structure of requirements is lost as requirements are refined into an implementation. Even in cases where requirements monitoring is explicitly included, high-level system requirements must be manually refined into low-level runtime artefacts during the design process so that they can be monitored. There is a lack of approaches supporting for runtime representation, evolution and assessment of requirements. Currently, the approaches mainly assume that it is possible to predefine and envisage the requirements for the total set of target behaviours. Such estimations

and beliefs may not be appropriate, if the system is to recover during execution from unforeseen situations, or adapt dynamically to new environmental conditions or to satisfy new requirements that were not foreseen during development. A self-adaptive system is able, at run time, to satisfy new requirements and behaviors. Our research focuses on approaches to support the runtime representation of requirements that will underpin the way a system can reason and assess them during execution.

Our research has been carried out within the research project Marie Curie Fellowship called Requirements-aware Systems (nickname: Requirements@run.time). The research is based on a new paradigm for SE, called requirements-awareness (also known as requirements reflection), in which requirements are reified as runtime entities. Requirements-awareness allows systems to dynamically reason about themselves at the level of the requirements - in much the same way that architectural reflection currently allows runtime reasoning at the level of software. We believe that requirements-awareness (i.e. requirements reflection) will support the development and management of SASs because it will raise the level of discourse at which a software system is able to reflect upon itself.

In the above context, we have been working on the design and implementation of systems with the ability to dynamically observe and reason about their requirements. The results will contribute towards the development of conceptual foundations, engineering techniques, and computing infrastructure for the access and manipulations of runtime abstractions of requirements. Currently, a prototype for the use of runtime goals has been developed. The RELAX language has been proposed to make requirements more tolerant to environmental uncertainty. Design assumptions, called Claims), are applied as markers of uncertainty that document how design assumptions affect goals. Monitoring Claims at runtime has been used to drive self-adaptation. By monitoring Claims during the execution of the systems, their veracity can be tested. If a Claim is falsified, the effect can be propagated to the system's goal model and an alternative (more suitable) means of goal realization will be selected, resulting in dynamic adaptation of the system to a configuration that better satisfies the goals under the prevailing environmental context.

ASAP Project-Team

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. Efficient shared memory consensus

Participants: Michel Raynal, Julien Stainer.

This work is on an efficient algorithm that builds a consensus object. This algorithm is based on an Ω failure detector (to obtain consensus liveness) and a store-collect object (to maintain its safety). A store-collect object provides the processes with two operations, a store operation which allows the invoking process to deposit a new value while discarding the previous value it has deposited and a collect operation that returns to the invoking process a set of pairs (i, val) where val is the last value deposited by the process p_i. A store-collect object has no sequential specification.

While store-collect objects have been used as base objects to design wait-free constructions of more sophisticated objects (such as snapshot or renaming objects), as far as we know, they have not been explicitly used to built consensus objects. The proposed store-collect-based algorithm, which is round-based, has several noteworthy features. First it uses a single store-collect object (and not an object per round). Second, during a round, a process invokes at most once the store operation and the value val it deposits is a simple pair $\langle r, v\rangle$ where r is a round number and v a proposed value. Third, a process is directed to skip rounds according to its view of the current global state (thereby saving useless computation rounds). Finally, the proposed algorithm benefits from the adaptive wait-free implementations that have been proposed for store-collect objects, namely, the number of shared memory accesses involved in a collect operation is $O(k)$ where k is the number of processes that have invoked the store operation. This makes this new algorithm particularly efficient and interesting for multiprocess programs made up of asynchronous crash-prone processes that run on top of multicore architectures.

6.1.2. A Contention-Friendly, Non-blocking Skip List

Participants: Tyler Crain, Michel Raynal.

This work [27] presents a new non-blocking skip list algorithm. The algorithm alleviates contention by localizing synchronization at the least contended part of the structure without altering consistency of the implemented abstraction. The key idea lies in decoupling a modification to the structure into two stages: an eager abstract modification that returns quickly and whose update affects only the bottom of the structure, and a lazy selective adaptation updating potentially the entire structure but executed continuously in the background. As non-blocking skip lists are becoming appealing alternatives to latch-based trees in modern main-memory databases, we integrated it into a main-memory database benchmark, SPECjbb. On SPECjbb as well as on micro-benchmarks, we compared the performance of our new non-blocking skip list against the performance of the JDK non-blocking skip list. Results indicate that our implementation is up to 2:5 faster than the JDK skip list.

6.1.3. STM Systems: Enforcing Strong Isolation between Transactions and Non-transactional Code

Participants: Tyler Crain, Eleni Kanellou, Michel Raynal.

Transactional memory (TM) systems implement the concept of an atomic execution unit called a transaction in order to discharge programmers from explicit synchronization management. But when shared data is atomically accessed by both transaction and non-transactional code, a TM system must provide strong isolation in order to overcome consistency problems. Strong isolation enforces ordering between non-transactional operations and transactions and preserves the atomicity of a transaction even with respect to non-transactional code. This work [29] presents a TM algorithm that implements strong isolation with the following features: (a) concurrency control of non-transactional operations is not based on locks and is particularly efficient, and (b) any non-transactional read or write operation always terminates (there is no notion of commit/abort associated with them).

6.1.4. A speculation-friendly binary search tree

Participants: Tyler Crain, Michel Raynal.

In this work [26], in collaboration with Vincent Gramoli, we introduce the first binary search tree algorithm designed for speculative executions. Prior to this work, tree structures were mainly designed for their pessimistic (non-speculative) accesses to have a bounded complexity. Researchers tried to evaluate transactional memory using such tree structures whose prominent example is the red-black tree library developed by Oracle Labs that is part of multiple benchmark distributions. Although well-engineered, such structures remain badly suited for speculative accesses, whose step complexity might raise dramatically with contention. We show that our speculation-friendly 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 in the background. In particular, the speculation-friendly tree is shown correct, reusable and it speeds up a transaction-based travel reservation application by up to 3:5.

6.1.5. 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 [28] presents 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 in 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).

6.1.6. A Tight RMR Lower Bound for Randomized Mutual Exclusion

Participant: George Giakkoupis.

The Cache Coherent (CC) and the Distributed Shared Memory (DSM) models are standard shared memory models, and the Remote Memory Reference (RMR) complexity is considered to accurately predict the actual performance of mutual exclusion algorithms in shared memory systems. Through a collaboration with Philipp Woelfel [32], we proved a tight lower bound for the RMR complexity of deadlock-free randomized mutual exclusion algorithms in both the CC and the DSM model with atomic registers and compare&swap objects and an adaptive adversary. Our lower bound establishes that an adaptive adversary can schedule n processes in such a way that each enters the critical section once, and the total number of RMRs is $\Omega(n \log n / \log \log n)$ in expectation. This matches an upper bound of Hendler and Woelfel (2011).

6.1.7. On the Time and Space Complexity of Randomized Test-And-Set

Participant: George Giakkoupis.

Through a collaboration with Philipp Woelfel [33] we studied the time and space complexity of randomized Test-And-Set (TAS) implementations from atomic read/write registers in asynchronous shared memory models with n processes. We presented an adaptive TAS implementation with an expected (individual) step complexity of $O(\log^* k)$, for contention k, against the oblivious adversary, improving a previous (non-adaptive) upper bound of $O(\log \log n)$ by Alistarh and Aspnes (2011). We also showed how to modify the adaptive RatRace TAS algorithm by Alistarh, Attiya, Gilbert, Giurgiu, and Guerraoui (2010) to improve the space complexity from $O(n^3)$ to $O(n)$, while maintaining logarithmic expected step complexity against the adaptive adversary. Finally, we proved that for any randomized 2-process TAS algorithm there exists a schedule determined by an oblivious adversary, such that with probability at least $1/4^t$ one of the processes does not finish its TAS operation in within fewer than t steps. This complements a lower bound by Attiya and Censor-Hillel (2010) of a similar result for $n \geq 3$ processes.

6.2. Large-scale and user-centric distributed system

6.2.1. WhatsUp: P2P news recommender

Participants: Antoine Boutet, Davide Frey, Arnaud Jegou, 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 and will appear in IPDPS 2013 [21].

6.2.2. Privacy in P2P recommenders

Participants: Antoine Boutet, Davide Frey, Arnaud Jegou, Anne-Marie Kermarrec.

We also propose a mechanism to preserve privacy in WhatsUp, which can also be used in any distributed recommendation system. Our approach relies on (i) an original obfuscation mechanism hiding the exact profiles of users without significantly decreasing their utility, as well as (ii) a randomized dissemination algorithm ensuring differential privacy during the dissemination process. Results show that our solution preserves accuracy without the need for users to reveal their preferences. Our approach is also flexible and robust to censorship.

6.2.3. BLIP: Non-interactive differentially-private similarity computation on Bloom filters

Participants: Mohammad Alaggan, Anne-Marie Kermarrec.

In this project [19], done in collaboration with Sébastien Gambs (team CIDRE), we consider the scenario in which the profile of a user is represented in a compact way, as a Bloom filter, and the main objective is to privately compute in a distributed manner the similarity between users by relying only on the Bloom filter representation. In particular, we aim at providing a high level of privacy with respect to the profile even if a potentially unbounded number of similarity computations take place, thus calling for a non-interactive mechanism. To achieve this, we propose a novel non-interactive differentially private mechanism called BLIP (for BLoom-and-flIP) for randomizing Bloom filters. This approach relies on a bit flipping mechanism and offers high privacy guarantees while maintaining a small communication cost. Another advantage of this non-interactive mechanism is that similarity computation can take place even when the user is offline, which is impossible to achieve with interactive mechanisms. Another of our contributions is the definition of a

probabilistic inference attack, called the “Profile Reconstruction Attack”, that can be used to reconstruct the profile of an individual from his Bloom filter representation, along with the “Profile Distinguishing Game”. More specifically, we provide an analysis of the protection offered by BLIP against this profile reconstruction attack by deriving an upper and lower bound for the required value of the differential privacy parameter ϵ.

6.2.4. Heterogeneous Differential Privacy

Participants: Mohammad Alaggan, Anne-Marie Kermarrec.

The massive collection of personal data by personalization systems has rendered the preservation of privacy of individuals more and more difficult. Most of the proposed approaches to preserve privacy in personalization systems usually address this issue uniformly across users, thus completely ignoring the fact that users have different privacy attitudes and expectations (even among their own personal data). In this project, in collaboration with Sébastien Gambs (team CIDRE), we propose to account for this non-uniformity of privacy expectations by introducing the concept of heterogeneous differential privacy. This notion captures both the variation of privacy expectations among users as well as across different pieces of information related to the same user. We also describe an explicit mechanism achieving heterogeneous differential privacy, which is a modification of the Laplacian mechanism due to Dwork [54], we evaluate on real datasets the impact of the proposed mechanism with respect to a semantic clustering task. The results of our experiments clearly demonstrate that heterogeneous differential privacy can account for different privacy attitudes while sustaining a good level of utility as measured by the recall.

6.2.5. Social Market

Participants: Davide Frey, Arnaud Jegou, Anne-Marie Kermarrec, Michel Raynal, Julien Stainer.

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 an explicit social network like Facebook. This year, we extended the contributions presented in 2011, by introducing two novel distributed protocols that combine interest-based connections between users with explicit links obtained from social networks a-la Facebook. Both protocols build trusted multi-hop paths between users in an explicit social network supporting the creation of semantic overlays backed up by social trust. The first protocol, TAPS2, extends our previous work on *TAPS* (Trust-Aware Peer Sampling), by improving the ability to locate trusted nodes. Yet, it remains vulnerable to attackers wishing to learn about trust values between arbitrary pairs of users. The second protocol, PTAPS (*Private TAPS*), improves TAPS2 with provable privacy guarantees by preventing users from revealing their friendship links to users that are more than two hops away in the social network. In addition to proving this privacy property, we evaluate the performance of our protocols through event-based simulations, showing significant improvements over the state of the art. We submitted this work for journal publication.

6.2.6. Geolocated Social Networks

Participants: Anne-Marie Kermarrec, François Taïani.

Geolocated social networks, that combine traditional social networking features with geolocation information, have grown tremendously over the last few years. Yet, very few works have looked at implementing geolocated social networks in a fully distributed manner, a promising avenue to handle the growing scalability challenges of these systems. In [25], we have focused on georecommendation, and showed that existing decentralized recommendation mechanisms perform in fact poorly on geodata. In this work, we have proposed a set of novel gossip-based mechanisms to address this problem, and captured these mechanisms in a modular similarity framework called "Geology". The resulting platform is lightweight, efficient, and scalable. More precisely, we have shown its benefits in terms of recommendation quality and communication overhead on a real data set of 15,694 users from Foursquare, a leading geolocated social network.

6.2.7. Content and Geographical Locality in User-Generated Content Sharing Systems

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

User Generated Content (UGC), such as YouTube videos, accounts for a substantial fraction of the Internet traffic. To optimize their performance, UGC services usually rely on both proactive and reactive approaches that exploit spatial and temporal locality in access patterns. Alternative types of locality are also relevant and hardly ever considered together. In [34], we show on a large (more than $650,000$ videos) YouTube dataset that content locality (induced by the related videos feature) and geographic locality, are in fact correlated. More specifically, we show how the geographic view distribution of a video can be inferred to a large extent from that of its related videos. We leverage these findings to propose a UGC storage system that proactively places videos close to the expected requests. Compared to a caching-based solution, our system decreases by 16% the number of requests served from a different country than that of the requesting user, and even in this case, the distance between the user and the server is 29% shorter on average.

6.2.8. Probabilistic Deduplication for Cluster-Based Storage Systems

Participants: Davide Frey, Anne-Marie Kermarrec, Konstantinos Kloudas.

The need to backup huge quantities of data has led to the development of a number of distributed deduplication techniques that aim to reproduce the operation of centralized, single-node backup systems in a cluster-based environment. At one extreme, stateful solutions rely on indexing mechanisms to maximize deduplication. However the cost of these strategies in terms of computation and memory resources makes them unsuitable for large-scale storage systems. At the other extreme, stateless strategies store data blocks based only on their content, without taking into account previous placement decisions, thus reducing the cost but also the effectiveness of deduplication. In [30], we propose, Produck, a stateful, yet lightweight cluster-based backup system that provides deduplication rates close to those of a single-node system at a very low computational cost and with minimal memory overhead. In doing so, we provide two main contributions: a lightweight probabilistic node-assignment mechanism and a new bucketbased load-balancing strategy. The former allows Produck to quickly identify the servers that can provide the highest deduplication rates for a given data block. The latter efficiently spreads the load equally among the nodes. Our experiments compare Produck against state-of-the-art alternatives over a publicly available dataset consisting of 16 full *Wikipedia* backups, as well as over a private one consisting of images of the environments available for deployment on the Grid5000 experimental platform. Our results show that, on average, Produck provides (i) up to 18% better deduplication compared to a stateless minhash-based technique, and (ii) an 18-fold reduction in computational cost with respect to a stateful BloomFilter-based solution.

6.2.9. Large scale analysis of HTTP adaptive streaming in mobile networks

Participants: Ali Gouta, Anne-Marie Kermarrec.

In collaboration with Yannick Le Louedec and Nathalie Amann we have been working in the context of adaptive streaming in mobile networks. HTTP Adaptive bitrate video Streaming (HAS) is now widely adopted by Content Delivery Network Providers (CDNPs) and Telecom Operators (Telcos) to improve user Quality of Experience (QoE). In HAS, several versions of videos are made available in the network so that the quality of the video can be chosen to better fit the bandwidth capacity of users. These delivery requirements raise new challenges with respect to content caching strategies, since several versions of the content may compete to be cached. We used a real HAS dataset collected in France and provided by a mobile telecom operator involving more than 485,000 users requesting adaptive video contents through more than 8 million video sessions over a 6 week measurement period. Firstly, we proposed a fine-grained definition of content popularity by exploiting the segmented nature of video streams. We also provided analysis about the behavior of clients when requesting such HAS streams. We proposed novel caching policies tailored for chunk-based streaming. Then we studied the relationship between the requested video bitrates and radio constraints. Finally, we studied the users' patterns when selecting different bitrates of the same video content. Our findings provide useful insights that can be leveraged by the main actors of video content distribution to improve their content caching strategy for adaptive streaming contents as well as to model users' behavior in this context.

6.2.10. Regenerating Codes: A System Perspective

Participants: Anne-Marie Kermarrec, Alexandre van Kempen.

The explosion of the amount of data stored in cloud systems calls for more efficient paradigms for redundancy. While replication is widely used to ensure data availability, erasure correcting codes provide a much better trade-off between storage and availability. Regenerating codes are good candidates for they also offer low repair costs in term of network bandwidth. While they have been proven optimal, they are difficult to understand and parameterize. In collaboration with Nicolas Le Scouarnec, Gilles Straub and Steve Jiekak from Technicolor, we performed an analysis of regenerating codes, which enables practitioners to grasp the various trade-offs. More specifically we made two contributions: (i) we studied the impact of the parameters by conducting an analysis at the level of the system, rather than at the level of a single device; (ii) we compared the computational costs of various implementations of codes and highlight the most efficient ones. Our goal is to provide system designers with concrete information to help them choose the best parameters and design for regenerating codes.

6.2.11. Availability-based methods for distributed storage systems

Participants: Anne-Marie Kermarrec, Alexandre van Kempen.

Distributed storage systems rely heavily on redundancy to ensure data availability as well as durability. In networked systems subject to intermittent node unavailability, the level of redundancy introduced in the system should be minimized and maintained upon failures. Repairs are well- known to be extremely bandwidth-consuming and it has been shown that, without care, they may significantly congest the system. In collaboration with Gilles Straub and Erwan Le Merrer from Technicolor, we proposed an approach to redundancy management accounting for nodes heterogeneity with respect to availability. We show that by using the availability history of nodes, the performance of two important faces of distributed storage (replica placement and repair) can be significantly improved. Replica placement is achieved based on complementary nodes with respect to nodes availability, improving the overall data availability. Repairs can be scheduled thanks to an adaptive per-node timeout according to node availability, so as to decrease the number of repairs while reaching comparable availability. We propose practical heuristics for those two issues. We evaluate our approach through extensive simulations based on real and well-known availability traces. Results clearly show the benefits of our approach with regards to the critical trade-off between data availability, load-balancing and bandwidth consumption.

6.2.12. On The Impact of Users Availability In OSNs

Participants: Antoine Boutet, Anne-Marie Kermarrec, Alexandre van Kempen.

Availability of computing resources has been extensively studied in literature with respect to uptime, session lengths and inter-arrival times of hardware devices or software applications. Interestingly enough, information related to the presence of users in online applications has attracted less attention. Consequently, only a few attempts have been made to leverage user availability pattern to improve such applications. In collaboration with Erwan Le Merrer from Technicolor, we studied an availability trace collected from MySpace. Our results show that the online presence of users tends to be correlated to that of their friends. User availability also plays an important role in some algorithms and focus on information spreading. In fact, identifying central users i.e. those located in central positions in a network, is key to achieve a fast dissemination and the importance of users in a social graph precisely vary depending on their availability.

6.2.13. Chemical programming model

Participant: Marin Bertier.

This work, done in collaboration with the Myriads project team, focuses on chemical programming, a promising paradigm to design autonomic systems. The metaphor envisions a computation as a set of concurrent reactions between molecules of data arising non-deterministically, until no more reactions can take place, in which case, the solution contains the final outcome of the computation.

More formally, such models strongly rely on concurrent multiset rewriting: the data are a multiset of molecules, and reactions are the application of a set of conditioned rewrite rules. At run time, these rewritings are applied concurrently, until no rule can be applied anymore (the elements they need do not exist anymore in the multiset). One of the main barriers towards the actual adoption of such models come from their complexity at run time: each computation step may require a complexity in $O(n^k)$ where n denotes the number of elements in the multiset, and k the size of the subset of elements needed to trigger one rule.

Our objective is to design a distributed chemical platform implementing such concepts. This platform should be adapted to large scale distributed system to benefit at his best the inherent distribution of chemical program.

Within this context, we proposed a protocol for the atomic capture of objects in a DHT [20]. This protocol is distributed and evolving over a large scale platform. As the density of potential request has a significant impact on the liveness and efficiency of such a capture, the protocol proposed is made up of two sub-protocols, each of them aimed at addressing different levels of densities of potential reactions in the solution. While the decision to choose one or the other is local to each node participating in a program's execution, a global coherent behavior is obtained.

ASCOLA Project-Team

6. New Results

6.1. Software composition

Participants: Akram Ajouli, Diana Allam, Omar Chebaro, Rémi Douence, Hervé Grall, Jean-Claude Royer, Mario Südholt.

We have produced results on service-oriented computing, language support for software composition, program transformation for composition, as well as the analysis of C programs.

6.1.1. Program transformation and formal properties

We have proposed an extension of the type theory underlying the Coq theorem prover and studied invertible transformations as a means to decompose object-oriented properties.

6.1.1.1. Forcing in the Calculus of Constructions and Coq

We have developed an intuitionistic forcing translation for the Calculus of Constructions (CoC), a translation that corresponds to an internalization of the presheaf construction in CoC [22]. Depending on the chosen set of forcing conditions, the resulting type theory can be extended with extra logical principles. The translation is proven correct—in the sense that it preserves type checking—and has been implemented in Coq. As a case study, we have shown how the forcing translation on integers (which corresponds to the internalization of the topos of trees) allows us to define general inductive types in Coq, without the strict positivity condition.

6.1.1.2. Invertible transformations for program decompositions

When one chooses a main axis of structural decomposition for a software, such as function- or data-oriented decompositions, the other axes become secondary, which can be harmful when one of these secondary axes becomes of main importance. In the context of modular maintenance, we have tackled this problem using invertible program transformations [19]. We have experimented our approach for Java [29] and Haskell programs.

In [29] we have presented such a transformation for Java. Precisely, we build a reversible transformation between Composite and Visitor design patterns in Java programs, based on chains of basic refactoring operations. Such transformations represent an automatic reversible switching between different program architectures with a guarantee of semantic preservation. The transformation is automated with the refactoring tool of a popular IDE: JetBrains Intellij Idea.

As seen in that paper, basic refactoring operations can be combined to perform complex program transformations. But the resulting composed operations are rarely reused, even partially, because popular tools have few support for composition. In [45] we have formalized the composition of refactoring operations of our Composite/Visitor transformation by the means of a static type system. That type system is based on two previous calculi for composition of refactoring, which we recast in one single calculus. The type system is used to prove non-failure and correctness of transformations. This kind of formalization yields a validation of transformations and, if integrated in existing IDEs, should help to reuse existing transformations.

6.1.2. Service-oriented computing

In the field of service-oriented computing, we have developed three contributions: a model for web services that enables WS*/SOAP-based heavyweight services and RESTful lightweight services to be handled uniformly, a type system that is safe in the presence of malicious agents and insecure communication channels, as well as a pivot language that provides a common abstraction for very different web query languages.

6.1.2.1. Uniform service model

Service-oriented applications are frequently used in highly dynamic contexts: service compositions may change dynamically, in particular, because new services are discovered at runtime. Moreover, subtyping has recently been identified as a strong requirement for service discovery. Correctness guarantees over service compositions, provided in particular by type systems, are highly desirable in this context. However, while service oriented applications can be built using various technologies and protocols, none of them provides decent support ensuring that well-typed services cannot go wrong. Currently, Service-Oriented Architecture applications are typically built using either the SOAP/WS or REST service models. Although there is a clear need for a model integrating both in multiple real-world contexts, no integrated model does (yet) exist. Therefore, in [15] we have introduced a model as a foundation for heterogeneous services, therefore unifying the SOAP/WSand RESTmodels.

6.1.2.2. A type system for services

We have presented a formal model in [14] for service compositions and defined a type system [33] with subtyping that ensures type soundness by combining static and dynamic checks. Our model allows channel mobility and inference of the type of discovered channels. This type system is based on the notion of semantic typing and proved to be sound. We have also demonstrated how to get type soundness in presence of malicious agents and insecure communication channels.

6.1.2.3. Criojo: a pivot language for services

Interoperability remains a significant challenge in service-oriented computing. After proposing a pivot architecture to solve three interoperability problems, namely adaptation, integration and coordination problems between clients and servers, we explore the theoretical foundations for this architecture. A pivot architecture requires a universal language for orchestrating services and a universal language for interfacing resources. Since there is no evidence today that Web Services technologies can provide this basis, we have proposed a new language called Criojo and shown that it can be considered as a pivot language. We have formalized the language Criojo and its operational semantics by resorting to a chemical abstract machine, and given an account of formal translations into Criojo: in a distributed context, we have dealt with idiomatic languages for four major programming paradigms: imperative programming, logic programming, functional programming and concurrent programming.

6.1.3. Languages and composition models

We have contributed new results in the domains of software product lines, model-based composition and language support for numerical constraint-based programming.

6.1.3.1. Software product lines and model composition

Many approaches to creating Software Product Lines have emerged that are based on Model-Driven Engineering. Our book [32] introduces both Software Product Lines and Model-Driven Engineering, which have separate success stories in industry, and focuses on the practical combination of them. It describes the challenges and benefits of merging these two software development trends and provides the reader with a novel approach and practical mechanisms to improve variability. Advanced concepts like fine-grained variability and decision models based on aspect-oriented programming techniques are illustrated. The concepts and methods are detailed with two product line examples: the classic smart-home systems and a collection manager information system.

6.1.3.2. Expressive language support for numerical constraint based programming

A combinatorial search can either be performed by using an implicit or an explicit search tree. We have proposed a functional DSL [35] for explicit search trees in the field of numerical constraints. The first advantage of our approach is expressiveness: we can write new algorithms or reformulate existing ones in a simple and unified way. The second advantage is efficiency, since an implicit search may also lead to a blowup of redundant computations. We illustrate this through various examples.

6.1.4. Analysis and test of C programs

Ascola members have participated, in cooperation with researchers from CEA List institute, in the development of analyses and corresponding tool support for C programs.

We have studied combinations of static and dynamic analysis techniques that enable the detection of out-of-bounds memory accesses in C programs and generate corresponding concrete test data [17]. This is particularly problematic for input arrays and pointers in C functions. We have presented a specific technique allowing the interpretation and execution of assertions involving the size of an input array (pointer) of a C function. We have successfully applied this technique in the Sante tool from the CEA where it allows potential out-of-bounds access errors to be detected and classified in several real-life programs.

PathCrawler is a test generation tool developed at CEA LIST for structural testing of C programs. The new version of PathCrawler [18] we have contributed to is developed in an entirely novel form: that of a test-case generation web service which is freely accessible at PathCrawler-online.com. This service allows many test-case generation sessions to be run in parallel in a completely robust and secure way. We have presented PathCrawler-online.com in the form of a lesson on structural software testing, showing its benefits, limitations and illustrating the usage of the tool on a series of examples.

6.2. Aspect-Oriented Programming

Participants: Rémi Douence, Guilhem Jaber, Ismael Mejía, Jacques Noyé, Mario Südholt, Nicolas Tabareau.

We have contributed to the foundations of aspect-oriented programming and presented new programming languages for aspects and related paradigms.

6.2.1. Formal models for AOP

We have presented two calculi contributing to the foundations of AOP: the A Calculus, a parameterized calculus encompassing AspectJ-like and history based aspect languages, and a category-theoretic definition of AOP in terms of 2-categories.

6.2.1.1. The A Calculus

With partners from Vrije Universiteit Brussel and Aarhus University, we have extended the foundational calculus for AOP (introduced in 2010) that supports the most general aspect model to-date compared to existing calculi and the deepest integration with plain OO concepts [12]. Integration with OOP is achieved essentially by modeling proceed using first-class closures. Two well-known pointcut categories, call and execution that are commonly considered similar are shown to be significantly different; our calculus enables the resolution of the associated soundness problems. The A-calculus also includes type ranges, an intuitive and concise alternative to explicit type variables that allows advices to be polymorphic over intercepted methods. Finally, our calculus is the first aspect calculus to use calculus parameters to cover type safety for a wide design space of other features. The soundness of the resulting type system has been verified in Coq.

In 2012, we have covered a range of choices with respect to evaluation order and non-determinism. We have studied one version that enforces a deterministic call-by-value semantics, and another one that omits restrictions on evaluation order and allows many kinds of non-determinism. Furthermore, we have provided a mechanized complete type soundness proof using the theorem prover Coq.

6.2.1.2. A category-theoretic foundation of aspects

Aspect-Oriented Programming (AOP) started fifteen years ago with the remark that modularization of so-called crosscutting functionalities is a fundamental problem for the engineering of large-scale applications. However, theoretical foundations of AOP have been much less studied than its applicability.

We have proposed [26] to put a bridge between AOP and the notion of 2-category to enhance the conceptual understanding of AOP. Starting from the connection between the λ-calculus and the theory of categories, we have defined 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. We have then used this categorical framework to introduce the notion of computational 2-monads for AOP. 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.

6.2.2. Programming languages for aspects and related paradigms

We have introduced three results related to aspect-based programming languages: an extension of EScala for multi-paradigm programming; Monascheme, a language for modular prototyping of aspect-based languages and language support for membranes, an aspect-based means for structuring computations.

6.2.2.1. Concurrent multi-paradigm programming with EScala

EScala integrates, around the notion of *declarative events*, object-oriented, aspect-oriented and event-based programming [30]. However, in spite of the fact that events naturally evoke some form of concurrency, there is no specific support for concurrency in EScala. It is up to the programmer to understand how to combine declarative events and Scala's support for concurrent programming. We have started working on injecting concurrency at the heart of declarative events so that events can indeed be naturally concurrent [28].

6.2.2.2. Monascheme: modular prototyping of aspect languages

We have then developed Monascheme [21], an extensible aspect-oriented programming language based on monadic aspect weaving. Extensions to the aspect language are defined as monads, enabling easy, simple and modular prototyping. The language is implemented as an embedded language in Racket. We illustrate the approach with an execution level monad and a level-aware exception transformer. Semantic variations can be obtained through monad combinations.

6.2.2.3. Structuring computations with aspect-based membranes

In most aspect-oriented languages, aspects have an unrestricted global view of computation. Several approaches for aspect scoping and more strongly encapsulated modules have been formulated to restrict the power of aspects. Our approach [27] leverages the concept of programmable membranes of Boudol, Schmitt and Stefani, as a means to tame aspects by customizing the semantics of aspect weaving locally. Membranes have the potential to subsume previous proposals in a uniform framework. Because membranes give structure to computation, they enable flexible scoping of aspects; because they are programmable, they enable visibility and safety constraints, both for the advised program and for the aspects. The power and simplicity of membranes open interesting perspectives to unify multiple approaches that tackle the unrestricted power of aspects.

6.3. Cloud applications and infrastructures

Participants: Frederico Alvares, Gustavo Bervian Brand, Yousri Kouki, Adrien Lèbre, Thomas Ledoux, Guillaume Le Louët, Jean-Marc Menaud, Jonathan Pastor, Rémy Pottier, Flavien Quesnel, Mario Südholt.

We have contributed on SLA management for Cloud elasticity, fully distributed and autonomous virtual machine scheduling, and energy-efficient Cloud infrastructures.

6.3.1. SLA Management for Cloud elasticity

In [23], we have introduced *CSLA*, the Cloud Service Level Agreement language. The CSLA language has been influenced by related work, in particular WSLA and SLA@SOI. It allows to describe the SLA between a cloud service provider and a cloud customer. One of the novelties of CSLA is that it integrates features dealing with QoS uncertainty and cloud fluctuations, such as *confidence*, *penalty* and *fuzziness*.

Cloud computing is a model for enabling on-demand access to a shared pool of configurable resources as services. However, the management of such elastic resources is a complex issue. In [24], we have proposed a SLA-driven approach for optimizing the resources capacity planning for Cloud applications. We have modeled Cloud services using a closed queuing network model and proposed an extension of a Mean Value Analysis (MVA) algorithm to take into account the concept of SLA. Then, based on capacity planning method, our solution calculates the optimal configuration of a Cloud application.

6.3.2. *Fully distributed and autonomous virtualized environments*

Extending previous preliminary results of the DVMS prototype, we have consolidated this system to obtain a fully distributed virtual machine scheduler [13]. This system makes it possible to schedule VMs cooperatively and dynamically in large scale distributed systems. Simulations (up to 64K VMs) and real experiments both conducted on the Grid'5000 large-scale distributed system [34] showed that DVMS is scalable. This building block is a first element of a more complete cloud OS, entitled DISCOVERY (DIStributed and COoperative mechanisms to manage Virtual EnviRonments autonomicallY) [66]. The ultimate goal of this system is to overcome the main limitations of the traditional server-centric solutions. The system, currently under investigation in the context of the Jonathan Pastor's PhD, relies on a peer-to-peer model where each agent can efficiently deploy, dynamically schedule and periodically checkpoint the virtual environments it manages.

6.3.3. *Energy-efficient Cloud applications and infrastructures*

As a direct consequence of the increasing popularity of Cloud Computing solutions, data centers are amazingly growing and hence have to urgently face with the energy consumption issue. Available solutions rely on Cloud Computing models and virtualization techniques to scale up/down application 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 applications Quality of Service and energy footprint. We have proposed a self-adaptation approach that considers both application internals and system to reduce the energy footprint in cloud infrastructure [31], [11]. Each application and the infrastructure are equipped with their own control loop, which allows them to autonomously optimize their executions. In addition, these autonomic loops are coordinated to avoid inconsistent states. This coordination improves the synergy between applications and infrastructure in order to optimize the infrastructure energy consumption [16].

We have extended our previous work on Entropy, a virtual machine placement manager, by the development of btrScript, a safe autonomic system for virtual machine management that includes actions and placement rules. Actions are imperative operations to reconfigure the data center and declarative rules specify the virtual machine placement. Administrators schedule both actions and rules, to manage their data center(s). They can also interact with the btrScript system in order to monitor the data center and compute the correct virtual machine placement [25]. btrScript and Entropy have been packaged in a common software btrCloud.

ATLANMOD Team

6. New Results

6.1. Core Modeling technologies

AtlanMod has continued to improve the core model and model transformation technologies that are reused in the other more domain-oriented research lines of the team. Main results in this area have been:

- Model-to-Model Transformation Refactorings. In object-oriented programming, continuous refactorings are used as the main mechanism to increase the maintainability of the code base. Unfortunately, in the field of model transformations, such refactoring support is so far missing. We have tackled this limitation by adapting the notion of refactorings to model-to-model (M2M) transformations. Particularly, in [17] we present a dedicated catalogue of refactorings for improving the quality of M2M transformations. This catalogue is the result of analyzing existing ATL transformations; its scope is beyond ATL, covering other M2M transformation languages.
- Reactive Model Transformations. Model-driven applications manipulate models by executing model transformations that are seen by host applications as black-box functions returning the computed target models. We propose a paradigm shift where a network of reactive transformations defines persistent data-flows among models. A reactive transformation engine takes care of activating only the strictly needed computation in response to updates or requests of model elements. Computation is updated when necessary, in an autonomous and optimized way. The application architecture results deeply changed, since the host application does not directly control the execution of the transformations anymore, but only accesses or updates the underlying models. We experiment this paradigm by implementing a reactive engine for ATL.
- EMF Profiles. There are many situations in which one needs to extend or annotate a model with additional information. Nevertheless, changing the metamodel to include this new information is very costly (e.g. you'll need to recreate the modeling environment and, possibly, to migrate other existing models). As a solution, we have proposed the idea of EMF Profiles [16] as a way to reuse the idea of UML Profiles for general EMF Models. UML profiles have been a key enabler for the success of UML by providing a lightweight language-inherent extension mechanism which is expressive enough to cover an important subset of adaptation scenarios. We believe a similar concept for DSMLs would provide an easier extension mechanism for EMF.

6.2. Domain-Specific languages

In the field of Domain-Specific Languages (DSLs), we have focused on the improvement of the DSLs definition process. During 2012 the new results in this area have been:

- Software development processes are becoming more collaborative, trying to integrate end-users as much as possible. The idea is to advance towards a community-driven process where all actors (both technical and nontechnical) work together to ensure that the system-to-be will satisfy all expectations. This seems specially appropriate in the field of Domain-Specific Languages (DSLs) typically designed to facilitate the development of software for a particular domain. We have designed a collaborative infrastructure for the development of DSLs where end-users have a direct and active participation in the evolution of the language [22], [32]. This infrastructure is based on Collaboro, a DSL to represent change proposals, possible solutions and comments arisen during the development and evolution of a language.

- When developing DSLs, a number of design decisions must be made, such as those related to its concrete syntax, how the language semantics is going to be defined and in which form (interpreted or compiled), or whether there will be an underlying abstract syntax. However, deciding whether the DSL will be internal or external will have an impact on the other aspects of the language. Making an effective choice between these two options therefore requires a careful evaluation of the pros and cons of each alternative. Some important aspects that should be evaluated are the following, which are related to the three elements of a DSL: abstract and concrete syntaxes, and semantics (executability and optimizations), and to quality criteria (extensibility and efficiency) and DSL tooling (tools for developing DSL and tools for using DSL). In [40] we presented the results of this work.

6.3. Model Verification

Guranteeing the correctness of models is a very important element of the MDE infrastructure. We made several contributions to the model verification field in 2012:

- Automated verification of declarative, rule-based model transformations. Having sound transformations is essential, as they are the compilers in MDE. Because transformations are created frequently, e.g., on a per-project basis, it is important that we can check their correctness automatically. We have developed a novel, automatic proof technique based on Satisfiability Modulo Theories (SMT) solving [20] for this, as well as a bounded-search verification approach using relational logic and Alloy [21]. Both Yices and Z3 have been used as SMT solvers for this work.
- Improving EMFtoCSP, the AtlanMod model finder. Model finding is a central, recurring task in MDE. It subsumes both metamodel consistency checking (i.e., metamodel verification) and metamodel instantiation (e.g., test case generation). Even when using a bounded search approach, the underlying research problem is computational hard and calls for flexibile solutions and heuristics. We have generalized and improved the EMFtoCSP model finder (formerly: UMLtoCSP), which is based on Constraint Logic Programming (CLP). It now supports both UML and Ecore (and OCL constraints) and is open for further modeling languages [26]. As the first available MDE model finder, it now supports reasoning over string constraints. Such constraints are common in practical applications of MDE, but none of the existing model finding approaches could handle them. We have developed a flexible string constraint solver (based on multi-head constraint handling rules) that seamlessly integrates into EMTtoCSP [19].

6.4. Model Transformation Testing

White-box testing for model transformations is a technique that involves the extraction of knowledge embedded in the transformation code to generate test models. In [31], we manually extract such knowledge and we represent it in the form of partial models that can drive the generation of highly effective test models. In other works we go a step further and use static analysis to automatically extract testing knowledge from transformation code. We propose two tool-supported methodologies to automatically generate test cases using structural information from a model transformation. In [27] we have developed an approach that optimizes the test coverage while testing rule-based model transformation languages like ATL. The approach is based on analyzing the dependencies among the OCL queries that are used within the transformation code. The methodology in [29] makes use of the metamodel footprinting mechanism, generates partial models representing the testing intent and uses the ALLOY solver to create complete usable models. The experimental results show that a limited amount of white-box information on the model transformation (i.e., our footprints) can provide remarkable improvements on the efficiency of the generated tests.

6.5. Reverse Engineering

Model Driven Reverse Engineering (MDRE), and its applications such as software modernization, is a discipline in which model-driven development (MDD) techniques are used to treat legacy systems. During 2012, Atlanmod has continued working actively on this research area. The main contributions are the following:

Grammar-to-Model Bridging When existing software artifacts are treated in MDRE, they must be first transformed into models to apply MDD techniques such as model transformations. Since most scenarios involve dealing with code in general-purpose programming languages (GPL), the extraction of models from GPL code is an essential task. We designed Grammar-to-Model Transformation Language (Gra2MoL) as a domain-specific language (DSL) tailored to the extraction of models from GPL code. Gra2MoL aims to reduce the effort needed to implement grammarware-MDD bridges, since building dedicated parsers is a complex and time-consuming task. The language also provides a powerful query language which eases the retrieval of scattered information in syntax trees. Moreover, it incorporates extensibility and grammar reuse mechanisms. In [13], Gra2MoL is described in detail and a case study based on the application of the language in the extraction of models from Delphi code is included.

API-to-Model Bridging Software systems usually manage many Application Programming Interfaces (APIs) to access different software assets (e.g., databases, middleware, etc). A MDRE process therefore also normally involves extracting models from legacy artifacts using API. Thus, we devised API2MoL [14], a DSL which allows developers defining technological bridges between the model and the API technologies. API2MoL is, to the best of our knowledge, the first generic proposal to deal with the integration of MDE and APIs which automates the creation of the API-MDE bridge. Our proposal includes a complete prototype of a toolkit focused on Java APIs, although an adaptation of the approach to deal with APIs for other statically-typed object-oriented languages (such as C sharp) could be easily implemented.

Security Information Discovery Most companies information systems are composed by heterogeneous components responsible of hosting, creating or manipulating critical information for the day-to-day operation of the company. Securing this information is therefore one of their main concerns, more particularly specifying Access Control (AC) policies. However, the task of implementing an AC security policy (sometimes relying on several mechanisms) remains complex and error prone as it requires knowing low level and vendor-specific facilities. In this context, discovering and understanding which security policies are actually being enforced by the Information System (IS) becomes critical. Thus, the main challenge consists in bridging the gap between the vendor-dependent security features and a higher-level representation. This representation has to express the policies by abstracting from the specificities of the system components, allowing security experts to better understand the policy and to implement all related evolution, refactoring and manipulation operations in a reusable way. As a first result, in [28] a method to extract AC policies from firewall configuration files is proposed.

Business Rules Discovery In order to react to the ever-changing market, every organization needs to periodically reevaluate and evolve its company policies. These policies must be enforced by its Information System (IS) by means of a set of so-called business rules that drive the system behavior and data. Clearly, policies and rules must be aligned at all times but unfortunately this is a challenging task. In most ISs, the implementation of business rules is scattered among the code so appropriate techniques must be provided for the discovery and evolution of changing business rules. In [24], we describe a MDRE framework aiming at extracting business rules out of Java source code. The use of modeling techniques facilitate the representation of the rules at a higher-abstraction level which enables stakeholders to understand and manipulate them more easily.

Software Modernization Software modernization processes usually follow the well-known horse-shoe model, which provides a framework to integrate different abstraction levels and reverse engineering tools. The Architecture-Driven Modernization (ADM) is an OMG's initiative which aims at defining and standardizing techniques, methods and tools for software modernization. It incorporates the horse-shoe framework as its reference model and uses MDE techniques as the implementation foundation. Since ADM proposes applying the modernization process at the most abstract level, we believe that, to some extent, the ADM initiative has misinterpreted the original horse-shoe model [34].

Legacy Data Federation The fast evolution of technologies (SOA, Cloud, mobile environments), ISs complexity and the growing need for agility require to be able to represent information systems as a whole. In this context, Enterprise Architecture (EA) approaches intend to address all the systems dimensions: software components, associated physical resources, relationships with the companies requirements and business processes, implied actors/roles/structures, etc. Within the TEAP FUI project (cf. corresponding section), we have started studying the reverse engineering capabilities required when dealing with such high-level views of an IS. More particularly, the focus has been put on features for allowing federating the relevant data coming from different existing sources, as well as for integrating them efficiently. To this intent, a prototype is currently being developed based on several technologies from the team (e.g. Virtual EMF, ATL, MoDisco).

6.6. Empirical software modeling

A new line started this year was the evaluation of how software modeling techniques (and in general software engineering methods) are used in practice. As the first area of study, we have focused on how software architects deal with non-functional requirements. Based on a set of interviews with software architects, we have analyzed whether all the languages, patterns and methodologies proposed by researchers have had any impact on the way software architect choose the best architecture for a given system. Results of the study can be read in these publications [18] [11].

CIDRE Project-Team

6. New Results

6.1. Intrusion Detection

6.1.1. Intrusion Detection based on an Analysis of the Flow Control

In 2012, we strengthened our research efforts around intrusion detection parameterized by a security policy.

In [22] we formally study information flows that occur during the executions of a system implementing a classical access control mechanism. More precisely, we detail how the generic access control model we proposed defines two sets of illegal information flows: the first set corresponds to the flows resulting from the accesses authorized by the access control policy while the second set corresponds to the information flow policy deduced from the access control policy interpretation. We show that these two sets may coincide for some policies and we propose a mechanism dedicated to illegal information flow detection that can be useful in other cases. Finally, we describe a real implementation for the Linux operating system.

In [38], we extended our previous illegal information flow detector to track network exchanges. A confidentiality policy is defined by labeling sensitive information and defining which information may leave the local system through network exchanges. Furthermore, per application profiles can be defined to restrict the sets of information each application may access and/or send through the network. An example application of this extension in the context of a compromised web browser showed that our implementation can detect a confidentiality violation when the browser attempts to leak private information to a remote host over the network.

In [30], we adapted our detection model to the Android operating system. Mobile phones nowadays evolve as data repositories in which pieces of data belong to different owners and can or must be protected by different security policies. These pieces of data are used on an open environment controlled by a non-specialist user. The dynamic monitoring of information flows is well adapted for protecting information on an embedded system as a mobile phone. Nevertheless the main difficulty relies on the definition of the information flow policy. We proposed a way to define such a policy for the Android operating system.

6.1.2. Detecting Attacks against Data in Web Applications

In [41] we present RRABIDS (Ruby on Rails Anomaly Based Intrusion Detection System) an application level intrusion detection system for applications implemented with the Ruby on Rails framework. This IDS has been developed in the context of a collaborative project funded by ANR and called DALI.

This work aims at detecting attacks against data in the context of web applications. This anomaly based IDS focuses on the modeling of the application profile in the absence of attacks (called normal 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. We showed on simple examples how the approach detects well known categories of web attacks that involve a state violation of the application, such as SQL injections. An assessment phase was performed to evaluate the accuracy of the detection provided by the proposed approach. We learned two lessons during this assessment. First this approach provides excellent results in term of false negatives. Second it demonstrates the importance of the learning phase in terms of false positives.

6.1.3. Visualization of Security Events

After having performed in the begining of the year an extensive state of the art of the current visualisation tools dedicated to security, it now clearly appears that there is an important lack of proposals in the context of security data analytics: most of the current visualization proposals build representations for real-time monitoring and only a few of them really allow the user to crawl its data sources in details. Due to this fact, we decided to focus on visualization for security data analytics.

We also built a new visualisation platform in order to lead experiments. Our new directions and the platform have been presented in [20].

6.1.4. Intrusion Detection System Assessement

In [32], we present Netzob [1], a tool dedicated to semi-automatic network protocol reverse-engineering. Such a tool is useful to understand proprietary or non-documented protocols, which is often the case in security analysis or security product assessments. Netzob leverages different algorithms from the fields of bio-informatics and automata theory to infer both the vocabulary and the grammar of undocumented protocols. The vocabulary is inferred from message sequences previously captured (network packets, function call traces, etc.) whereas the grammar inference needs a working implementation of the protocol, which is executed in a confined environment and is used as an oracle. The inferred model could be used to automatically build a client or server implementation of the protocol to generate realistic network traffic.

6.2. Privacy

6.2.1. Geoprivacy

Recent advances in geolocated capacities, secure and verified positioning techniques, ubiquitous connectivity, as well as mobile and embedded systems, have led to the development of a plethora of Location-Based Services (LBS), personalizing the services they deliver according to the location of the user querying the service. However, beyond the benefits they provide, users have started to be worried about the privacy breaches caused by such systems. Among all the Personally Identifiable Information (PII), learning the location of an individual is one of the greatest threats against privacy. In particular, an inference attack [19], can use mobility data (together with some auxiliary information) to deduce the points of interests characterizing his mobility, to predict his past, current and future locations [34] or even to identify his social network.

In order to address and mitigate these privacy issues, within the AMORES project [31], we aim at developing an architecture for the provision of privacy-preserving and resilient collaborative services for “mobiquitous” (*i.e.*, mobile and ubiquitous) systems. The project is built around three uses-cases from the area of publication transportation: (1) dynamic carpooling, (2) real-time computation of multimodal transportation itineraries and (3) mobile social networking. Recently, we have introduced the concept of locanym [35], which corresponds to a pseudonym linked to a particular location that could be used as a basis for developing privacy-preserving LBS.

6.2.2. Privacy-enhanced Social Networks

In [49], we have introduced a new research track focusing on the protection of privacy in distributed social networks, which corresponds to the PhD thesis of Regina Paiva Melo Marin. Our first step has been a study of the needs and practices regarding privacy and personal data policies in social networking frameworks. The commonly accepted requirements for general privacy policies are evaluated with respect to the corresponding notions found in European regulations, and then interpreted in the context of social networking applications. One of the main finding of this study is that some of these requirements are not met by the existing social networks (be they widely used or in development, centralized or distributed, focusing on personal data monetization or on user privacy). The concept of *purpose*, as well as the associated notions of minimization, finality and proportionality, in particular, appears to be insufficiently described in the various policy models. Finally, we have proposed a set of minimal requirements that a privacy policy framework designed for distributed social networks should meet for it to be sufficiently expressive with regards to the current regulations.

[1] http://www.netzob.org

6.2.3. Privacy Enhancing Technologies

Even though they integrate some blind submission functionalities, current conference review systems, such as EasyChair and EDAS, do not fully protect the privacy of authors and reviewers, in particular from the eyes of the program chair. As a consequence, their use may cause a lack of objectivity in the decision process. To address this issue, we have proposed in collaboration with researchers from the Université de Montréal, P3ERS (for Privacy-Preserving PEer Review System) [17], a distributed conference review system based on group signatures, which aims at preserving the privacy of all participants involved in the peer review process. One of the main ideas of P3ERS is to ensure the privacy of both the authors and the reviewers (and this even from the point of view of the conference provider and the conference chair) by using two different groups of users. In particular, the authors can submit anonymized papers on behalf of the author group to the program chair, who then dispatches the papers according to the declared skills of the reviewer group members in an oblivious manner. In this way, the program chair knows neither the identity of the authors (until a paper is accepted, if it is) nor the correspondence between papers and reviewers.

In [25], we have considered the setting in which the profile of a user is represented in a compact way, as a Bloom filter, and the main objective is to privately compute in a distributed manner the similarity between users by relying only on the Bloom filter representation. In particular, our main objective is to provide a high level of privacy with respect to the profile even if a potentially unbounded number of similarity computations take place, thus calling for a non-interactive mechanism. To achieve this, we have proposed a novel non-interactive differentially private mechanism called BLIP (for BLoom-and-flIP) for randomizing Bloom filters. This approach relies on a bit flipping mechanism and offers high privacy guarantees while maintaining a small communication cost. Another advantage of this non-interactive mechanism is that similarity computation can take place even when the user is offline, which is impossible to achieve with interactive mechanisms. Another contribution of this work is the definition of a probabilistic inference attack, called the "Profile Reconstruction attack", that can be used to reconstruct the profile of an individual from his Bloom filter representation. More specifically, we provided an analysis of the protection offered by BLIP against this profile reconstruction attack by deriving an upper and lower bound for the required value of the differential privacy parameter ϵ.

In order to contribute to solve the personalization/privacy paradox, we have proposed a privacy-preserving architecture for one of the state of the art recommendation algorithm, Slope One [36]. More precisely, we designed SlopPy (for *Slope One with Privacy*), a privacy-preserving version of Slope One in which a user never releases directly his personal information (*i.e.*, his ratings). Rather, each user first perturbs locally his information by applying a Randomized Response Technique before sending this perturbed data to a semi-trusted entity responsible for storing it. While there is a trade-off to set between the desired privacy level and the utility of the resulting recommendation, our preliminary experiments clearly demonstrate that SlopPy is able to provide a high level of privacy at the cost of a small decrease of utility.

A privacy-preserving identity card is a personal device device that allows its owner to prove some binary statements about himself (such as his right of access to some resources or a property linked to his identity) while minimizing personal information leakage. As a follow-up of previous works, we have discussed a taxonomy of threats against the card. Finally, we also proposed for security and cryptography experts some novel challenges and research directions raised by the privacy-preserving identity card [50].

6.2.4. Privacy and Data Mining

In [44], [33], we have introduced a novel inference attack that we coined as the reconstruction attack whose objective is to reconstruct a probabilistic version of the original dataset on which a classifier was learnt from the description of this classifier and possibly some auxiliary information. In a nutshell, the reconstruction attack exploits the structure of the classifier in order to derive a probabilistic version of dataset on which this model has been trained. Moreover, we proposed a general framework that can be used to assess the success of a reconstruction attack in terms of a novel distance between the reconstructed and original datasets. In case of multiple releases of classifiers, we also gave a strategy that can be used to merge the different reconstructed datasets into a single coherent one that is closer to the original dataset than any of the simple reconstructed datasets. Finally, we gave an instantiation of this reconstruction attack on a decision tree classifier that was

learnt using the algorithm C4.5 and evaluated experimentally its efficiency. The results of this experimentation demonstrate that the proposed attack is able to reconstruct a significant part of the original dataset, thus highlighting the need to develop new learning algorithms whose output is specifically tailored to mitigate the success of this type of attack.

6.2.5. Privacy and Web Services

We have proposed [18] a new model of security policy based for a first part on our previous works in information flow policy and for a second part on a model of Myers and Liskov. This new model of information flow serves web services security and allows a user to precisely define where its own sensitive pieces of data are allowed to flow through the definition of an information flow policy. A novel feature of such policy is that they can be dynamically updated, which is fundamental in the context of web services that allow the dynamic discovery of services. We have also presented an implementation of this model in a web services orchestration in BPEL (Business Process Execution Language) [18].

6.3. Trust

6.3.1. Privacy Preserving Digital Reputation Mechanism

Digital reputation mechanisms have recently emerged as a promising approach to cope with the specificities of large scale and dynamic systems. Similarly to real world reputation, a digital reputation mechanism expresses a collective opinion about a target user based on aggregated feedback about his past behavior. The resulting reputation score is usually a mathematical object, *e.g.* a number or a percentage. It is used to help entities in deciding whether an interaction with a target user should be considered. Digital reputation mechanisms are thus a powerful tool to incite users to trustworthily behave. Indeed, a user who behaves correctly improves his reputation score, encouraging more users to interact with him. In contrast, misbehaving users have lower reputation scores, which makes it harder for them to interact with other users. To be useful, a reputation mechanism must itself be accurate against adversarial behaviors. Indeed, a user may attack the mechanism to increase his own reputation score or to reduce the reputation of a competitor. A user may also free-ride the mechanism and estimate the reputation of other users without providing his own feedback. From what has been said, it should be clear that reputation is beneficial in order to reduce the potential risk of communicating with almost or completely unknown entities. Unfortunately, the user privacy may easily be jeopardized by reputation mechanisms which is clearly a strong argument to compromise the use of such a mechanism. Indeed, by collecting and aggregating user feedback, or by simply interacting with someone, reputation systems can be easily manipulated in order to deduce user profiles. Thus preserving user privacy while computing robust reputation is a real and important issue that we address in ou work [48], [52]. Our proposition combines techniques and algorithms coming from both distributed systems and privacy research domains. Specifically, we propose to self-organize agents over a logical structured graph, and to exploit properties of these graphs to anonymously store interactions feedback. By relying on robust reputation scores functions we tolerate ballot stuffing, bad mouthing and repudiation attacks. Finally, we guarantee error bounds on the reputation estimation score.

6.4. Other Topics Related to Security and Distributed Computing

6.4.1. Network Monitoring and Fault Detection

Monitoring a system is the ability of collecting and analyzing relevant information provided by the monitored devices so as to be continuously aware of the system state. However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. We propose in [29] to address this issue by pushing the monitoring task at the edge of the network. We present a peer-to-peer based architecture, which enables nodes to adaptively and efficiently self-organize according to their ”health” indicators. By exploiting both temporal and spatial correlations that exist between a device and its vicinity,

our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator. We show that the end-to-end detection process, *i.e.*, from the local detection to the management operator reporting, requires a logarithmic number of messages in the size of the network.

6.4.2. Metrics Estimation on Very Large Data Streams

In [27] and [28], we consider the setting of large scale distributed systems, in which each node needs to quickly process a huge amount of data received in the form of a stream that may have been tampered with by an adversary. In this situation, a fundamental problem is how to detect and quantify the amount of work performed by the adversary. To address this issue, we propose AnKLe (for Attack-tolerant eNhanced Kullback-Leibler divergence Estimator), a novel algorithm for estimating the KL divergence of an observed stream compared to the expected one. AnKLe com- bines sampling techniques and information-theoretic methods. It is very efficient, both in terms of space and time complexities, and requires only a single pass over the data stream. Experimental results show that the estimation provided by AnKLe remains accurate even for different adversarial settings for which the quality of other methods dramatically decreases. In [26], considering n as the number of distinct data items in a stream, we show that AnKLe is an (ε, δ)-approximation algorithm with a space complexity $\widetilde{\mathcal{O}}(\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2})$ bits in "most" cases, and $\widetilde{\mathcal{O}}(\frac{1}{\varepsilon} + \frac{n-\varepsilon^{-1}}{\varepsilon^2})$ otherwise. To the best of our knowledge, an approximation algorithm for estimating the Kullback-Leibler divergence has never been analyzed before. We go a step further by considering in [51] the problem of estimating the distance between any two large data streams in small-space constraint. This problem is of utmost importance in data intensive monitoring applications where input streams are generated rapidly. These streams need to be processed on the fly and accurately to quickly determine any deviance from nominal behavior. We present a new metric, the *Sketch ☆-metric*, which allows to define a distance between updatable summaries (or sketches) of large data streams. An important feature of the *Sketch ☆-metric* is that, given a measure on the entire initial data streams, the *Sketch ☆-metric* preserves the axioms of the latter measure on the sketch (such as the non-negativity, the identity, the symmetry, the triangle inequality but also specific properties of the f-divergence or the Bregman one). Extensive experiments conducted on both synthetic traces and real data sets allow us to validate the robustness and accuracy of the *Sketch ☆-metric*.

6.4.3. Robustness Analysis of Large Scale Distributed Systems

In [14] we present an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, and in particular of peer-to-peer systems. When designing such systems, two major issues need to be faced. First, population of these systems evolves continuously (nodes can join and leave the system as often as they wish without any central authority in charge of their control), and second, these systems being open, one needs to defend against the presence of malicious nodes that try to subvert the system. Given robust operations and adversarial strategies, we propose an analytical model of the local behavior of clusters, based on Markov chains. This local model provides an evaluation of the impact of malicious behaviors on the correctness of the system. Moreover, this local model is used to evaluate analytically the performance of the global system, allowing to characterize the global behavior of the system with respect to its dynamics and to the presence of malicious nodes and then to validate our approach. We complete this work by considering in [13], the behavior of a stochastic system composed of several identically distributed, but non independent, discrete-time absorbing Markov chains competing at each instant for a transition. The competition consists in determining at each instant, using a given probability distribution, the only Markov chain allowed to make a transition. We analyze the first time at which one of the Markov chains reaches its absorbing state. When the number of Markov chains goes to infinity, we analyze the asymptotic behavior of the system for an arbitrary probability mass function governing the competition. We give conditions for the existence of the asymptotic distribution and we show how these results apply to cluster-based distributed systems when the competition between the Markov chains is handled by using a geometric distribution.

6.4.4. Secure Multiparty Computation in Dynamic Networks

In [37] in collaboration with researchers from EPFL, we consider the problem of securely conducting a poll in synchronous dynamic networks equipped with a Public Key Infrastructure (PKI). Whereas previous

distributed solutions had a communication cost of $O(n^2)$ in an n nodes system, we present SPP (Secure and Private Polling), the first distributed polling protocol requiring only a communication complexity of $O(nlog^3n)$, which we prove is near-optimal. Our protocol ensures perfect security against a computationally-bounded adversary, tolerates $(1/2?\epsilon)n$ Byzantine nodes for any constant $1/2 > \epsilon > 0$ (not depending on n), and outputs the exact value of the poll with high probability. SPP is composed of two sub-protocols, which we believe to be interesting on their own: SPP-Overlay maintains a structured overlay when nodes leave or join the network, and SPP-Computation conducts the actual poll. We validate the practicality of our approach through experimental evaluations and describe briefly two possible applications of SPP: (1) an optimal Byzantine Agreement protocol whose communication complexity is $\Theta(nlogn)$ and (2) a protocol solving an open question of King and Saia in the context of aggregation functions, namely on the feasibility of performing multiparty secure aggregations with a communication complexity of $o(n^2)$.

6.4.5. Agreement Problems in Unreliable Systems

In distributed systems, replication techniques are used to mask occurrences of accidental and malicious failures. To coordinate efficiently the different replicas, different approaches can be adopted (state machine mechanisms, group communication services, ...). Most solutions are based on agreement protocols. 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. We have proposed a protocol designed to solve several consecutive consensus instances in an asynchronous distributed system prone to crash failures and message omissions. The protocol follows the Paxos approach and integrates two different optimizations to reduce the latency of learning a decision value. As one optimization is risky, dynamics triggering criterion are defined to check at runtime if the context seems to be favorable or not. The proposed protocol 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 [39] focus on the impact of the prediction of collisions (i.e. , the cases where the use of the risky optimization is counterproductive).

We consider also the problem of approximate consensus in mobile ad hoc networks in the presence of Byzantine nodes. Each node begins to participate by providing a real number called its initial value. Eventually all correct nodes must obtain final values that are different from each other within a maximum value denoted ϵ (convergence property) and must be in the range of initial values proposed by the correct nodes (validity property). Due to nodes' mobility, the topology is dynamic and unpredictable. In [40], [53], we propose an approximate Byzantine consensus protocol which is based on the linear iteration method. Each node repeatedly executes rounds. During a round, a node moves to a new location, broadcasts its current value, gathers values from its neighbors, and possibly updates its value. In our protocol, nodes are allowed to collect information during several consecutive rounds: thus moving gives them the opportunity to gather progressively enough values. An integer parameter Rc is used to define the maximal number of rounds during which values can be gathered and stored while waiting to be used. A novel sufficient and necessary condition guarantees the final convergence of the consensus protocol. At each stage of the computation, a single correct node is concerned by the requirement expressed by this new condition (the condition is not universal as it is the case in all previous related works). Moreover the condition considers both the topology and the values proposed by correct nodes. If less than one third of the nodes are faulty, the condition can be satisfied. We are working on mobility scenarios (random trajectories, predefined trajectories, meeting points) to assert that the condition can be satisfied for reasonable values of Rc.

FOCUS Project-Team

6. New Results

6.1. Service-oriented computing

Participants: Mila Dalla Preda, Ornela Dardha, Maurizio Gabbrielli, Elena Giachino, Claudio Guidi, Jacopo Mauro, Fabrizio Montesi, Davide Sangiorgi.

6.1.1. Primitives

In the context of Service-Oriented Architectures (SOAs), the integration of services is an important aspect that is usually addressed by using specific tools, such as Enterprise Service Bus (ESB). Although widely used these ad-hoc solutions do not exploit the possibility of using a mechanism of interface extension to foster the rapid prototyping and deployment of in-the-middle entities that compose services abstracting from the order in which they exchange messages. We have proposed [29] a framework to perform service integration, based on the extension of service interfaces, capturing a class of service integrators that are decoupled from the services they integrate in an SOA. We also provide a reference implementation for the primitive of service integration by extending the Jolie language, thus allowing for the experimentation with real SOA scenarios. We have shown [30] how our methodology differs from the standard practice with ESB.

6.1.2. Contracts and sessions

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.

Contracts specify the expected dialogue in a session and they can be expressed as types, usually called *session types* in this context.

A session type describes communication by specifying type and direction of data exchanged between two parties. When session types and session primitives are added to the syntax of the types and terms of a language, they give rise to additional separate syntactic categories. As a consequence, there may be duplication of efforts in the theory: the proofs of properties must be checked both on ordinary types and on session types. We have shown [32] that this duplication is not necessary, by exhibiting an encoding of (dyadic) session types into ordinary types. Using the encoding, the properties of session types are derived as straightforward corollaries.

We have also studied the problem of handling unexpected or unwanted conditions in sessions, that may change the default execution of distributed communication protocols. We have proposed [14] a global escape mechanism; it can handle such events while preserving compatibility of multiparty conversations. This flexibility enables us to model complex exceptions such as criss-crossing global interactions and error handling for distributed cooperating threads. Guided by multiparty session types, our semantics is proven to provide a termination algorithm for global escapes, as well as further safety properties, such as progress within the session and atomicity of escapes with respect to the subset of involved participants.

6.2. Adaptability and faults

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

6.2.1. Reversibility

We have continued the study of reversibility started in the past years, aimed at developing programming abstractions for reliable distributed systems. We have shown [39] preliminary results on the interplay between reversibility and compensations, which are a main ingredient in many existing techniques for reliability, in particular long running transactions.

We have then applied [43] our reversibility theory to μOz, a concurrent programming language defined by a stack-based abstract machine, and we make it reversible. This is a first step towards the definition of reversible variants of more complex languages. As additional result we show that the memory overhead due to reversibility is optimal as an order of magnitude.

6.2.2. Primitives for adaptable and evolving components

We study primitives for adaptable and evolving components both in an abstract algebraic setting and in a more concrete setting based on the ABS object-oriented language.

We have defined [13] adaptable processes, a concurrent higher-order calculus where processes have a location, and are sensible to actions of update at runtime. This allows us to express a wide range of evolvability patterns. We have also defined [24] a temporal logic over adaptable processes, with examples of the expressiveness of the logic and of its significance in relation to the calculus of adaptable processes.

A different direction has focused on ABS, a concurrent object-oriented language based on futures for asynchronous method invocations and on object groups for concurrency control. We have given [38] an overview of the architectural aspects of ABS: a feature-driven development workflow, a formal notion of deployment components for specifying environmental constraints, and a dynamic component model that is integrated into the language. We have employed an industrial case study to demonstrate how the various aspects work together in practice. In [40] we have focused on the component model and studied techniques allowing safe dynamic reconfiguration. Our approach adds to ABS: i) output ports to represent variability points, ii) critical sections to control when updates of the software can be made and iii) hierarchy to model locations and distribution. These different notions work together to allow dynamic and safe update of a system.

6.2.3. Reconfigurability in the cloud

The cloud is a relevant application domain for FOCUS. We have considered [35] the problem of deploying and (re)configuring resources in a cloud setting, where interconnected software components and services can be deployed on clusters of heterogeneous (virtual) machines that can be created and connected on-the-fly. We introduce a component model to capture similar scenarii from realistic cloud deployments, and instrument automated planning of day-to-day activities such as software upgrade planning, service deployment, elastic scaling, etc. We formalize the model and characterize the feasibility and complexity of configuration achievability.

6.2.4. Delta-Oriented Programming and Software Product Lines

Delta-oriented programming (DOP) provides a technique for implementing Software Product Lines based on modifications (add, remove, modify) to a core program. Unfortunately, such modifications can introduce errors into a program, especially when type signatures of classes are modified in a non-monotonic fashion. To deal with this problem in we have designed [42] a type system for delta-oriented programs based on row polymorphism. This exercise elucidates the close correspondence between delta-oriented programs and row polymorphism.

In [41] we have studied the notion of conflict for a variant of DOP without features, separating out the notions of hard and soft conflict. Specifically, we have defined a language for this subset of DOP and give a precise, formal definition of these notions. We have then extended the type system in [42] to ensure that the computation of a well-typed product will always succeed and has an unambiguous result.

6.3. Resource Control

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

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.

During 2012, we have continued our work on intensionally complete techniques for the complexity analysis of functional programs. In [15] a relatively complete type system from which complexity of call-by-name terms has been introduced, while in [25] the same approach is used in a call-by-value setting. The introduced methodology allows us to reduce the problem at hand to the verification of a set of first-order proof-obligations. No information is lost along the reduction.

The interpretation method has been the object of a couple of investigations. On the one hand, we have proved a necessary condition for a first-order program to admit a quasi-interpretation [12]: it must be blind, namely it must be somehow insensible to its argument value, but only sensible to their length. Moreover, we have introduced a new methodology for the complexity analysis of higher-order programs based on an higher-order generalizations of ordinary polynomial interpretations and quasi-interpretations [23].

Among the most foundational works in this area, we should also mention those about invariance results on cost models [16], [22]: we proved that in many different cases, the number of beta-reduction steps is an adequate cost-model for the lambda calculus. These results are potentially useful for complexity analysis, in that they show that a natural and quite intuitive cost model is indeed reasonable, meaning that evaluation can be simulated by finer-grained models of computation within a polynomial overhead.

Finally, some of our works have to do with the semantics of various sorts of lambda calculi with linearity constraints: a non-deterministic extension of the call-by-value lambda calculus, which corresponds to the additive fragment of the linear-algebraic lambda-calculus [36]; a lambda calculus with constructors that decomposes the pattern matching à la ML into some atomic rules [44]; a categorical approach to model the programming language SlPCF that has been conceived in order to program only linear functions between Coherence Spaces [20].

We have also continued our work on techniques for ensuring termination of programs, studying [34] how to transport techniques initially devised for processes onto sequential higher-order languages with imperative features (e.g., λ-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.

6.4. Verification of extensional properties

Participants: Mario Bravetti, Daniel Hirschkoff, Cosimo Laneve, Jean-Marie Madiot, 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. A substantial amount of the work carried out this year has to do with the transfer of techniques from the area of concurrency theory to the investigation of properties in adaptable systems, obiect-oriented concurrent systems, and systems based on specific synchronization mechanisms.

6.4.1. Adaptability

We mentioned earlier the process calculus of adaptable processes [13] and the related temporal logic [24]. In the same papers, we have addressed the (un)decidability of two safety properties related to error occurrence, and we have explained how the proof techniques in [13] can be extended to prove (un)decidability results for the temporal logic.

6.4.2. Object-orientation

We have considered concurrent object-oriented languages with futures and cooperative scheduling. Verification of deadlock in such systems is a nontrivial task due to the dynamic and unbounded creation of futures. We have introduced [45] a technique to prove deadlock freedom for such systems by translating a concrete program to an abstract version of the program, and then encoding such abstract program into a Petri net. Deadlock can be detected on Petri nets via checking the reachability of a distinct marking: absence of deadlocks in the Petri net constitutes deadlock freedom of the concrete system.

6.4.3. Synchronization primitives

We have investigated [33] the impact of node and communication failures on the decidability and complexity of parametric verification of a formal model of ad hoc networks, in which finite state processes communicate via selective broadcast. We have considered three possible kinds of node failures –intermittence, restart, and crash– and three cases of communication failures –nondeterministic message loss, message loss due to conflicting emissions, and detectable conflicts. Interestingly, we have proved that the considered decision problem (reachability of a control state) is decidable for node intermittence and message loss (either nondeterministic or due to conflicts) while it turns out to be undecidable for node restart/crash, and conflict detection. The conclusion is that verification is decidable only when processes are unaware of the occurrence of a failure.

In another line of work, we have studied the impact of dualities and symmetries in synchronization primitives for message-passing processes [37]. We have shown that in languages where input and output are dualisable (e.g., variants of the π-calculus such as πI and fusion), duality breaks with the addition of ordinary input/output types. We have then considered the minimal symmetrical conservative extension of π-calculus input/output types. We have proved duality properties for it. As example of application of the dualities, we have used this language to relate two encodings of λ-calculus, by Milner and by van Bakel and Vigliotti, syntactically quite different from each other. Thus, results on one encoding can be transferred onto the other one.

6.4.4. Coinduction

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 [48], [51] two comprehensive textbooks on bisimulation and coinduction (in [51], Sangiorgi is an editor, and author of two chapter contributions [49], [47]). 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.

6.5. Expressiveness of computational models

Participants: Cosimo Laneve, Maurizio Gabbrielli, Gianluigi Zavattaro.

Expressiveness refers to the study of the expressive power of computational models.

We have studied [46] the expressiveness of an actor-based language similar to the language ABS developed in Hats. We have identified the presence/absence of fields as a relevant feature: the dynamic creation of names in combination with fields gives rise to Turing completeness. On the other hand, restricting to stateless actors gives rise to systems for which properties such as termination are decidable. Such decidability result holds in actors with states when the number of actors is finite and the state is read-only.

Our other study of expressiveness has been made on Constraint Handling Rules (CHR), a committed-choice declarative language originally designed for writing constraint solvers and that is nowadays a general purpose language. The study of CHR is interesting within Focus as this kind of languages, having both constraint solving and concurrency features, allow us to express in a natural way quantitative aspects related to resources. Moreover, constraints may be used to describe dynamic adaptation and evolution of systems. CHR programs

consist of multi-headed guarded rules which allow one to rewrite constraints into simpler ones until a solved form is reached. Many empirical evidences suggest that multiple heads augment the expressive power of the language (somehow, it can be considered similar to multiple synchronization gathering n processes simultaneously), however no formal result in this direction had been proved so far. We have proved [18] a number of expressiveness results to support such a claim. First, we have analyzed the Turing completeness of CHR with respect to the underlying constraint theory. If the constraint theory is powerful enough then restricting to single head rules does not affect Turing completeness. On the other hand, differently from the case of the multi-headed language, the single head CHR language is not Turing powerful when the underlying signature (for the constraint theory) does not contain function symbols. Then we have proved that, no matter which constraint theory is considered, under some reasonable assumptions it is not possible to encode the CHR language (with multi-headed rules) into a single headed language while preserving the semantics of the programs. Moreover, under some stronger assumptions, considering an increasing number of atoms in the head of a rule augments the expressive power of the language.

INDES Project-Team

6. New Results

6.1. Security

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

6.1.1. Session types with security

We have pursued our work on integrating security constraints within session types, in collaboration with our colleagues from Torino University. This resulted in the journal paper [8]. This article extends a previous conference paper with full proofs, additional examples and further results. In particular, [8] presents new properties of information-flow security, which is stronger and more compositional (*i.e.*, more robust with respect to parallel composition of processes) than that originally proposed, while being still ensured by the same session type system.

All the work on session types was partially funded by the ANR-08- EMER-010 grant PARTOUT. It is expected to continue within the starting COST Action BETTY.

6.1.2. Mashic Compiler: Mashup Sandboxing Based on Inter-frame Communication

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 constraints 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.

This work appeared in CSF'12 [14]. 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 [6].

6.2. Models, semantics, and languages

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

6.2.1. Memory Models

As regards the theory of multithreading, we have extended our operational approach to capture more relaxed memory models than simple write buffering. A step was made in this direction by formalizing the notion of a speculative computation, but this was not fully satisfactory as an operational approach to the theory of memory models: indeed, in the speculative framework one has to reject a posteriori some sequences of executions as invalid. In [13] we have defined a truly operational semantics, by means of an abstract machine, for extremely relaxed memory models like the one of PowerPC. In our new framework, the relaxed abstract machine features a "temporary store" where the memory operations issued by the threads are recorded, in program order. A memory model then specifies the conditions under which a pending operation from this sequence is allowed to be globally performed, possibly out of order. The memory model also involves a "write grain," accounting for architectures where a thread may read a write that is not yet globally visible. Our model is also flexible enough to account for a form of speculation used in PowerPC machines, namely branch prediction. To experiment with our framework, we found it useful to design and implement a simulator that allows us to exhaustively explore all the possible relaxed behaviors of (simple) programs. The main problem was to tame the combinatory explosion due to the massive non-deterministic interleaving of the relaxed semantics. Introducing several optimizations described in [13], we were able to run a large number of litmus tests successfully.

6.2.2. Dynamic Synchronous Language with Memory

We have investigated the language DSLM (Dynamic Synchronous Language with Memory), based on the synchronous reactive model. In DSLM, systems are composed of several sites, each of which runs a number of agents. An agent consists of a memory and a script. This script is made of several parallel components which share the agent's memory. A simple form of migration is provided: agents can migrate from one site to another. Since sites have different clocks, a migrating agent resumes execution at the start of the next instant in the destination site. Communication between a migrating agent and the agents of the destination site occurs via (dynamically bound) events. The language uses three kinds of parallelism: 1) synchronous, cooperative and deterministic parallelism among scripts within an agent, 2) synchronous, nondeterministic and confluent parallelism among agents within a site, and 3) asynchronous and nondeterministic parallelism among sites. Communication occurs via both shared memory and events in the first case, and exclusively via events in the other two cases. Scripts may call functions or modules which are handled in a host language. Two properties are assured by DSLM: reactivity of each agent and absence of data-races between agents. Moreover, the language offers a way to benefit from multi-core and multi-processor architectures, by means of the notion of synchronized scheduler which abstractly models a computing resource. Each site may be expanded and contracted dynamically by varying its number of synchronized schedulers. In this way one can model the load-balancing of agents over a site.

A secure extension of the language DSLM, called DSSLM (Dynamic Secure Synchronous Language with Memory), is currently under investigation. This language uses the same deterministic parallel operator for scripts as DSLM. It adds to DSLM a let operator that assigns a security level to the defined variable. Security levels are also assigned to events and sites, to allow information flow control during interactions and migrations. The study of different security properties (both sensitive and insensitive to the passage of the instants) and of type systems ensuring these properties is currently under way.

6.2.3. jthread

The jthread library (working name) is a Bigloo library featuring threads and mutexes and most notably a deadlock-free locking primitive. The jthread library appears as an alternative to Bigloo's pthread (POSIX threads) library and relies on it for its implementation.

The locking primitive is the following: `(synchronize* ml [:prelock mlp] expr1 expr2 ...)` where ml and mlp are lists of mutexes.

This primitive evaluates the expressions that constitute its body after having locked the mutexes in ml and before unlocking them back. The meaning of the *prelock* argument is to be explained below.

The absence of deadlocks is guaranteed by two complementary mechanisms:

- Each mutex belongs to a *region* defined by the programmer. Regions form a lattice which is inferred at runtime. A thread owning a mutex belonging to region R0 can only lock a mutex belonging to region R1 if R1 is lower than R0 in the lattice. This rule is enforced at runtime and guarantees the absence of deadlocks involving mutexes belonging to different regions.
- Under the previous condition, a thread owning a mutex M1 can lock a mutex M2 belonging to the same region only provided that M2 appeared in the *prelock* list of the `synchronize*` that locked M1. This rule is enforced at runtime and allows a *deadlock-avoiding* scheduling of threads based on previous work by Gérard Boudol and on Lamport's Bakery algorithm.

The library has been implemented. It is currently being integrated to Bigloo and benchmarked. It has not been released yet.

6.3. Web programming

Participants: Zhengqin Luo, Cyprien Nicolas, Tamara Rezk, Bernard Serpette, Manuel Serrano [correspondant].

6.3.1. Reasoning about Web Applications: An Operational Semantics for HOP

We propose a small-step operational semantics to support reasoning about web applications written in the multi-tier language HOP. The semantics covers both server side and client side computations, as well as their interactions, and includes creation of web services, distributed client-server communications, concurrent evaluation of service requests at server side, elaboration of HTML documents, DOM operations, evaluation of script nodes in HTML documents and actions from HTML pages at client side. We also model the browser same origin policy (SOP) in the semantics. We propose a safety property by which programs do not get stuck due to a violation of the SOP and a type system to enforce it. This work appeared in TOPLAS [7].

6.3.1.1. Hiphop

We pursued the development of the Hiphop orchestration language. The first version was written as a DSL with very few connection to Hop. During this year, we changed Hiphop syntax to blend it better with Hop. All Hiphop objects are now Hop values, and thus Hiphop programs can benefit from all Hop featues. The Hiphop development has enabled us to improve Hop stability and quality in client code generation.

We have found a new use-case for Hiphop: Robotics. We are currently working with the Inria Coprin team to pilot their robot using Hop and Hiphop. We have already used Hop to program with low-level motors API (using the Phidget libraries). Hop enabled us to distribute the robot control application over HTTP, in order to control the robot from a smartphone or tablet.

6.3.2. A CPS definition of HipHop

Since the Esterel model is used very dynamically in the HipHop framework, we have begun studying new frameworks of computations. We designed a definition of a HipHop-core, which is similar to Esterel-core, based on continuations. This approach allows a specification close to the implementation. The main problem was to define a predicate assuring the *absence* of a specific signal in the current instant. For this, we have designed a static analysis that predicts, for each program point and for each signal, the number of emissions remaining to be done until the end of the instant. The prediction may be over-estimated but when a null value is reached the corresponding signal can be considered as absent for the analyzed instant.

Contrary to existing analyses, this prediction can be done at compile time. Nevertheless some extra computations must be inserted in the evaluator to adjust a runtime prediction. For example. this is done when one branch of a conditional is dynamically taken, but adjusting the prediction only involves subtractions on global counters.

The continuation based definition doesn't prevent a space efficiency implementation. Esterel is known to be compiled to hardware and thus able to run a program in a fixed space of silicon; in the same manner we have implemented an evaluator than doesn't allocate extra memories while running a program: all the continuations can be allocated at compile time.

We have also extended the language to reach the HipHop definition. Some dynamic extensions (`mappar`) may dynamically allocate some resources but we were able to tune the static analysis to insure both confluence and constructive absence detection.

LOGNET Team

5. New Results

5.1. A Backward-Compatible Protocol for Inter-routing over Heterogeneous Overlay Networks

Participants: Giang Ngo Hoang, Luigi Liquori, Vincenzo Ciancaglini, Petar Maksimovic, Hung Nguyen Chan [HUST, Vietnam].

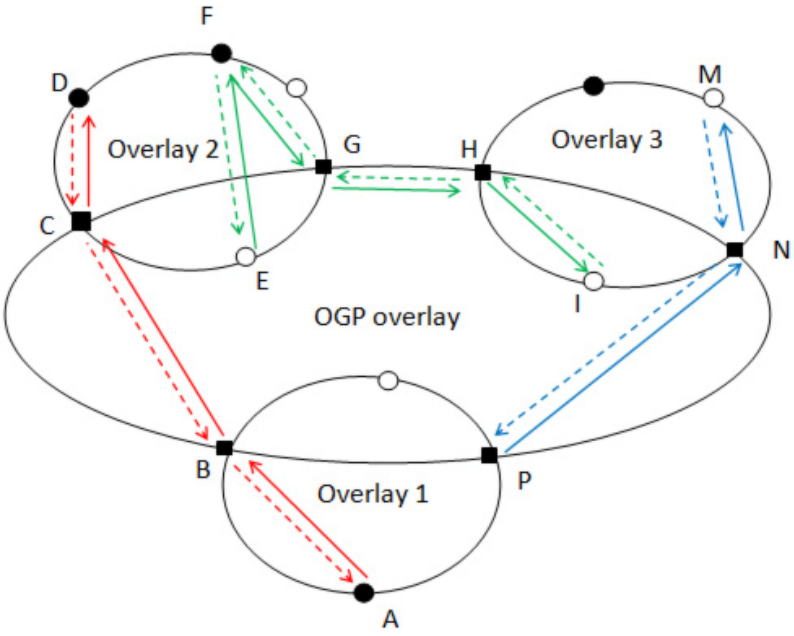

Figure 11. An Overlay Gateway Protocol Topology

Overlay networks are logical networks running on the highest level of the OSI stack: they are applicative networks used by millions of users everyday. In many scenarios, it would be desirable for peers belonging to overlays running different protocols to communicate with each other and exchange certain information. However, due to differences in their respective protocols, this communication is often difficult or even impossible to be achieved efficiently, even if the overlays are sharing common objectives and functionalities. In this paper, we address this problem by presenting a new overlay protocol, called OGP (Overlay Gateway Protocol), allowing different existing networks to route messages between each other in a backward-compatible fashion, by making use of specialized peers joined together into a super-overlay. Experimental results on a large scale Grid5000 infrastructure show that having only a small number of nodes running the OGP protocol is sufficient for achieving efficient routing between heterogeneous overlay networks.

The three scenarios in Figure 11 are shown to illustrate the routing of three lookup queries, in which full OGP peers, lightweight OGP peers and blind peers interact in order to reach across overlays represent requests, while dashed lines represent responses. using the OGP super-overlay. The three smaller ovals represent standard overlays, while the largest oval represents the OGP super-overlay, forwarding messages back and forth between standard overlays. The black squares B; C; G; N and P represent full OGP peers, the black circles A; D and F represent lightweight OGP peers, while the white circles E; H, and M represent blind peers. Solid lines requests, while dashed lines represent responses. The paper is the continuation of the work of HotPost 2011 [7] and it has been accepted to ACM SAC 2013 [33] and a long version will be submitted in a high level conference [34].

5.2. Interconnection of large scale unstructured P2P networks: modeling and analysis

Participants: Rossano Gaeta [Univ. Turin], Riccardo Loti, Luigi Liquori, Vincenzo Ciancaglini [contact].

Interconnection of multiple P2P networks has recently emerged as a viable solution to increase system reliability and fault-tolerance as well as to increase resource availability. In this paper we consider interconnection of large scale unstructured P2P networks by means of special nodes (called *Synapses*) that are co-located in more than one overlay. Synapses act as *trait d'union* by sending/forwarding a query to all the P2P networks they belong to. Modeling and analysis of the resulting interconnected system is crucial to design efficient and effective search algorithms and to control the cost of interconnection. To this end, we develop a generalized random graph based model that is validated against simulations and it is used to investigate the performance of search algorithms for different interconnection costs and to provide some insight in the characteristics of the interconnection of a large number of P2P networks. To overcome this strong limitation, we develop a generalized random graph based model to represent the topology of one unstructured P2P network, the partition of nodes into Synapses, the probabilistic flooding based search algorithms, and the resource popularity. We validate our model against simulations and prove that its predictions are reliable and accurate. We use the model to investigate the performance and the cost of different search strategies in terms of the probability of successfully locating at least one copy of the resource and the number of queries as well as the interconnection cost. We also gain interesting insights on the dependency between interconnection cost and statistical properties of the distribution of Synapses. Finally, we show that thanks to our model we can analyze the performance of a system composed of a large number of P2P networks.

To the best of our knowledge, this is the first paper on model-based analysis of interconnection of large scale unstructured P2P networks [27], [28]

5.3. SIEVE: a distributed, accurate, and robust technique to identify malicious nodes in data dissemination on MANET

Participants: Rossano Gaeta [Univ. Turin], Riccardo Loti [contact], Marco Grangetto [Univ Turin].

We consider the following problem: nodes in a MANET must disseminate data chunks using rateless codes but some nodes are assumed to be malicious, i.e., before transmitting a coded packet they may modify its payload. Nodes receiving corrupted coded packets are prevented from correctly decoding the original chunk. We propose SIEVE, a fully distributed technique to identify malicious nodes.

SIEVE is based on special messages called *checks* that nodes periodically transmit. A check contains the list of nodes identifiers that provided coded packets of a chunk as well as a flag to signal if the chunk has been corrupted. SIEVE operates on top of an otherwise reliable architecture and it is based on the construction of a *factor graph* obtained from the collected checks on which an incremental belief propagation algorithm is run to compute the probability of a node being malicious. Analysis is carried out by detailed simulations using ns-3. We show that SIEVE is very accurate and discuss how nodes speed impacts on its accuracy. We also show SIEVE robustness under several attack scenarios and deceiving actions. The paper has been accepted to [20]

5.4. CCN-TV: a data-centric approach to real-time video services

Participants: Luigi Liquori, Vincenzo Ciancaglini [contact], Riccardo Loti, Giuseppe Piro [Politech Bari], Alfredo Grieco [Politech Bari].

Content Centric Networking is a promising data- centric architecture, based on in-network caching, name-driven routing, and receiver-initiated sessions, which can greatly enhance the way Internet resources are currently used, thus making the support for a broader set of users with increasing traffic demands possible. The CCN vision is, currently, attracting the attention of many researchers across the world, because it has all the potential to become ready to the market, to be gradually deployed in the Internet of today, and to facilitate a graceful transition from a host-centric networking rationale to a more effective data-centric working behavior. At the same time, several issues have to be investigated before CCN can be safely deployed at the Internet scale. They include routing, congestion control, caching operations, name-space planning, and application design. With reference to application-related facets, it is worth to notice that the demand for TV services is growing at an exponential rate over the time, thus requiring a very careful analysis of their performance in CCN architectures. To this end, in the present contribution we deploy a CCN-TV system, able to deliver real-time streaming TV services and we evaluate its performance through a simulation campaign based on real topologies. The paper has been accepted to [19].

5.5. Towards a Trust and Reputation Framework for Social Web Platforms and @-economy

Participants: Thao Nguyen [contact], Bruno Martin [Unice], Luigi Liquori, Karl Hanks.

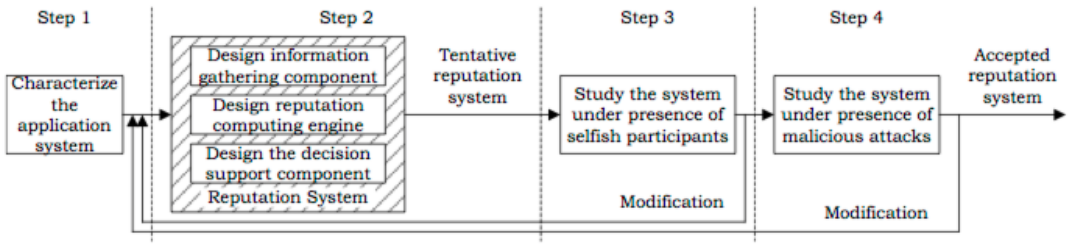

Figure 12. Process of designing a robust trust and reputation system

Trust and reputation systems (TRSs) have recently seen as a vital asset for the safety of online interaction environment. They are present in many practical applications, e.g., e-commerce and social web. A lot of more complicated systems in numerous disciplines also have been studied and proposed in academia. They work as a decision support tool for participants in the system, helping them decide whom to trust and how trustworthy the person is in fulfilling a transaction. They are also an effective mechanism to encourage honesty and cooperation among users, resulting in healthy online markets or communities. The basic idea is to let parties rate each other so that new public knowledge can be created from personal experiences. The greatest challenge in designing a TRS is making it robust against malicious attacks. In this paper, we provide readers an overview on the research topic of TRSs, propose a consistent research agenda in studying and designing a robust TRS, and present an implemented reputation computing engine alongside simulation results, which is our preliminary work to acquire the target of a trust and reputation framework for social web applications.

Information concerning the reputation of individuals has always been spread by word-of-mouth and has been used as an enabler of numerous economic and social activities. Especially now, with the development of technology and, in particular, the Internet, reputation information can be broadcast more easily and faster than ever before. Trust and Reputation Systems (TRSs) have gained the attention of many information and computer scientists since the early 2000s. TRSs have a wide range of applications and are domain specific. The multiple areas where they are applied, include social web platforms, e-commerce, peer-to-peer networks, sensor networks, ad-hoc network routing, and so on. Among these, we are most interested in social web platforms. We observe that trust and reputation is used in many online systems, such as online auction and shopping websites, including eBay, where people buy and sell a broad variety of goods and services, and Amazon, which is a world famous online retailer. Online services with TRSs provide a better safety to their users. A good TRS can also create incentives for good behavior and penalize damaging actions. Markets with the support of TRSs will be healthier, with a variety of prices and quality of service. TRSs are very important for an online community, with respect to the safety of participants, robustness of the network against malicious behavior and for fostering a healthy market.

From a functional point of view, a TRS can be split into three components The first component gathers feedback on participants' past behavior from the transactions that they were involved in. This component includes storing feedback from users after each transaction they take part in. The second component computes reputation scores for participants through a Reputation Computing Engine (RCE), based on the gathered information. The third component processes the reputation scores, implementing appropriate reward and punishment policies if needed, and representing reputation scores in a way which gives as much support as possible to users' decision-making. A TRS can be centralized or distributed. In centralized TRSs, there is a central authority responsible for collecting ratings and computing reputation scores for users. Most of the TRSs currently on the Internet are centralized, for example the feedback system on eBay and customer reviews on Amazon. On the other hand, a distributed TRS has no central authority. Each user has to collect ratings and compute reputation scores for other users himself. Almost all proposed TRSs in the literature are distributed.

Some of the main unwanted behaviors of users that might appear in TRSs are: *free riding* (people are usually not willing to give feedback if they are not given an incentive to do so), *untruthful rating* (users give incorrect feedback either because of malicious intent or because of unintended and uncontrolled variables), *colluding* (a group of users coordinate their behavior to inflate each other's reputation scores or bad-mouth other competitors. Colluding motives are only clear in a specific application), *whitewashing* (a user creates a new identity in the system to replace his old one when the reputation of the old one has gone bad), *milking reputation* (at first, a participant behaves correctly to get a high reputation and then turns bad to make a profit from their high reputation score). The milking reputation behavior is more harmful to social network services and e-commerce than to the others.

This research aims to build on these studies and systematize the process of designing a TRS in general as depicted in Fig. 12 . First, we characterize the application system into which we want to integrate a TRS, and find and identify new elements of information which substitute for traditional signs of trust and reputation in the physical world. Second, based on the characteristics of the application, we find suitable working mechanisms and processes for each component of the TRS. This step should answer the following questions: "What kind of information do we need to collect and how?", "How should the reputation scores be computed using the collected information?", and "How should they be represented and processed to lead users to a correct decision?". To answer the first question, which corresponds to the information gathering component, we should take advantage of information technology to collect the vast amounts of necessary data. An RCE should meet these criteria: *accuracy* for long-term performance (distinguishing a newcomer with unknown quality from a low-quality participant who has stayed in the system for a long time), *weighting* towards recent behavior, *smoothness* (adding any single rating should not change the score significantly), and *robustness* against attacks. Third, we study the tentative design obtained after the second step in the presence of selfish behaviors. During the third step, we can repeatedly return to Step 2 whenever appropriate until the system reaches a desired performance. The fourth step will refine the TRS and make it more robust against malicious attacks. If a modification is made, we should return to Step 2 and check all the conditions in steps 2 and 3 before accepting the modification. The paper has been accepted to [19]

5.6. An Open Logical Framework

Participants: Luigi Liquori [contact], Marina Lenisa [Univ. Udine], Furio Honsell [Univ. Udine], Petar Maksimovic, Ivan Scagnetto [Univ. Udine].

The LFP Framework is an extension of the Harper-Honsell-Plotkin's Edinburgh Logical Framework LF with external predicates, hence the name Open Logical Framework. This is accomplished by defining lock type constructors, which are a sort of "diamond"-modality constructors, releasing their argument under the condition that a possibly external predicate is satisfied on an appropriate typed judgement. Lock types are defined using the standard pattern of constructive type theory, i.e. via introduction, elimination, and equality rules. Using LFP, one can factor out the complexity of encoding specific features of logical systems which would otherwise be awkwardly encoded in LF, e.g. side-conditions in the application of rules in Modal Logics, and sub-structural rules, as in non-commutative Linear Logic. The idea of LFP is that these conditions need only to be specified, while their verification can be delegated to an external proof engine, in the style of the Poincaré Principle or Deduction Modulo. Indeed such paradigms can be adequately formalized in LFP. We investigate and characterize the meta-theoretical properties of the calculus underpinning LFP: strong normalization, confluence, and subject reduction. This latter property holds under the assumption that the predicates are well-behaved, i.e. closed under weakening, permutation, substitution, and reduction in the arguments. Moreover, we provide a canonical presentation of LFP, based on a suitable extension of the notion of $\beta\eta$-long normal form, allowing for smooth formulations of adequacy statements.

LFP is parametric over a potentially unlimited set of (well-behaved) predicates P, which are defined on derivable typing judgements of the form $\Gamma \vdash_\Sigma N : \sigma$, see Fig 13 .

$$\frac{\Gamma \vdash_\Sigma M : \rho \qquad \Gamma \vdash_\Sigma N : \sigma}{\Gamma \vdash_\Sigma \mathcal{L}^{\mathcal{P}}_{N,\sigma}[M] : \mathcal{L}^{\mathcal{P}}_{N,\sigma}[\rho]} \quad \text{(O·Lock)}$$

$$\frac{\Gamma \vdash_\Sigma M : \mathcal{L}^{\mathcal{P}}_{N,\sigma}[\rho] \qquad \Gamma \vdash_\Sigma N : \sigma \qquad \mathcal{P}(\Gamma \vdash_\Sigma N : \sigma)}{\Gamma \vdash_\Sigma \mathcal{U}^{\mathcal{P}}_{N,\sigma}[M] : \rho} \quad \text{(O·Unlock)}$$

Figure 13. Some rule of the Open Logical Framework

The syntax of LFP predicates is not specified, with the main idea being that their truth is to be verified via a call to an external validation tool; one can view this externalization as an oracle call. Thus, LFP allows for the invocation of external "modules" which, in principle, can be executed elsewhere, and whose successful verification can be acknowledged in the system via L-reduction. Pragmatically, lock types allow for the factoring out of the complexity of derivations by delegating the {checking, verification, computation} of such predicates to an external proof engine or tool. The proof terms themselves do not contain explicit evidence for external predicates, but just record that a verification {has to be (lock), has been successfully (unlock)} carried out. In this manner, we combine the reliability of formal proof systems based on constructive type theory with the efficiency of other computer tools, in the style of the Poincaré Principle. In this paper, we develop the meta-theory of LFP. Strong normalization and confluence are proven without any additional assumptions on predicates. For subject reduction, we require the predicates to be well-behaved, i.e. closed under weakening, permutation, substitution, and $\beta\mathcal{L}$-reduction in the arguments. LFP is decidable, if the external predicates are

decidable. We also provide a canonical presentation of LFP, based on a suitable extension of the notion of $\beta\eta$-long normal form. This allows for simple proofs of adequacy of the encodings. In particular, we encode in LFP the call-by-value λ-calculus and discuss a possible extension which supports the design-by-contract paradigm. We provide smooth encodings of side conditions in the rules of Modal Logics, both in Hilbert and Natural Deduction styles. We also encode sub-structural logics, i.e. non-commutative Linear Logic. We also illustrate how LFP can naturally support program correctness systems and Hoare-like logics. In our encodings, we utilize a library of *external predicates*. As far as expressiveness is concerned, LFP is a stepping stone towards a general theory of shallow vs deep encodings, with our encodings being shallow by definition. Clearly, by Church's thesis, all external decidable predicates in LFP can be encoded, possibly with very deep encodings, in standard LF. It would be interesting to state in a precise categorical setting the relationship between such deep internal encodings and the encodings in LFP. LFP can also be viewed as a neat methodology for separating the logical-deductive contents from, on one hand, the verification of structural and syntactical properties, which are often needlessly cumbersome but ultimately computable, or, on the other hand, from more general means of validation.

MYRIADS Project-Team

6. New Results

6.1. Autonomous Management of Virtualized Infrastructures

Participants: Amine Belhaj, Alexandra Carpen-Amarie, Roberto-Gioacchino Cascella, Stefania Costache, Djawida Dib, Florian Dudouet, Eugen Feller, Piyush Harsh, Rémy Garrigue, Filippo Gaudenzi, Ancuta Iordache, Yvon Jégou, Sajith Kalathingal, Christine Morin, Anne-Cécile Orgerie, Nikos Parlavantzas, Yann Radenac.

6.1.1. Application Deployment in Cloud Federations

Participants: Roberto-Gioacchino Cascella, Florian Dudouet, Piyush Harsh, Filippo Gaudenzi, Yvon Jégou, Christine Morin.

The move of users and organizations to Cloud computing will become possible when they will be able to exploit their own applications, applications and services provided by cloud providers as well as applications from third party providers in a trustful way on different cloud infrastructures. In the framework of the Contrail European project [17], we have designed and implemented the Virtual Execution Platform (VEP) service in charge of managing the whole life cycle of OVF distributed applications under Service Level Agreement rules on different infrastructure providers [43]. In 2012, we designed the CIMI inspired REST-API for VEP 2.0 with support for Constrained Execution Environment (CEE), advance reservation and scheduling service, and support for SLAs [40], [29], [32]. We integrated support for delegated certificates and provided test scripts to the Virtual Infrastructure Network (VIN) team. VEP 1.1 was slightly modified to integrate the usage control (Policy Enforcement Point (PEP)) solution developed by CNR. Work is in full progress to implement the CEE management interface and a complete web-based platform for all tasks.

6.1.2. Energy Management in IaaS Clouds: A Holistic Approach

Participants: Eugen Feller, Christine Morin.

Energy efficiency has now become one of the major design constraints for current and future cloud data center operators. One way to conserve energy is to transition idle servers into a lower power-state (e.g. suspend). Therefore, virtual machine (VM) placement and dynamic VM scheduling algorithms are proposed to facilitate the creation of idle times. However, these algorithms are rarely integrated in a holistic approach and experimentally evaluated in a realistic environment. We have designed overload and underload detection and mitigation algorithms and implemented them as well as a modified version of the Sercon existing consolidation algorithm [69] and power management algorithms and mechanisms in a novel holistic energy-efficient VM management framework for IaaS clouds called Snooze [25], [39]. In collaboration with David Margery and Cyril Rohr, we have conducted an extensive evaluation of the energy and performance implications of our system on 34 power-metered machines of the Grid'5000 experimentation testbed under dynamic web workloads. The results show that the energy saving mechanisms allow Snooze to dynamically scale data center energy consumption proportionally to the load, thus achieving substantial energy savings with only limited impact on application performance [26], [48]. Snooze has been released as an open source software since May 2012. It will be further developed and maintained as part of the Snooze ADT. This work has been carried out in the framework of Eugen Feller's PhD thesis [24], [8] funded by the ECO-GRAPPE ANR project.

6.1.3. A Case for Fully Decentralized Dynamic VM Consolidation in Clouds

Participants: Eugen Feller, Christine Morin.

One way to conserve energy in cloud data centers is to transition idle servers into a power saving state during periods of low utilization. Dynamic virtual machine (VM) consolidation (VMC) algorithms are proposed to create idle times by periodically repacking VMs on the least number of physical machines (PMs). Existing works mostly apply VMC on top of centralized, hierarchical, or ring-based system topologies, which result in poor scalability and/or packing efficiency with increasing number of PMs and VMs. We have proposed a novel fully decentralized dynamic VMC schema based on an unstructured peer-to-peer (P2P) network of PMs. The proposed schema is validated using three well known VMC algorithms: First-Fit Decreasing (FFD), Sercon, V-MAN, and a novel migration-cost aware ACO-based algorithm we have designed. Extensive experiments performed on the Grid'5000 testbed show that once integrated in our fully decentralized VMC schema, traditional VMC algorithms achieve a global packing efficiency very close to a centralized system. Moreover, the system remains scalable with increasing numbers of PMs and VMs. Finally, the migration-cost aware ACO-based algorithm outperforms FFD and Sercon in the number of released PMs and requires less migrations than FFD and V-MAN [23], [47]. This work has been done in the context of Armel Esnault's Master internship [57].

6.1.4. Market-Based Automatic Resource and Application management in the Cloud

Participants: Stefania Costache, Nikos Parlavantzas, Christine Morin.

Themis is a market-based Platform-as-a-Service system for private clouds. Themis dynamically shares resources between competing applications to ensure a fair resource utilization in terms of application priority and actual resource needs. Resources are allocated through a proportional-share auction while autonomous controllers apply elasticity rules to scale application demand according to resource availability and user priority. Themis provides users the flexibility to adapt controllers to their application types, and thus it can support diverse application types and performance goals. We have evaluated Themis through simulation and the obtained results demonstrated the effectiveness of the market-based mechanism[19], [20]. We have recently improved Themis in three ways. First, we extended the resource allocation algorithms to support multiple resources (CPU and memory) and to perform load-balancing between physical nodes while considering the migration cost. Second, we improved the management of applications. We added generic support for virtual cluster deployment, configuration and runtime management and also for application monitoring. Finally, we implemented several adaptation policies to scale elastically applications in term of number of provisioned virtual machines and in term of allocated CPU and memory per virtual machine. Themis is implemented in Python and uses OpenNebula for virtual machine operations. We used Themis to scale elastically two resource management frameworks (Torque and Condor) according to their current workload and also MPI scientific codes according to user-given deadlines. Themis has been deployed on Grid'5000 and also on EDF's testbed, HPSLAB. This work is carried out in the fraemwork of Stefania Costache's PhD thesis.

6.1.5. Autonomous PaaS-level resource management

Participants: Djawida Dib, Christine Morin, Nikos Parlavantzas.

PaaS providers host client applications on provider-owned resources or resources leased from public IaaS clouds. The providers have service-level agreements (SLAs) with their clients specifying application quality requirements and prices. A main concern for providers is sharing their private and leased resources among client applications in order to reduce incurred costs. We have proposed a PaaS architecture based on multiple elastic virtual clusters (VCs), each associated with a specific application type (e.g., batch, MapReduce). The VCs dynamically share the private resources using a decentralised allocation scheme and, when necessary, lease remote resources from public clouds. Resource allocation is guided by the SLAs of hosted applications and resource costs. We have implemented a prototype of this architecture that supports batch and MapReduce applications; the application SLAs constrain completion times and prices. The prototype is currently being evaluated on Grid'5000. This work is performed as part of Djawida Dib's thesis.

6.1.6. Elastic MapReduce on Top of Multiple Clouds

Participants: Ancuta Iordache, Yvon Jégou, Christine Morin, Nikos Parlavantzas.

We have worked on the design and implementation of Resilin. To the best of our knowledge Resilin is the first system which is capable of leveraging resources distributed across multiple potentially geographically distinct locations. Unlike the Amazon s proprietary Elastic Map Reduce (EMR) system, Resilin allows users to perform MapReduce computations across a wide range of resources from private, community, and public clouds such as Amazon EC2. Indeed, Resilin can be deployed on top of most of the open-source and commercial IaaS cloud management systems. Once deployed, Resilin takes care of provisioning Hadoop clusters and submitting MapReduce jobs thus allowing the users to focus on writing their MapReduce applications rather than managing cloud resources. In 2012 we designed and implemented a new version of Resilin based on a service-based architecture, which enables system recovery from errors and can be easily extended and maintained. Important functionalities were added to the system: scaling down the platform, deployment of data analysis systems (Apache Hive, Apache Pig). We have also started to work on the design of policies and mechanisms for the autonomous scaling of the virtual Hadoop clusters managed by Resilin. We perfomed an extensive experimental evaluation of Resilin on top of Nimbus and OpenNebula clouds deployed on multiple clusters of the Grid 5000 experimentation testbed. Our results show that Resilin enables the execution of MapReduce jobs across geographically distributed resources with only a limited impact on the jobs execution time, which is the result of intercloud network latencies [51], [31]. Resilin has been released as an open source software since September 2012. This work was carried out in the framework of the RMAC EIT ICT Labs activity.

6.1.7. Adaptation of the CooRM architecture into XtreemOS

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

In the framework of the COOP ANR project, we have mainly worked on the adaptation and on the implementation of the CooRM architecture (resulting from the work of the Avalon team at Inria Grenoble - Rhône Alpes in the context of the COOP project) into XtreemOS. The main results include a first version of the design of a decentralized version of CooRM, the modification of XtreemOS to support distributed applications (tested with OpenMPI and MPICH2), and the implementation of a launcher of moldable MPI applications using the modified XtreemOS API. A demonstration was presented to the COOP consortium in December 2012.

To get an operational prototype for evaluation purposes, we also had to fix many bugs in XtreemOS, revise its build chain, help clean the distribution package dependencies in collaboration with Rémy Garrigue (engineer from the ADT XtreemOS Easy), rewrite the code generator, help fix issues related to configuration commands in collaboration with Amine Belhaj (engineer from ADT XtreemOS Easy).

6.1.8. Extending a Grid with Virtual Resources Provisioned from IaaS Clouds

Participants: Amine Belhaj, Alexandra Carpen-Amarie, Rémy Garrigue, Sajith Kalathingal, Yvon Jégou, Christine Morin, Yann Radenac.

XtreemOS is a Grid operating system designed to facilitate the execution of grid applications by aggregating resources on multiple sites. XtreemOS provides virtual organization support and enables Grid users to run applications on the resources made available by their virtual organization. As the number of scientific applications that need access to Grid platforms increases, as well as their requirements in terms of processing power, the limited amount of resources that XtreemOS gathers from its virtual organizations may become a bottleneck. To address this limitation, we extended XtreemOS with the capability to acquire virtual resources from cloud service providers. To this end, we enable XtreemOS to provision and configure cloud resources both on behalf of a user and of a virtual organization. This can be done either on-demand, when a user specifically requires cloud resources, or in a dynamic fashion, when the local grid resources cannot comply with the application needs. Furthermore, we devised a selection mechanism for the cloud service providers, allowing users to rent resources from the providers that best match the requirements of their applications. We implemented our approach as a set of extension modules for XtreemOS and we evaluated the prototype in Grid'5000, using cloud resources provisioned from a private OpenNebula cloud. For this evaluation, we made a extensive use of tools developed jointly by Ascola and Myriads project-teams to easily manage large number of VMs on top of IaaS cloud management software (e.g. OpenNebula, Nimbus, OpenStack) deployed on the Grid'5000 platform. This work was carried out as part of the ANR Cloud project [60], [58] and an EIT ICT Labs activity.

6.1.9. Data Management Frameworks for Scientific Applications in Cloud Environments

Participants: Eugen Feller, Christine Morin.

During Eugen Feller's internship at LBNL, we have worked with Lavanya Ramakrishnan from the Advanced Computing for Science department on the evaluation of Hadoop MapReduce jobs in a virtualized environment. We have investigated the performance and power consumption of scientific MapReduce jobs executed in an environment with separated Hadoop compute and data nodes. This enables data sharing across multiple users and is key to support elastic MapReduce. Snooze cloud management stack was used to manage the VMs. Preliminary experimental results on top of Snooze demonstrate the feasibility of our approach.

6.1.10. Energy Consumption Models and Predictions for Large-scale Systems

Participant: Christine Morin.

We have collaborated with Taghrid Samak from the Advanced Computing for Science department at LBNL on the initial investigation of energy consumption models for Grid'5000 sites using Pig and Hadoop, and data from 6 months logs on 135 resources in the Lyon site. The initial results investigate time-series summarization for the entire dataset. For each resource the average power consumption is evaluated and compared with statistically estimated thresholds. A paper is under preparation.

6.1.11. Management of Large Data Sets

Participant: Christine Morin.

Moderate Resolution Imaging Spectroradiometer (MODIS) aboard NASAs satellites continuously generates data important to many scientific analyses. A dataprocessing pipeline that downloads the MODIS products, reprojects them on HPC systems or clouds and make them available to users through a web portal has been developed. In collaboration with Valerie Hendrix and Lavanya Ramakrishnan from the Advanced Computing for Science department at LBNL we have worked on providing community access to MODIS Satellite Reprojection and Reduction Pipeline and Data Sets. In a future version of the system, users will be able to reproject data on demand and/or run algorithms on the reprojected MODIS data such as an evapotranspiration calculation [30].

6.2. Dynamic Adaptation of Service-based Applications

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

6.2.1. Adaptation for Service-Oriented Architectures

Participants: 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.

6.2.1.1. Services adaptation in distributed and heterogeneous systems

Participants: Erwan Daubert, Guillaume Gauvrit, Jean-Louis Pazat.

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.

6.2.1.2. Quality Assurance for Distributed Services

Participants: André Lage, Nikos Parlavantzas, Jean-Louis Pazat.

In the context of the service-centric paradigm, we have designed and developed the Qu4DS (Quality Assurance for Distributed Services) system. Qu4DS is a cloud PaaS solution which fills the gap between SaaS service providers and IaaS infrastructures. Qu4DS provides automatic support for service execution management, aiming at increasing service providers' profits by reducing resource costs as well as fines owning to SLA violations. More specifically, Qu4DS dynamically acquires resources according to the customer demand, deploys service instances and implements QoS assurance mechanisms in order to prevent SLA violations. Qu4DS has been evaluated on Grid'5000 and shown to be effective in reducing service provider's costs [33]. This work has been done in the context of André Lage-Freitas' PhD thesis [10].

6.2.1.3. Self-configuration for Cloud Platforms

Participants: Jean-Louis Pazat, Chen Wang.

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.

6.2.1.4. Dynamic Adaptation of Chemical services

Participants: Jean-Louis Pazat, Chen Wang.

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. This year, we have studied and implemented fragment replacement in workflows within this environment.

6.2.2. Multi-level Adaptation for Distributed Operating Systems

Participants: Djawida Dib, Christine Morin, Nikos Parlavantzas.

This work focused on enhancing distributed operating systems with the ability to continually adapt to their changing environments. Two challenges arise in this context: how to design the distributed operating system (OS) in order to facilitate dynamic adaptation, and how to ensure that OS-level adaptation does not conflict with application-level adaptation. This work proposed to address these challenges by (1) building the distributed OS as an assembly of adaptable services following the service-oriented architecture; and (2) using a common multi-level adaptation framework to adapt both the OS and the application layers in a coordinated way. To demonstrate the usefulness of the proposed architecture, the work focused on distributed shared memory services and provided examples and experimental results using the Kerrighed distributed OS. The work was performed as part of Djawida Dib's thesis [22].

6.3. A Chemical Approach for Autonomous Service Computing

Participants: Héctor Fernández, Marko Obrovac, Cédric Tedeschi.

6.3.1. Chemical Computing for the Simulation of Agile-Based Software Engineering

Participants: Héctor Fernández, Cédric Tedeschi.

In the framework of Héctor Fernández' internship at Vrije University, we applied the chemical programming model to simulate the behavior of a team developing software with Agile methods. Although an unexpected application, it has been the occasion to widen the range of applications and users of the software prototype developed during Héctor's thesis.

6.3.2. 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. While the problem resembles the classic resource allocation problem, it differs from it by different aspects. One of the main difference stands in the fact that the probability of a conflict varies during the course of execution. When the number of possible reactions is high, then there is no need for a complex conflict resolution scheme, as it would lead to a useless additional cost. In contrary, when this number drops, the probability of a conflict increases, and a systematic conflict resolution is mandatory to ensure at least one reaction will take place.

An adaptive protocol has been proposed, based on the dynamic combination of several strategies. Based on simulations, we have shown that, by dynamically switching from one strategy to another one, even by locally deciding which protocol to use, it is possible to combine the good properties of the strategies without suffering from their drawbacks [18].

The work was recently extended to take several rules into account. Rules have been defined to be able, not only to choose a strategy, but also to choose the rule to be executed, with the constant objective of maximizing the number of reactions executed in a given time.

6.3.3. DHT-based Runtime for the Chemical Programming Model

Participants: Marko Obrovac, Cédric Tedeschi.

The development of a distributed chemical machine entered its experimental phase with the development of a software prototype containing the following building blocks:

- A distributed hash table structures the network and allows any node to communicate with any other node in a logarithmic number of hops in this logical overlay.
- On top of the distributed hash table, a set of discovery mechanisms allows to find molecules needed in reactions, whatever their location is. These mechanisms are based on complex distribution and retrieval scheme borrowed from the P2P literature.
- The atomic capture protocol described before has been fully integrated in this framework.
- The discovery of molecules has been extended in order to detect the termination of the program and to be able to send the results of the computation back to the requester.

This software prototype has been deployed over the Grid'5000 platform [36].

OASIS Project-Team

6. New Results

6.1. Programming and Composition Models for Large-Scale Distributed Computing

6.1.1. Multi-active Objects

Participants: L. Henrio, F. Huet, A. Bourdin.

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

- improvement on the model and its formalisation
- use of the new model in our CAN P2P network (see below); this was also the opportunity to improve our implementation.

This year, we also spent considerable efforts to publish this work; a conference paper is currently under review.

6.1.2. Events for Algorithmic skeletons

Participant: L. Henrio.

In the context of the SCADA associated team, we worked on the algorithmic skeleton programming model. The structured parallelism approach (skeletons) takes advantage of common patterns used in parallel and distributed applications. The skeleton paradigm separates concerns: the distribution aspect can be considered separately from the functional aspect of an application.

- This year we focused on the handling of events in algorithmic skeletons: adding the possibility for a skeleton to output an event should increase the control and monitoring capabilities of algorithmic skeletons. The ultimate goal is to improve autonomicity for algorithmic skeletons.

6.1.3. Behavioural models for Distributed Components

Participants: E. Madelaine, N. Gaspar, A. Savu, L. Henrio.

In the past [3], we defined the behavioural semantics of active objects and components. This year we extended this work to address 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 the CADP model-checker, that also supports distributed generation of state-space. This year our main achievements are the following:

- We entirely formalised the specification of the behavioural model generation for component systems. This should provide us both a stronger formal background for our works in this area, and a specification for the automatic generation of behavioural models for our component systems.
- We additionally have put considerable efforts on the improvement of the Vercors platform and its integration with the Papyrus framework (see Section 5.2).

The formal work has been published as a research report [40]. A journal version is under submission. This work was done in collaboration with Rabéa Ameur-Boulifa from Télécom-Paristech.

In parallel with core developments of the behavioural specification environment, our collaborations led us to the study of the following application domain. In the context of the Spinnaker project, we are interested in developing a component-based distributed application to manage and monitor some pre-existing component-based distributed application - and hence, we called it The HyperManager. Our in-house component model (GCM) provides all the means to define, compose and dynamically reconfigure such applications. However, special care must be taken for this kind of undertaking. To this end, this year:

- We made the first steps towards a platform for the mechanized specification and verification, in the Coq Proof Assistant, of GCM applications. This work was published in [33], and is progressively being updated [1] to cope with behavioural specification, and to seamlessly combine deductive and model-checking techniques.
- We studied a real-life application scenario for our HyperManager prototype using distributed model-checking techniques in order to cope with the huge space state generated from reconfigurable applications.

6.1.4. Autonomic Monitoring and Management of Components

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

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. For this purpose, we finalized the implementation of a factory which, from any GCM ADL description can instantiate the requested non functional components of a GCM application.

The framework for autonomic computing allows the designer to describe in a separate way each phase of the MAPE autonomic control loop (Monitoring, Analysis, Planning, and Execution), and to plug them or unplug them dynamically. We have demonstrated how such a control loop can be relevant to drive the dynamic reconfiguration of services part of a SOA application, considering as in the SCA standard, that services are components [15].

Our objective now is to exemplify such autonomic and structured approach in the management of any distributed middleware or application, e.g. in the Spinnaker industrial context.

6.1.5. Optimization of data transfer in SOA and EDA models

Participants: I. Alshabani, F. Baude, L. Pellegrino, B. Sauvan, Q. Zagarese.

Traditional client-server interactions rely upon method invocations with copy of the parameters. This can be useless in particular if the receiver does not effectively uses them. On the contrary, copying and transferring parameters lazily so to allow the receiver to proceed even without all of them is a meaningful idea that we proved to be effective for active objects in the past [56]. This idea wasn't so far realized in the context of the web services technology, the most popular one used today for client-server SOAP-based interactions.

- To such an aim, we contributed to the offloading of objects representing parameters of the web service Java Apache CXF API [29]. It is innovative notably in the way the offloading of parameters for on-demand access can be delegated from services to services, which resembles the concept of first-class futures.
- Relying upon such an effective approach, we have applied a similar idea of "lazy copying and transfer" to the data parts of events in the context of event-driven architecture applications [28]. The middleware dynamically off-loads data (generally of huge size) attached to an event, according to some user-level policy expressed as annotation in the Java code at the subscriber side. The event itself, without its attachments, gets forwarded into the publish/subscribe brokering system (in our case, the event cloud middleware, that is the subject of section 6.2.1) and its attachements are transferred to the subscriber only on-demand. Compared to some existing propositions geared towards a data centric publish-subscribe pattern (e.g. the DDS OMG standard), ours is more user-friendly as it does not require the user code to explicitly program when to get the data attached to notified events.

[1] http://www-sop.inria.fr/members/Nuno.Gaspar/Mefresa.php

Overall, this work opens the way towards a strong convergence between service oriented and event-driven technologies.

6.1.6. Multi-layer component architectures

Participant: O. Dalle.

Since a few years, we have been investigating the decomposition of a simulation application into multiple layers corresponding to the various concerns commonly found in a simulation: in addition to the various modeling domains that may be found in a single simulation application (e.g. telecommunications networks, road-networks, power-grids, and so on), a typical simulation includes various orthogonal concerns such as system modelling, simulation scenario, instrumentation and observation, distribution, and so on. This large number of concerns has put in light some limits of the traditional hierarchical component-based architectures and their associated ADL, as found in the FCM and GCM. In order order to cope with these limitations, we started a new component architecture model called Binding Layers centered on the binding rather than the component, with no hierarchy but advanced layering capabilities, and offering advanced support for dynamic structures[32].

6.2. Middleware for Grid and Cloud computing

6.2.1. Publish-Subscribe in Distributed Environments

Participants: F. Baude, F. Huet, F. Bongiovanni, L. Pellegrino, B. Sauvan, I. Alshabani, A. Bourdin, M. Antoine, A. Alshabani.

In the context of the SOA4ALL FP7-IP project, we designed and implemented a hierarchical Semantic Space infrastructure based on Structured Overlay Networks (SONS) [62], [63]. It originally aimed at the storage and the retrieval of the semantic description of services at the Web scale [57]. This infrastructure combines the strengths of both the P2P paradigm at the architectural level and the Resource Description Framework (RDF) data model at the knowledge representation level. The achievements of this year are the following:

- In the context of the FP7 Strep PLAY and French ANR SocEDA research projects, we have been extending the aforementioned work with a content-based Publish/Subscribe abstraction in order to support asynchronous queries for RDF-based events in large scale settings, which raises some interesting challenges [26]. The goal is to build a platform for large scale distributed reasoning[25]. Such an integrated working platform [39], [38] has been presented in two tutorials [27], [54].
- We have also investigated the Publish/Subscribe paradigm in the MapReduce programming model. We have proposed the concept of continuous job which allows MapReduce jobs to be re-executed when new data are added to the system. To maintain the correctness of the execution, we have introduced the notion of carried data, i.e. data which are kept between subsequent executions. An implementation has been written on top of Hadoop and a paper submitted.

6.2.2. Distributed algorithms for CAN-like P2P networks

Participants: L. Henrio, F. Bongiovanni, F. Huet.

The nature of some large-scale applications, such as content delivery systems or publish/subscribe systems, built on top of SONs, demands application-level dissemination primitives which do not overwhelm the overlay, i.e. efficient, 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. Our objective is to design and prove an efficient and reliable broadcast algorithm for CAN-like P2P networks. To this aim, in 2012 we:

- Improved the formalisation in Isabelle/HOL of a CAN-like P2P system, devised formalised tools to reason on CAN topologies, and on communication protocols on top of CANs. We designed and proved the efficiency of a first naive algorithm.
- Sketched on paper the proof of completeness and efficiency for the algorithm we designed and implemented last year.

Part of this work was done in the PhD thesis of F. Bongiovanni [10]

We are also investigating the new algorithms to efficiently build a SONs in the presence of existing data. Most of the work on SONs assume that new peers joining the network will arrive without data or fail to take into account the cost of distributing these data. Indeed, depending on the key subspace given to the new peer, some or all its data will have to be distributed in the network. In 2012:

- We proposed a first version of new join algorithms which try to allocate key sub-spaces to peers so that the amount of data that needs to be moved is minimal. An expected benefit of this work is that it should allow for fast and efficient reconstruction of a SON in case of a crash, without having to use distributed snaphshots.

6.2.3. Network Aware Cloud Computing

Participants: S. Malik, F. Huet.

We have worked on the Resource Aware Cloud Computing project. Its primary purpose is to address different issues which can help the scheduler to make more efficient scheduling decisions. These issues are related to the resource characteristics.

- We introduce a framework, which increases the performance of the application and ensures high level of reliability during the scheduling of application onto the cloud. It is a cloud scheduler module named as Resource Aware Cloud Scheduling (RACS) module. It helps the scheduler in making the scheduling decisions on the basis of different characteristics of cloud resources. These characteristics are reliability, network latency, and monetary cost. RACS consists of multiple sub modules, which are responsible for their corresponding tasks. In RACS, we have done the implementation for the different issues.
- We worked on a model for the reliability assessment of the cloud's computing nodes. This reliability assessment mechanism helps to do the scheduling on cloud infrastructure and perform fault tolerance on the basis of the reliability values acquired during reliability assessment. The model has different algorithms for different types of applications. Thus it has multiple reliability values for each computing node. For real time applications, the model has time based reliability assessment algorithms.

This work is part of S. Malik's PhD thesis [12]

6.2.4. Testbed Designs from Experimenters Requirements

Participant: F. Hermenier.

The physical design of the Emulab facility, and many other testbeds like it, has been based on the facility operators' expectations regarding user needs and behavior. If operators' assumptions are incorrect, the resulting facility can exhibit inefficient use patterns and sub-optimal resource allocation.

- We have collaborated with Robert Ricci from the University of Utah on the study of the Utah' Emulab facility to provide better testbed designs. Our study gained insight into the needs and behaviors of networking researchers by analyzing more than 500,000 topologies from 13,000 experiments submitted to Emulab.
- Using this dataset, we re-visited the assumptions that went into the physical design of the Emulab facility and considered improvements to it. Through extensive simulations with real workloads, we evaluated alternative testbeds designs for their ability to improve testbed utilization and reduce hardware costs.

The results have been published to TridentCom [22], the reference conference related to testbeds and research infrastructures, and the article received the best paper award.

6.2.5. Energy Efficient Virtual Machines Placement in Data Centers

Participant: F. Hermenier.

Data centres are powerful ICT facilities which constantly evolve in size, complexity, and energy consumption. At the same time, tenants' and operators' requirements become more and more complex. The data centre operators may target different energy-related objectives while the workload volatility may alter the data centre capacity at supporting load spikes. Finally, clients of data centres are looking for dependable infrastructures that can comply with their SLA requirements.

To stay attractive, a data centre should then support these expectations. These constraints are however very specific to each of the tenants but also to the infrastructure. They also cover a large range of concerns (hardware requirements, performance, security ...) that are continuously evolving according to new trends and new technologies. Existing solutions are however ad-hoc and can not be updated easily to fit the data centres and the workload specificities.

We proposed a flexible energy-aware framework to address the multiple facets of an energy-aware consolidation of VMs in a cloud data centre.[21] This framework extended BtrPlace to make it able to address specific energy concerns. We integrated a fine grain energy model reducing either gas emissions or power consumption. We also proposed constraints to control the aggressiveness of these objectives to let the data centre reactive when a load spike occurs. We finally proposed various constraints to satisfy the hardware and the resource requirements of the tenants. The evaluation on a testbed running an industrial workload validated the practical benefits provided by the usage of our framework.

6.2.6. GPU-based High Performance Cloud Computing

Participants: M. Benguigui, F. Baude, F. Huet.

To address HPC, GPU devices are now considered as unavoidable cheap, energy efficient and very efficient alternative computing units. The barrier to handle such devices is the programming model: it is both very fine grained and synchronous.

Our long term goal is to devise some generic solutions in order to incorporate GPU-specific code whenever relevant into a parallel and distributed computation. The first step towards this objective is to gain some insight on how to efficiently program a non trivial but well known algorithm. We selected the American basked option pricing non embarrassingly parallel problem that was previously parallelized and distributed using ProActive master-slave approach [60], achieving an almost linear speedup and good performances (64 CPUs based computation allowed us to solve the problem in about 8 hours). The same algorithm has been reorganized for running on **a single GPU** [17] and achieved the same option pricing computation in about 9 hours. The current work is to succeed to take advantage of GPUs, even if non homogeneous, hired from a Cloud or a federation of clouds at once, orchestrated by an active object acting as a GPU task delegator. The goal is to drastically lower the overall computation time for such highly time consuming stochastic simulation problems.

6.3. Large-scale Simulation Platform: Techniques and methodologies

Participants: O. Dalle, E. Mancini.

In the domain of simulation techniques and methodologies, this year, we conducted research in the three following areas:

Simulation in the Cloud In recent years, numerous applications have been deployed into mobile devices. However, until now, there have been no attempts to run simulations on handheld devices. In the context of the DISSIMINET Associate Team, we work in collaboration with our partners at the Carleton University to investigate different architectures for running and managing simulations on handheld devices, and putting the simulation services in the Cloud[24]. We propose a hybrid simulation and visualization approach, where a dedicated mobile application is running on the client side and the RISE simulation server is hosted in the Cloud.

Simulation Methodology In the context of the ANR INFRA SONGS project, we are involved (as coordinators) in a Work-Package called "Open Science" whose aim is to investigate and contribute means to ensure the long term visibility and reproductibility of simulation results obtained using the SimGrid simulation platform. Our preliminary work in this direction consisted in identifying the issues, trends and potential solutions to ensure the long-term reproducibility of simulations[16].

Peer-to-peer Simulation In order to evaluate the performance and estimate the resource usage of peer-to-peer backup systems, it is important to analyze the time they spend in storing, retrieving and keeping the redundancy of the stored files. The analysis of such systems is difficult due to the random behavior of the peers and the variations of network conditions. In the context of the ANR USS-SIMGRID and INFRA-SONGS projects, we investigated means for reproducing such varying conditions in a controlled way. We worked on the design of a general simulation meta-model for peer-to-peer backup systems and a tool-chain, based on SimGrid, to help in their analysis[20]. We validated the meta-model and tool-chain through the analysis of a common scenario, and verified that they can be used, for example, for retrieving the relations between the storage size, the saved data fragment sizes and the induced network workload. We also started to investigate a new simulation technique for very-large scale distributed simulation of peer-to-peer systems based on the decomposition of a simulation into many micro-simulation steps[31] in order to optimize the overlap between communications and computations.

PHOENIX Project-Team

6. New Results

6.1. Design-driven Testing by simulation

Previously, we have introduced a paradigm-oriented development approach that revolves around a conceptual framework concretized by a design language [26]. A design description is used to generate high-level programming support, to perform a range of verifications, and to abstract over underlying technologies.

This approach is illustrated with the Sense-Compute-Control (SCC) paradigm [48], where an SCC software system gathers information about an environment via sensors (whether hardware or software) and issues orders to impact the environment via actuators. The SCC paradigm has a wide spectrum of applicability; we have used it successfully in the domains of home/building automation, multimedia, avionics and networking.

SCC systems involve both software concerns, like any software system, and integration concerns, for the constituent networked entities forming the environment of the SCC-loop. This situation is problematic for testing because it requires acquiring, testing and interfacing a variety of software and hardware entities. This process can rapidly become costly and time-consuming when the target environment involves many entities.

We have developed a simulation approach and a tool named *DiaSim* that leverage the DiaSpec description of an environment [15]. This description is used to generate both a programming framework to develop the simulation logic and an emulation layer to execute applications. The generic nature of our approach has been illustrated by leveraging two different simulation tools, namely, Siafu for 2-D rendering of home/building spaces, and FlightGear for avionics.

To fuel the simulation of an environment with accurate stimuli, we need to model real systems, including natural phenomena (*e.g.,* heat transfer in a building) or mechanical systems (*e.g.,* aircraft models). These physical models are typically defined as continuous systems using differential equations. To facilitate the reuse of off-the-shelf physical models, we have used a DSL named Acumen for describing differential equations. Acumen continuous models are coupled with the DiaSim discrete simulator, forming a hybrid system fueled by accurate stimulus producers [18].

These major accomplishments were conducted by Julien Bruneau, in the context of his PhD studies [11].

6.2. Design-driven Development of Dependable Software Systems

Dependability of a system is the ability to avoid service failures that are more frequent and more severe than is acceptable [22]. This generic concept includes attributes such as availability, integrity and reliability. Dependable systems are now pervasive in a range of domains (*e.g.,* railway, avionics, automotive) and require a certification process. The main goal of certification is to demonstrate that a system is conform to its *high-level requirements*, resulting from functional and safety analyses.

Software plays an increasingly important role in dependable systems; software development is thus required to be certified. In particular, the stakeholders have to pay attention to the coherence of the functional and non-functional aspects of an application to demonstrate the conformance of the software with the high-level requirements. Non-functional aspects of a system refer to constraints on the manner in which this system implements and delivers its functionality (*e.g.,* performance, reliability, security) [48].
Coherence. Because functional and non-functional aspects are inherently coupled, ensuring their coherence is critical to avoid unpredicted failures [39]. For example, fault-tolerance mechanisms may significantly deteriorate the application performance. Generally, this kind of issues are detected at the late stages of the development process, increasing the development cost of applications [21].
Conformance. Ensuring that an application is in conformance with its high-level requirements is typically done by tracing their propagation across the development stages. In practice, this process is human-intensive and error prone because it is performed manually [37].

Certifying a development process requires a variety of activities. In industry, the usual procedures involve holding peer review sessions for coherence verification, and writing traceability documents for conformance certification. In this context, *design-driven development* approaches are of paramount importance because the design drives the development of the application and provides a basis for tracing requirements [53]. However, because most existing approaches are general purpose, their guidance is limited, causing inconsistencies to be introduced in the design and along the development process. This situation calls for an integrated development process centered around a conceptual framework that allows to guide the certification process in a systematic manner. In response to this situation, we proposed a design-driven development methodology, named DIASUITE [2], which is dedicated to the *Sense/Compute/Control (SCC) paradigm* [48]. As demonstrated by Shaw, the use of a specific paradigm provides a conceptual framework, leading to a more disciplined engineering process and guiding the verification process [47]. An SCC application is one that interacts with a physical environment. Such applications are typical of domains such as home/building automation, robotics and avionics.

In this work, we have shown the benefits of DIASUITE for the development of dependable SCC applications. This approach is applied to a realistic case study in the avionics domain, in the context of two non-functional aspects, namely time-related performance and reliability. The DIASUITE design language, named DIASPEC , offers declarations covering both functional and non-functional dimensions of an SCC application [2], [9] [32]. However, so far, the DIASUITE methodology has only been used to study each dimension in isolation, leaving open the problems of coherence and conformance when considering multiple dimensions. This work integrates all these dimensions, enabling the generation of validation support. More precisely, this work makes the following contributions:
Design coherence over functional and non-functional dimensions. We use the DIASPEC language to describe both functional and non-functional aspects of an application and apply this approach to a realistic case study. A DIASPEC description is verified at design time for coherence of its declarations. This verification is performed with respect to a formal model generated from a DIASPEC description.
Design conformance through the development process. At design time, we provide verification support to check the conformance between the specification and the formalized form of the high-level requirements. At implementation time, we guarantee the conformance between the application code and the previously verified requirements. This process is automatically done by leveraging the generative approach of DIASUITE . As some of the high-level requirements cannot be ensured at design time (*e.g.,* time-related performance), we provide further testing support to validate the implementation with respect to these remaining requirements. This support leverages a realistic flight simulator, namely FlightGear [44].
Validation in avionics. We validate our approach by developing a realistic case study in avionics. Following the DIASUITE methodology, we have developed an aircraft flight guidance system and tested it on FlightGear. Additionally, we have duplicated this case study in the context of a commercial drone system, namely Parrot AR.Drone. [2]

These accomplishments were conducted by Julien Bruneau, Quentin Enard and Stéphanie Gatti, in the context of their PhD studies. This work will be published at the International Conference on Pervasive and Embedded Computing and Computation Systems (PECCS'13).

6.3. Putting DiaSuite to Work

A continuing concern of the Phoenix research group is to put our work into practice by tackling realistic applications. We have validated DiaSuite on a variety of applications in areas including telecommunications, pervasive computing, and avionics.

Our expertise in smart home and building, combined with the maturity of DiaSuite, have given rise to the development of a dedicated instance of our technology called DiaSuiteBox. This instance is destined for technology transfer.

[2] http://ardrone.parrot.com

6.3.1. Applying DiaSuite to a Variety of Areas.

Let us examine the application of DiaSuite to two key areas: pervasive computing and avionics. In each case, demonstrations and posters have been presented to researchers and industrial partners [24], [25], [23], [35]

Smart Homes. Despite much progress, developing a pervasive computing application remains a challenge because of a lack of conceptual frameworks and supporting tools. This challenge involves coping with heterogeneous entities, overcoming the intricacies of distributed systems technologies, working out an architecture for the application, encoding it in a program, writing specific code to test the application, and finally deploying it.

At the beginning of this evaluation period, our research group was mainly interested in orchestrating applications in the telecommunications domain, leveraging new opportunities created by the emergence of Voice over IP (mainly based on SIP). Concurrently, a myriad of objects became networked, prompting a need to expand the scope of telecommunications beyond human-human interaction.

Two main industrial collaborations were instrumental to explore the scope of this evolution and to validate the Diasuite approach with realistic case studies. First, we collaborated with a French telecommunications company, in a two-year project named HomeSIP, to study the convergence between VoIP and networked objects in the context of home automation. During this project, we developed a range of applications, including remote appliance control through phone keypad, TV recording via SMS, and dynamic entry phone systems. Second, we contributed to a two-year project named SmartImmo, which gathered major French companies in the area of building construction, installation, and management. The goal of this project was to create a service infrastructure for building automation. SmartImmo gave us the opportunity to elaborate realistic building automation scenarios (*e.g.,* parking lot management, meeting room reservation, energy monitoring).

Our work on applying DiaSuite to the pervasive computing domain has leveraged key contributions by two PhD students of Phoenix, namely, Wilfried Jouve [36] and Nicolas Palix [42]. They both defended at the beginning of this evaluation period.

Avionics. Safety-critical applications have to fulfill stringent requirements, both functional and non-functional. These requirements have to be coherent with each other and must be preserved throughout the software development process. In this context, a design-driven development approach can play a critical role. However existing design-driven development approaches are often general purpose, providing little, if any, conceptual framework to guide the development. Previously, we explained how the DiaSuite approach was enriched with non-functional declarations such as QoS and error handing.

To validate the interest of DiaSuite for safety-critical applications, several avionics case studies have been realized in the context of a collaboration with Thales, a French airborne systems company. One case study was a flight guidance application; it is in charge of the plane navigation and is under the supervision of the pilot. For example, if the pilot specifies a heading to follow, the application compares it to the current heading, sensed by devices such as the Inertial Reference Unit, and maneuvers ailerons accordingly. To test this application, we have used the DiaSim tool coupled with the FlightGear simulator. A flight guidance application has also been developed for a commercial drone platform. The goal of this application was to make the drone autonomous by following a flight plan similar to the one in avionics.

This simulation work has been presented in the thesis of Julien Bruneau [11]. Non-functional concerns addressing error handling and QoS will be presented in two forthcoming dissertations.

6.3.2. DiaSuiteBox: an Ongoing Technology-Transfer Project.

The DiaSuiteBox platform runs an open-ended set of applications, leveraging a range of appliances and web services. Our solution consists of a dedicated development environment, an application store, and a lightweight runtime platform. This solution is based on DiaSuite.

DiaSuiteBox consists of three main components:

- A tool-based environment is dedicated to the development of applications, orchestrating networked entities. This environment leverages DiaSpec, its compiler and an Eclipse plugin.
- An application store is composed of two servers: (1) a server verifies and packages submitted applications of developers prior to making them available to users and (2) another server enables users to browse, select and install applications.
- An execution environment runs end-user applications and allows to manage and configure all aspects of a smart space. This environment can either be deployed on low-resource computing platform (*e.g.,* Plug-PC, set-top-box) at the end-user's home or in the Cloud, coupled with a gateway for controlling equipments on the end-user's side.

Thanks to the application store and a developer community, the platform should provide users with a stream of innovative applications. During the submission process, an application is automatically analyzed and checked in order to be certified. [3] The user is ensured that the behavior of its applications is innocuous and conform to their description. DiaSuiteBox supports several technology standards like UPnP, Bluetooth, USB...This platform can be easily extended by plugging appliances directly on the hardware platform or by connecting devices on the local network.

[3] This certification process is preliminary in the current version of DiaSuiteBox.

REGAL Project-Team

6. New Results

6.1. Introduction

In 2012, we focused our research on the following areas:

- *Management of distributed data.*
- *Performance and robustness of Systems Software in multicore architectures.*

6.2. Distributed algorithms for dynamic networks

Participants: Luciana Arantes [correspondent], Olivier Marin, Sébastien Monnet, Franck Petit [correspondent], Maria Potop-Butucaru, Pierre Sens, Julien Sopena, Raluca Diaconu, Ruijing Hu, Anissa Lamani, Sergey Legtchenko, Jonathan Lejeune, Karine Pires, Guthemberg Silvestre, Véronique Simon.

This objective aims to design distributed algorithms adapted to new large scale or dynamic distributed systems, such as mobile networks, sensor networks, P2P systems, Grids, Cloud environments, and robot networks. Efficiency in such demanding environments requires specialised protocols, providing features such as fault or heterogeneity tolerance, scalability, quality of service, and self-stabilization. Our approach covers the whole spectrum from theory to experimentation. We design algorithms, prove them correct, implement them, and evaluate them in simulation, using OMNeT++ or PeerSim, and on large-scale real platforms such as Grid'5000. The theory ensures that our solutions are correct and whenever possible optimal; experimental evidence is necessary to show that they are relevant and practical.

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

Since 2008, we have obtained results both on fundamental aspects of distributed algorithms and on specific emerging large-scale applications.

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

6.2.1. Mutual Exclusion and Failure Detection.

Mutual Exclusion and Fault Tolerance are two major basic building blocks in the design of distributed systems. Most of the current mutual exclusion algorithms are not suitable for modern distributed architectures because they are not scalable, they ignore the network topology, and they do not consider application quality of service constraints. Under the ANR Project *MyCloud* and the FSE *Nu@age*, we study locking algorithms fulfilling some QoS constraints often found in Cloud Computing [38].

A classical way for a distributed system to tolerate failures is to detect them and then recover. It is now well recognized that the dominant factor in system unavailability lies in the failure detection phase. Regal has worked for many years on practical and theoretical aspects of failure detections and pioneered hierarchical scalable failure detectors. [2] Since 2008, we have studied the adaptation of failure detectors to dynamic networks. Following the model introduced in [18], we have proposed new algorithms to detect crashes and Byzantine behaviors [32].

These algorithms were designed as part of the ANR Project SHAMAN.

[2] Recent work by Leners et al published in SOSP 2011 uses our DSN 2003 paper as basis for performance comparison

6.2.2. Self-Stabilization and Self-* Services.

We have also approached fault tolerance through self-stabilization. Self-stabilization is a versatile technique to design distributed algorithms that withstand transient faults. In particular, we have worked on the unison problem, [3] i.e., the design of self-stabilizing algorithms to synchronize a distributed clock. As part of the ANR project *SPADES*, we have proposed several snap-stabilizing algorithms for the message forwarding problem that are optimal in terms of number of required buffers [36]. A snap-stabilizing algorithm is a self-stabilizing algorithm that stabilizes in 0 steps; in other words, such an algorithm always behaves according to its specification.

Finally, we have applied our expertise in distributed algorithms for dynamic and self-* systems in domains that at first glance seem quite far from the core expertise of the team, namely ad-hoc systems and swarms of mobile robots. In the latter, as part of ANR project *R-Discover*, we have studied various problems such as exploration [29], and gathering [15].

6.2.3. Dissemination and Data Finding in Large Scale Systems.

In the area of large-scale P2P networks, we have studied the problems of data dissemination and overlay maintenance, i.e., maintenance of a logical network built over the a P2P network. First, we have proposed efficient distributed algorithms to ensure data dissemination to a large set of nodes. Also, we have introduced a new method to compare dissemination algorithms over various topologies [35].

6.2.4. MMOGs.

Peer-to-peer overlay networks can be used to build scalable infrastructures for MMOGs. Our work on MMOGs has primarily focused on the impact of latency constraints in dynamic distributed systems. In online P2P games, players are connected by a logical graph, implemented as an overlay network. Latency constraints imply that players that interact must remain close in the overlay, even when the mobility of players induces rapid changes in the graph.

We have also addressed problems related to cheating and arbitration. In a distributed system, certification of entities makes it possible to circumscribe malicious behavior, such as cheating in games. Certification requires the use of a trusted third party and is traditionally done centrally. At a large scale, however, centralized certification represents a bottleneck and a single point of attack or failure. We have proposed solutions based on distributed reputations to identify trusted nodes and use them as game referees to detect and prevent cheating [46]. Our method relies on previous work on the subject of trusted node collaboration to ensure reliable distributed certification [4].

6.3. Management of distributed data

Participants: Mesaac Makpangou, Olivier Marin, Sébastien Monnet, Pierre Sens, Marc Shapiro, Julien Sopena, Gaël Thomas, Pierpaolo Cincilla, Raluca Diaconu, Sergey Legtchenko, Jonathan Lejeune, Karine Pires, Thomas Preud homme, Masoud Saeida Ardekani, Guthemberg Silvestre, Pierre Sutra, Marek Zawirski, Annette Bieniusa, Pierpaolo Cincilla, Véronique Simon, Mathieu Valero.

Sharing information is one of the major reasons for the use of large-scale distributed computer systems. Replicating data at multiple locations ensures that the information persists despite the occurrence of faults, and improves application performance by bringing data close to its point of use, enabling parallel reads, and balancing load. This raises numerous issues: where to store or replicate the data, in order to ensure that it is available quickly and remains persistent despite failures and disconnections; how to ensure consistency between replicas; when and how to move data to computation, or computation to data, in order to improve response time while minimizing storage or energy usage; etc. The Regal group works on several key issues related to replication:

[3] C. Boulinier, F. Petit, and V. Villain. Synchronous vs. asynchronous unison. Algorithmica, 51(1):61-80, 2008

[4] Erika Rosas, Olivier Marin and Xavier Bonnaire. CORPS: Building a Community Of Reputable PeerS in Distributed Hash Tables. The Computer Journal, 54(10):1721-1735(2011)

- Replica placement for fault tolerance and latency in the presence of churn,
- scalable strong consistency for replicated databases, and
- theory and practice of eventual consistency.

6.3.1. Distributed hash tables

A DHTs replicates data and spreads the replicas uniformly across a large number of nodes. Being very scalable and fault-tolerant, DHTs are a key component for dependable and secure applications, such as backup systems, distributed file systems, multi-range query systems, and content distribution systems.

Despite the advantages of DHTs, several studies show that they become inefficient in environments subject to churn, i.e., with many node arrivals and departures. We therefore propose a new replication mechanism for DHTs that is churn resilient [20]. RelaxDHT relaxes placement constraints, in order to avoid redundant data transfers and to increase parallelism. RelaxDHT loses up to 50% fewer data blocks that the well-known PAST DHT.

6.3.2. Strong consistency

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

To reduce the latency of consensus, we study *Generalised Consensus* algorithms, i.e., ones that leverage the commutativity of operations or the spontaneous ordering of messages by the network. We propose a novel protocol for generalised consensus that is optimal, both in message complexity and in faults tolerated, and that switches optimally between its fast path (which avoids ordering commuting requests) and its classical path (which generates a total order). Experimental evaluation shows that our algorithm is much more efficient and scales better than competing protocols.

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

Another research direction leverages isolation levels, particularly Snapshot Isolation (SI), in order to parallelize non-conflicting transactions on databases. We prove a novel impossibility result, namely that a system cannot have both genuine partial replication and SI. We designed an efficient protocol that maintains the most important features of SI, but side-steps this impossibility. Finally, we study the trade-offs between freshness (and hence low abort rates) and space complexity in computing snapshots, as required by SI and its variants.

Parallel transactions in distributed DBs incur high overhead for concurrency control and aborts. Our Gargamel system proposes an alternative approach by pre-serializing possibly conflicting transactions, and parallelizing non-conflicting update transactions to different replicas. It system provides strong transactional guarantees. In effect, Gargamel partitions the database dynamically according to the update workload. Each database replica runs sequentially, at full bandwidth; mutual synchronisation between replicas remains minimal. Our simulations show that Gargamel improves both response time and load by an order of magnitude when contention is high (highly loaded system with bounded resources), and that otherwise slow-down is negligible. This is published at ICPADS 2012 [27].

Our current experiments aim to compare the practical pros and cons of different approaches to designing large-scale replicated databases, by implementing and benchmarking a number of different protocols.

Our study the trade-offs between freshness and meta-date overhead, is published in HotCDP 2012 [43].

6.3.3. Eventual consistency

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

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

This new model is suited to large-scale systems, such as P2P or cloud computing. For instance, we propose a "sequence" CRDT type called Treedoc that supports concurrent text editing at a large scale, e.g., for a wikipedia-style concurrent editing application. We designed a number of CRDTs such as counters (supporting concurrent increments and decrements), sets (adding and removing elements), graphs (adding and removing vertices and edges), and maps (adding, removing, and setting key-value pairs). In particular, we publish a study of the concurrency semantics of sets in DISC 2012 [48], [22].

On the theoretical side, we identified sufficient correctness conditions for CRDTs, viz., that concurrent updates commute, or that the state is a monotonic semi-lattice. CRDTs raise challenging research issues: What is the power of CRDTs? Are the sufficient conditions necessary? How to engineer interesting data types to be CRDTs? How to garbage collect obsolete state without synchronisation, and without violating the monotonic semi-lattice requirement?

We are currently developing a very large-scale CRDT platform called SwiftCloud, which aims to scale to millions of clients, deployed inside and outside the cloud.

6.4. Improving the Performance and Robustness of Systems Software in Multicore Architectures

6.4.1. Managed Runtime Environments

Participants: Bertil Folliot, Julia Lawall, Gilles Muller [correspondent], Marc Shapiro, Julien Sopena, Gaël Thomas, Florian David, Lokesh Gidra, Jean-Pierre Lozi, Thomas Preud homme, Suman Saha, Harris Bakiras, Arie Middelkoop, Koutheir Attouchi.

Today, multicore architectures are becoming ubiquitous, found even in embedded systems, and thus it is essential that managed languages can scale on multicore processors. We have found that a major scalability bottleneck is the implementation of high contention locks, which can overload the bus, eliminating all performance benefits from adding more cores. To address this issue, as part of the PhD of Jean-Pierre Lozi, we have developed remote core locking (RCL), in which highly contended locks are implemented on a dedicated server, minimizing bus traffic and improving application scalability (USENIX ATC 2012 [24]). This work initially targeted C code but is now being adapted to the needs of Java applications in the PhD of Florian David. Another bottleneck in the support for managed languages is the garbage collector. As part of the PhD of Lokesh Gidra, we have identified the main sources of overhead.

6.4.2. Systems software robustness

A new area of research for Regal, with the arrival of Gilles Muller in 2009 as Inria Senior Research Scientist and Julia Lawall in 2011 as Inria Senior Research Scientist, is on improving the reliability of operating systems code. Muller and Lawall previously developed Coccinelle, a scriptable program matching and transformation tool for C code that is now commonly used in the open-source development community, including by the developers of Linux, Wine and Dragonfly BSD. Based on Coccinelle, we have developed a new approach to inferring API function usage protocols from software, relying on knowledge of common code structures (Software – Practice and Experience [19]).

We have also proposed a method for automatically identifying bug-fixing patches, with the goal of helping developers maintain stable versions of the software (ICSE 2012 [45]) and have designed an approach to automatically generating a robust interface to the Linux kernel, to provide developers of new kernel-level code more feedback in the case of a misunderstanding of kernel API usage conventions (ASE 2012 [24]).

RMOD Project-Team

5. New Results

5.1. Object serializer

Participants: Martin Dias [Correspondant], Mariano Martinez-Peck, Stéphane Ducasse.

Fuel: A Fast General Purpose Object Graph Serializer Since objects need to be stored and reloaded on different environments, serializing object graphs is a very important activity. There is a plethora of serialization frameworks with different requirements and design trade-offs. Most of them are based on recursive parsing of the object graphs, an approach which often is too slow. In addition, most of them prioritize a language-agnostic format instead of speed and language-specific object serialization. For the same reason, such serializers usually do not support features like class-shape changes, global references or executing pre and post load actions. Looking for speed, some frameworks are partially implemented at Virtual Machine (VM) level, hampering code portability and making them difficult to understand, maintain and extend. That is why we work on Fuel, a general-purpose object serializer based on these principles: (1) speed, through a compact binary format and a pickling algorithm which invests time in serialization for obtaining the best performance on materialization; (2) good object-oriented design, without special help at VM; (3) serialize any object, thus have a full-featured language-specific format. We implement and validate this approach in Pharo, where we demonstrate that Fuel is faster than other serializers, even those with special VM support. The extensibility of Fuel made possible to successfully serialize various objects: classes in Newspeak, debugger stacks, and full CMS object graphs [11].

5.2. Cycles and dependencies

Participants: Stéphane Ducasse [Correspondant], Nicolas Anquetil, Muhammad Bhatti.

OZONE: Layer Identification in the presence of Cyclic Dependencies A layered software architecture helps understanding the role of software entities (e.g., packages or classes) in a system and hence, the impact of changes on these entities. However, the computation of an optimal layered organization in the presence of cyclic dependencies is difficult. We develop an approach that (i) provides a strategy supporting the automated detection of cyclic dependencies, (ii) proposes heuristics to break cyclic dependencies, and (iii) computes an organization of software entities in multiple layers even in presence of cyclic dependencies. Our approach performs better than the other existing approaches in terms of accuracy and interactivity, it supports human inputs and constraints. We compare this approach to existing solutions and apply it on two large software systems to identify package layers. The results are manually validated by software engineers of the two systems [12].

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 developers to remove cycles from their programs. We design an efficient algorithm that performs a fine-grained analysis of cycles among application packages. In addition, we define multiple metrics to rank cycles by their level of undesirability, prioritizing cycles that are the more undesired by developers. We compare these multiple ranking metrics on four large and mature software systems in Java and Smalltalk [14].

Resolving cyclic dependencies between packages with Enriched Dependency Structural Matrix Dependency Structural Matrix (DSM) is an approach originally developed for process optimization. It has been successfully applied to identify software dependencies among packages and subsystems. A number of algorithms have been proposed to compute the matrix so that it highlights patterns and problematic dependencies between subsystems. However, existing DSM implementations often miss important information to fully support reengineering effort. For example, they do not clearly qualify and quantify problematic relationships, information that is crucial to support remediation tasks. We propose Enriched Dependency Structural Matrix (eDSM), which provides small multiple views and micro-macro readings by adding fine-grained information in each cell of the matrix. Each cell is enriched with contextual information about (i) the type of dependencies (inheritance, class reference, etc.), (ii) the proportion of referencing entities, (iii) the proportion of referenced entities. We distinguish independent cycles and stress potentially simple fixes for cycles using coloring information. This work is language independent and has been implemented on top of the Moose software analysis platform.We improved the cell content information view based on user feedback and performed multiple validations: two different case studies on Moose and Seaside software; one user study for validating eDSM as a usable approach for developers. Solutions to problems identified with eDSM have been performed and retrofitted in analyzed software [13].

5.3. Warnings and bugs

Participants: Simon Allier [Correspondant], Andre Hora, Nicolas Anquetil, Muhammad Bhatti, Stéphane Ducasse.

A Framework to Compare Alert Ranking Algorithms To improve software quality, rule checkers statically check if a software contains violations of good programming practices. On a real sized system, the alerts (rule violations detected by the tool) may be numbered by the thousands. Unfortunately, these tools generate a high proportion of "false alerts", which in the context of a specific software, should not be fixed. Huge numbers of false alerts may render impossible the finding and correction of "true alerts" and dissuade developers from using these tools. In order to overcome this problem, the literature provides different ranking methods that aim at computing the probability of an alert being a "true one". We propose a framework for comparing these ranking algorithms and identify the best approach to rank alerts. We have selected six algorithms described in literature. For comparison, we use a benchmark covering two programming languages (Java and Smalltalk) and three rule checkers (FindBug, PMD, SmallLint). Results show that the best ranking methods are based on the history of past alerts and their location. We could not identify any significant advantage in using statistical tools such as linear regression or Bayesian networks or ad-hoc methods [15].

Uncovering Causal Relationships between Software Metrics and Bugs Bug prediction is an important challenge for software engineering research that consists in looking for possible early indicators of the presence of bugs in a software. However, despite the relevance of the issue, most experiments designed to evaluate bug prediction only investigate whether there is a linear relation between the predictor and the presence of bugs. However, it is well known that standard regression models can not filter out spurious relations. We describe an experiment to discover more robust evidences towards causality between software metrics (as predictors) and the occurrence of bugs. For this purpose, we have relied on Granger Causality Test to evaluate whether past changes in a given time series are useful to forecast changes in another series. As its name suggests, Granger Test is a better indication of causality between two variables. We present and discuss the results of experiments on four real world systems evaluated over a time frame of almost four years. Particularly, we have been able to discover in the history of metrics the causes - in the terms of the Granger Test - for 64% to 93% of the defects reported for the systems considered in our experiment [18].

BugMaps: A Tool for the Visual Exploration and Analysis of Bugs To harness the complexity of big legacy software, software engineering tools need more and more information on these systems. This information may come from analysis of the source code, study of execution traces, computing of metrics, etc. One source of information received less attention than source code: the bugs on the system. Little is known about the evolutionary behavior, lifetime, distribution, and stability of bugs. We propose to consider bugs as first class entities and a useful source of information that can answer such topics. Such analysis is inherently complex,

because bugs are intangible, invisible, and difficult to be traced. Therefore, our tool extracts information about bugs from bug tracking systems, link this information to other software artifacts, and explore interactive visualizations of bugs that we call bug maps [19].

A Catalog of Patterns for Concept Lattice Interpretation in Software Reengineering Formal Concept Analysis (FCA) provides an important approach in software reengineering for software understanding, design anomalies detection and correction. However, FCA-based approaches have two problems: (i) they produce lattices that must be interpreted by the user according to his/her understanding of the technique and different elements of the graph; and, (ii) the lattice can rapidly become so big that one is overwhelmed by the mass of information and possibilities. We make a catalog of important patterns in concept lattices, which can allow automating the task of lattice interpretation. The approach helps the reengineer to concentrate on the task of reengineering rather than understanding a complex lattice. We provide interpretation of these patterns in a generalized manner and illustrate them on various contexts constructed from program information of different open-source systems. We also present a tool that allows automated extraction of the patterns from concept lattices [16].

5.4. Reflective

Participants: Marcus Denker [Correspondant], Stéphane Ducasse.

DynamicSchema: a lightweight persistency framework for context-oriented data management While context-oriented programming technology so far has focused mostly on behavioral adaptation, context-oriented data management has received much less attention. We make a case for the problem of context-oriented data management, using a concrete example of a mobile application. We illustrate some of the issues involved and propose a lightweight persistency framework, called DynamicSchema, that resolves some of these issues. The solution consists in a flexible reification of the database schema, as a convenient dynamic data structure that can be adapted at execution time, according to sensed context changes. Implementing our mobile application using this framework enabled us to reduce the complexity of the domain modeling layer, to facilitate the production of code with low memory footprint, and to simplify the implementation of certain scenarios related to context-dependent security concerns [17].

SARDES Project-Team

5. New Results

5.1. Languages and Foundations: Process algebra

Participants: Damien Pous, Jean-Bernard Stefani, Barbara Petit.

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 and Rever projects.

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 primitives for controlling reversibility in a higher-order variant of the π-calculus. We have shown that the combination of a basic notion of message alternative coupled with a rollback primitive that respects causal consistency provides enough expresive power to encode various rollback policies. We have also started to study the expresive power of these primitives with respect to transactional constructs. In particular, we have shown that our primitives allow for a faithful encoding of a notion of communicating transaction proposed by Hennessy et al, while avoiding spurious rollbacks which mar Hennessy's approach. This work has been submitted for publication. A digest of our main ideas on controlling reversibility has appeared in [25].

We have also started a study on the cost of making a concurrent programming language reversible. More specifically, we have started from an abstract machine for a fragment of the Oz programming language and made it reversible. We have shown that the overhead of the reversible machine with respect to the original one in terms of space is at most linear in the number of execution steps, and that this bound is tight since some programs cannot be made reversible without storing a commensurate amount of information. This work has been published in [26].

5.2. Control for adaptive systems: Discrete control for adaptive and reconfigurable systems

Participants: Eric Rutten, Noël De Palma, Olivier Gruber, Fabienne Boyer, Xin An, Soguy Mak-Kare Gueye.

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. 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.

This year, we have started investigating the application of discrete controller synthesis to various problems in computer systems management and administration. The increasing complexity of computer systems has led to the automation of administration functions, in the form of autonomic managers. One important aspect requiring such management is the issue of energy consumption of computing systems, in the perspective of green computing. As these managers address each a specific aspect, there is a need for using several managers to cover all the domains of administration. However, coordinating them is necessary for proper and effective global administration. Such coordination is a problem of synchronization and logical control of administration operations that can be applied by autonomous managers on the managed system at a given time in response to events observed on the state of this system. We therefore propose to investigate the use of reactive models with events and states, and discrete control techniques to solve this problem. In [20], [21], [31], [30], we illustrate this approach by integrating a controller obtained by synchronous programming, based on Discrete Controller Synthesis, in an autonomic system administration infrastructure. The role of this controller is to orchestrate the execution of reconfiguration operations of all administration policies to satisfy properties of logical consistency. We have applied this approach to coordinate energy-aware managers for self-optimization, self-regulation of processor frequency and self-repair.

5.3. System support: System support for multicore machines

Participants: Vivien Quéma, Renaud Lachaize, Baptiste Lepers.

Multicore machines with Non-Uniform Memory Accesses (NUMA) are becoming commodity platforms. Efficiently exploiting their resources remains an open research problem. This line of work investigates system support to tackle various issues related to efficient resource management and programming support.

One of the key concerns in efficiently exploiting multicore NUMA architectures is to limit as much as possible the number of remote memory accesses (i.e., main memory accesses performed from a core to a memory bank that is not directly attached to it). However, in many cases, existing profilers do not provide enough information to help programmers achieve this goal. We have developed MemProf [24], the first profiler that allows programmers to choose and implement efficient application-level optimizations for NUMA systems. MemProf achieves this goal by allowing programmers to (i) precisely understand which memory objects are accessed remotely in memory, and (ii) building temporal flows of interactions between threads and objects. We evaluated MemProf using four applications (FaceRec, Streamcluster, Psearchy, and Apache) on three different machines. In each case, we showed how MemProf helped us choose and implement efficient optimizations, unlike existing profilers. These optimizations provide significant performance gains on the studied applications (up to 161%), while requiring very lightweight modifications (10 lines of code or less).

State-machine replication is a well-known fault-tolerance technique. Unfortunately existing state-machine replication schemes do not scale well on multicore machines. In collaboration with U. Texas at Austin (L. Alvisi), we have developed a new state-machine replication scheme [23], that departs from the standard agree-execute architecture of existing schemes, in favor of a more optimistic, and less deterministic, execute-verify replication scheme, which yields much better scalability. We have evaluated Eve's throughput gain compared with traditional sequential execution approaches, as well as Eve's overheads compared to unreplicated multithreaded execution and to alternative replication approaches.

5.4. System support: Performance and dependability benchmaking

Participants: Amit Sangroya, Damian Serrano-Garcia, Sara Bouchenak [correspondant].

MapReduce is a popular programming model for distributed data processing. Extensive research has been conducted on the reliability of MapReduce, ranging from adaptive and on-demand fault-tolerance to new fault-tolerance models. However, realistic benchmarks are still missing to analyze and compare the effectiveness of these proposals. To date, most MapReduce fault-tolerance solutions have been evaluated using micro benchmarks in an ad-hoc and overly simplified setting, which may not be representative of real-world applications. To remedy this situation, we have developed MRBS, a comprehensive benchmark suite for evaluating the dependability of MapReduce systems. MRBS includes five benchmarks covering several application domains and a wide range of execution scenarios such as data-intensive vs. compute-intensive applications, or batch applications vs. online interactive applications. MRBS allows to inject various types of faults at different rates. It also considers different application workloads and data loads, and produces extensive reliability, availability and performance statistics. We have shown the use of MRBS with Hadoop clusters running on Amazon EC2, and on a private cloud [29], [28].

SCORE Team

6. New Results

6.1. Collaborative Data Management

6.1.1. A Framework to Design Conflict-Free Replicated Data Types

Participants: Mehdi Ahmed-Nacer, Stéphane Martin, Pascal Urso.

Design new eventually consistent data types is difficult and error-prone as demonstrated by the numerous proposed approaches that fail to resolve conflicts for simple plain text document. Moreover, more the data type is complex, more conflicts types must be resolved. We have presented a layered approach to design new eventually consistent data types [21], [15]. This approach decouples eventual consistency management from data type constraints satisfaction. We compose one or several existing replicated data types which ensure eventual consistency, and adaptation layers to obtain a new eventually consistent data type. Each layer or replicated data type can be freely substituted by one providing the same interface. We have demonstrated that our approach is implementable and obtains acceptable performances. Our experiments and implementation are publicly available and re-playable (https://github.com/score-team/replication-benchmarker).

6.1.2. Enhancing Rich Content Wikis with Real-Time Collaboration

Participants: Luc André, Claudia-Lavinia Ignat, Gérald Oster.

Wikis are one of the most important tools of Web 2.0 allowing users to easily edit shared data. WYSIWYG editors for wiki pages avoid the impediments of learning wiki syntax. However, wikis offer poor support for merging concurrent contributions on the same pages. Users have to manually merge concurrent changes and there is no support for an automatic merging. As real-time collaborative editing reduces the number of conflicts as the time frame for concurrent work is very short, we proposed extending wiki systems with real-time collaboration [23]. We propose an automatic merging solution adapted for rich content wikis. Our solution is integrated as an extension of XWiki system (http://extensions.xwiki.org/xwiki/bin/view/Extension/RealTime+Wiki+Editor).

6.1.3. Rapid and Round-free Multi-pair Asynchronous Push-Pull Aggregation

Participants: Claudia-Lavinia Ignat, Hyun-Gul Roh.

In the context of STREAMS project we investigated gossip-based dissemination mechanisms in peer-to-peer real-time collaboration adapted for consistency maintenance algorithms based on CRDT (Commutative Replicated Data Types). These dissemination mechanisms need to compute the size of the network and therefore a suitable rapid protocol that aggregates data over network is essential. Iterative aggregation protocols, especially push-pull style aggregations, generally need prior configurations to synchronize rounds over all nodes, and messages should be exchanged in a synchronous/blocking way in order to ensure accurate estimates in push-pull or push-sum protocols. We proposed a multi-pair asynchronous push-pull aggregation (MAPPA) [22], which frees the push-pull aggregations from the synchronization constraints, and therefore accelerates the aggregation speed. MAPPA is resilient to network churns, and thus suitable for dynamic peer-to-peer networks.

6.1.4. Trustworthy contract based collaboration

Participants: Claudia-Lavinia Ignat, Hien Thi Thu Truong.

Availability of trustworthy environments is one of the main conditions that would lead to a greater acceptance and reliance on collaborative systems. In the context of large scale multi-synchronous collaboration where users work in parallel on different streams of activities a ”hard” security that would forbid many actions is unusable. We adopt instead a “soft” security where rather than adopting an a priori strict enforcement of security rules, access is given first to data without control but with restrictions that are verified a posteriori. We proposed a contract-based collaboration model [2], [4] where we establish and adjust trust in users based on detective enforcement of basic usage control requirements. Usage control requirements are specified as contracts. Contracts are specified by data owners when they share data in accordance with user trust levels. Observation of adherence to or violation of contracts is used to adjust trust levels. Our contract-based collaboration model allows the specification of contracts, merging of data and contracts and resolution of conflicting contracts. A trust metric for computing user trust levels was proposed based on auditing user compliance to the given contracts.

Multi-synchronous collaboration maintains multiple, simultaneous streams of activity which continually diverge and converge. These streams of activity are represented by means of logs of operations, i.e. user modifications. A malicious user might tamper his log of operations. At the moment of synchronization with other streams, the tampered log might generate wrong results. A trustworthy collaboration environment should detect if logs were tampered. We proposed a mechanism for establishment of trusted logs relying on hash-chain based authenticators [17], [18], [2]. Our solution ensures the authenticity, the integrity of logs, and the user accountability. We proposed algorithms to construct authenticators and verify logs. We proved their correctness and provided theoretical and practical evaluations.

6.1.5. Distributed activity management in crisis situation

Participants: François Charoy, Joern Franke.

Crisis management has been a very fruitful domain to investigate new approaches for 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 analyse requirements and confront our proposals. This paper present the final part of this work on the problem of replication of activities between several workspaces [3]. We are now looking for new vehicles to continue this research at an international level.

6.2. Data Centered Service Oriented Computing

6.2.1. Business process distribution on a SaaS architecture

Participants: Walid Fdhila, Claude Godart, Elio Goettelmann, Samir Youcef.

The objective of this work is to support the deployment of a business process as a set of distributed services provided partially or totally off-premises or even in the cloud. Direct applications in our target are:

- A methodological approach for choreographies elicitations and monitoring[12].
- An algorithm for optimized service providers selection (including cloud) [11], [9], [10].

In this objective, we have deployed two approaches. A first is based on heuristics (*greedy* algorithm to compute an initial solution, combined with a *tabu search*) for optimizing the selection of services assigned to activities in a decentralized composite service, both in terms of the overall QoS of the composite service and the communication overhead; in output, the initial business process model is translated in a set of interconnected business process fragments.

A second approach uses operational research techniques for optimizing a cloud selection taking into account two conflicting objectives, namely: the execution time (makespan) and the overall cost incurred using a set of resources. We proposed in [9] three complementary approaches to deal with the matching and scheduling scientific workflow tasks in Cloud computing environments. An extension of this first study was presented in [11], [10]. More precisely, we have extended the three proposed approaches to consider: (i) the business workflows and (ii) the concurrent access to resources by multiple instances of a given process. To achieve this goal, we proposed to use a predictive models in order to estimate the availability of the used resources. We

are currently working on the business processes execution in Cloud computing context taking into account workflow patterns such as *sequence, switch, multi-choice, etc.* patterns. Moreover, we plan to extend the proposed work to take into account others criteria like carbon emission and energy cost.

6.2.2. 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. Further than that we are trying to describe and validated an engineering method to help designer to maintain this alignment. Karim Dahman has defended is PhD on this matter in october 2012.

6.2.3. Monitoring and violations detections of choreographies or distributed compositions of services

Participants: Aymen Baouab, Ehtesham Zahoor, Olivier Perrin, Walid Fdhila, Claude Godart.

The dynamic nature of the cross-organizational business processes poses various challenges to their successful execution. Services choreographies or distributed compositions of services help to reduce such complexity by providing means for describing complex systems at a higher level. However, this does not necessarily guarantee that erroneous situations cannot occur due to inappropriately specified interactions. In [7], [6], we propose an approach for decentralized monitoring of cross-organizational choreographies, using a runtime event-based approach to deal with the problem of monitoring conformance of interaction sequences. Our approach allows for an automatic and optimized generation of rules. After parsing the choreography graph into a hierarchy of *canonical* blocks, tagging each event by its block ascendancy, an optimized set of monitoring queries is generated. We evaluate the concepts based on a scenario showing how much the number of queries can be significantly reduced. These results use our previous results about event-based framework DISC [33].

TRISKELL Project-Team

6. New Results

6.1. Distributed models at runtime

In the last two years we have developed a new models@runtime approach, named Kevoree. It supports extensive architecture evolution at runtime and enables the design of eternal systems with a continuous design process. The Kevoree type model supports dynamic types redefinition, allowing for complete redesign of specifications and implementations while the system is running. Communication channels between components are themselves first class dynamic entities. By combining our component metamodel and a *models@runtime* approach we have developed implementations of Kevoree for a wide range of computation nodes, ranging from inexpensive embedded microcontrollers to large commercial cloud implementations. We have shown that **applications based on the Kevoree component model are able to reconfigure their architecture completely on the fly** several times per second [40] on computation nodes with very limited resources.

Using the Kevoree platform, we demonstrated the use of *models@runtime* for large-scale distributed systems. We have shown that the *models@runtime* approach is applicable to pervasive distributed systems, even with volatile networks and continuously changing topologies [41]. Using *ad hoc* distributed algorithms, architectural models are propagated reliably in spite of frequent loss of connectivity, and **reconfigurations of a distributed application are managed in a continuous consistent manner**. Using colored Petri nets to describe quantititative properties we are building a toolchain to estimate the time related properties of assemblies at runtime [51].

6.2. Real scale platform for dynamic tactical decision system

Since mid 2011 the Triskell team is designing and implementing the DAUM platform that integrates a large range of technologies, ranging from wireless low cost sensors to clouds made of rugged field miniservers. Our application use case is a tactical decision system designed in cooperation with a large firefighter department of 3,500 firefighters. This platform is being used as a real life testbed for our results on dynamic, continuous design of distributed pervasive systems. It is also used as a concrete cooperation support within the Marie Curie Initial Training Network *Relate*.

By combining *models@runtime* techniques and component-based techniques, we have shown how we can apply model driven engineering to design large-scale, distributed, heterogeneous and adaptive systems [40].

6.3. Software Language Engineering

With the growing interest in MDE, more and more models are used during a software development to capture various aspects (both functional and extra-functional). Therefore, explicitly identifying and analyzing these relationships becomes a real challenge during a model-based software development. To address this challenge, we proposed a **formal language that captures relations between modeled things in order to reason and communicate about modeling activities** [19].

More recently, we started to explore the necessary breakthrough in software languages to support a global software engineering. Consequently, we investigate MDE-based tools and methods in software language engineering (SLE) for the design and implementation of collaborative, interoperable and composable modeling languages [32], [31], [30].

6.4. Model Typing

In recent years, the Triskell team established a formal theory of model typing, considering models as first class entities when modeling in the large [8]. Model typing was initially developed to support the reuse of both metamodels and model transformations [21]. It is now becoming the cornerstone of the various established metamodeling operators to ensure structural and behavioral properties [85][43].

[8] Model typing goes beyond the typing of individual model elements to actually deal with the type of graphs of model elements

The series of work on model typing was initially developed in the context of Jim Steel's PhD, defended in 2008. Then, it has continuously evolved in the scheme of the Naouel Moha's post doctoral position and the Clément Guy's PhD thesis [43]. Recently, work on model typing had a very strong application to the field of optimizing compilers [18]. This is the result of a close collaboration between Inria and Colorado State University (CSU), involving two teams in MDE (the Triskell team at Inria and the SE group at CSU), and two teams in optimizing compilers (the CAIRN team at Inria and the Mélange group at CSU). This collaboration was partially funded by the Inria associated teams MoCAA and LRS.

6.5. Model Footprint / Pruning / Slicing

During the previous evaluation period, we have established various facilities to ease the metamodeling activity.

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 [46]. Then, we propose a conjunct use of model typing and metamodel pruning to ease the reuse of model transformations on instances of different metamodels [21].

In general, many operators consist into extracting a subset of a model according to a language-based specification. 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 [70][66]. An extended version was recently published in SoSyM [14].

6.6. Model Composition

Triskell hence contributed to the software engineering community's effort to propose new ways of composing software from modeling elements, including for cross cutting concerns, that would unify the composition ideas behind Model Driven Engineering, Aspect Oriented Modeling, Software Product Lines etc [77]. Several research prototypes [9] have been built to provide new composition operators. In the Mickael Clavreul PhD [72], we define a framework to unify and classify existing model composition operator and ease the definition of new model composition operators. Theoretical basis to such a framework have been recently based on category theory in [48].

6.7. Model Variability

In the context of Aspects Oriented Modeling (AOM), one of the key challenge is the variability management leading to software product lines. Our work in this area has led to the involvement of the Triskell group in the ANR project MOVIDA, as well as in the OMG standardization process of the *Common Variability Language* where we developed a Kermeta-based implementation conforming to this future standard (called *kCVL*).

6.8. Testing software product lines

Nowadays, many applications are expected to run on a tremendous variety of execution environments. For example, network connection software must deliver the same functionalities on distinct physical platforms, which themselves run several distinct operating systems, with various applications and physical devices. Testing those applications is challenging as it is simply impossible to consider every possible environment configuration. We tackle this issue through the systematic selection of a subset of configurations for testing [45] and through model-based verification [37].

[9] http://www.kermeta.org/kompose/,http://www.kermeta.org/mdk/ModMap/

6.9. Testing service-oriented applications

The changes resulting from the evolution of Service Based Systems (SBSs) may degrade their design and quality of service (QoS) and may often cause the appearance of common poor solutions, called antipatterns. The automatic detection of antipatterns is thus important to assess the design and QoS of SBSs and ease their maintenance and evolution. Using our approach, we specify 10 well-known and common antipatterns, including Multi Service and Tiny Service, and we automatically generate their detection algorithms [50]. This work has received the best paper award at ICSOC 2012.

6.10. Testing aspect oriented programs

Aspect-oriented programming (AOP) promises better software quality through enhanced modularity.

Crosscutting concerns are encapsulated in separate units called aspects and are introduced at specific points in the base program at compile-time or runtime. However, aspect-oriented mechanisms also introduce new risks for reliability that must be tackled by specific testing techniques in order to fully benefit from the use of AOP. During the evaluation period, we proposed a series of work to analyze these new risks, let designers understand the interactions between the base and the aspects and test aspects. The major achievement is a **novel oracle to test the injection of aspects in a base program**. The oracle allows to capture new classes of errors that occur only in aspect-oriented programs. Its ability to capture these errors in a more efficient way than an object-oriented oracle (shorter test cases and written in less time), has been empirically demonstrated and was published in the Journal for Software Testing, Verification and Reliability [74].

6.11. Testing peer-to-peer systems

Peer-to-peer (P2P) is one of the major distributed platforms for many applications such as large data sharing and collaboration in social networks. However, building trustworthy P2P applications is difficult because they must be deployed on a large number of autonomous, volatile nodes, which may refuse to answer to some requests and even leave the system unexpectedly. This volatility of nodes is a common behavior in P2P systems and may be interpreted as a fault during tests (*i.e.*, failed node). In this context, we have developed a **novel framework and a methodology for testing P2P applications**. The framework is based on the individual control of nodes, allowing test cases to precisely control the volatility of nodes during their execution. We validated this framework through an experimentation on the FreePastry distributed hashtable. The experimentation tests the behavior of the system in different conditions of volatility and shows how the tests were able to detect complex implementation errors. This work, published in the Empirical Software Engineering journal [73], in collaboration with the ATLAS Inria team, is directly related to Triskell's goal to apply software engineering to distributed systems.

6.12. Testing the boundaries of a specific domain

The increasing use of domain-specific modeling to increase efficiency in modeling multiple concerns, increases the need to correctly formalize domain models. Domains are modeled as metamodels, which capture the domain's modeling spaces, *i.e.* the set of all models which structure conforms to the description specified in the metamodel. However, there is currently no systematic method to test that a metamodel captures all the correct models of the domain and no more. Our most recent contribution to testing focuses on the **automatic selection of models in the modeling space captured by a metamodel**. We adapt metaheuristic search to generate a set that covers as many representative situations as possible, while staying as small as possible. This work was published in the International Conference on Software Testing, verification and validation [27].

6.13. Testing interactive systems

While model-based design of interactive systems is moving from pure event-based models of WIMP interactions to stateful models of post-WIMP interactions, model-based test generation techniques for HCI currently consider only WIMP interaction testing. We proposed an original model-based test generation technique,

which aims at providing test cases to test post-WIMP behavior (*e.g.* multi-touch). We leverage the Malai architecture to model the system under test to establish two contributions: the definition of novel adequacy criteria to generate test cases that cover Malai models; an algorithm for the automatic generation of test suites that satisfy the adequacy criteria. We applied the novel approach to two open-source interactive systems to validate the ability of generated test cases to reveal bugs. This early work is part of the project Connexion (*cf.* Section 8.1.3) which notably focuses on testing interactive parts of critical systems.

ALGORILLE Project-Team

6. New Results

6.1. Structuring applications for scalability

In this domain we have been active on several research subjects: efficient locking interfaces, data management, asynchronism, algorithms for large scale discrete structures and the use of accelerators, namely GPU.

In addition to these direct contributions within our own domain, numerous collaborations have permitted us to test our algorithmic ideas in connection with academics of different application domains and through our association with SUPÉLEC with some industrial partners: physics and geology, biology and medicine, machine learning or finance.

6.1.1. Efficient linear algebra on accelerators.

Graphics Processing Units have evolved to fully programmable parallel vector-processor sub-systems. We have designed several parallel algorithms on GPUs, and integrated that level of parallelism into larger applications including several other levels of parallelism (multi-core, multi-node,...). In this context, we also have studied energy issues and designed some energy performance models for GPU clusters, in order to model and predict energy consumption of GPU clusters.

The PhD thesis of Wilfried Kirschenmann, has been a collaboration with EDF R&D and was co-supervised by S. Vialle and Laurent Plagne (EDF SINETICS). It has given rise to a DSEL based on C++ and to a unified generic library that adapts to multi-core CPUs, multi-core CPUs with vector units (SSE or AVX), and GPUs. This framework allows to implement linear algebra operations originating from neutronic computations, see [22].

The PhD thesis of Thomas Jost, co-supervised by S. Contassot-Vivier and Bruno Lévy (Alice INRIA team) deals with specific algorithms for GPUs, in particular linear solvers. He has also worked on the use of GPUs within clusters of workstations via the study of a solver of non-linear problems [17]. The defense of this thesis is planned in January 2013.

6.1.2. Combining locking and data management interfaces.

Handling data consistency in parallel and distributed settings is a challenging task, in particular if we want to allow for an easy to handle asynchronism between tasks. Our publication [4] shows how to produce deadlock-free iterative programs that implement strong overlapping between communication, IO and computation; [21] extends distributed lock mechanisms and combines them with implicit data management.

A new implementation (ORWL) of our ideas of combining control and data management in C has been undertaken, see 5.5 . A first work has demonstrated its efficiency for a benchmark application [18]. Our current efforts concentrate on the implementation of a complete application (an American Option Pricer) that was chosen because it presents a non-trivial data transfer and control between different compute nodes and their GPU. ORWL is now able to handle such an application seamlessly and efficiently, a real alternative to home made interactions between MPI and CUDA.

6.1.3. Discrete and continuous dynamical systems.

The continuous aspect of dynamical systems has been intensively studied through the development of asynchronous algorithms for solving PDE problems. We have focused our studies on the interest of GPUs in asynchronous algorithms [17]. Also, 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 sized clusters with efficient networks. Finally, we investigate other optimizations like load balancing. For this last subject, the SimGrid environment has revealed itself to be a precious tool to perform feasibility tests and benchmarks for this kind of algorithms on large scale systems. It has been successfully used to evaluate an asynchronous load balancing algorithm [37].

In 2011, the PhD thesis of Marion Guthmuller, supervised by M. Quinson and S. Contassot-Vivier, has started on the subject of model-checking distributed applications inside the SimGrid simulator [20]. This is also the opportunity of designing new tools to study more precisely the dynamics of discrete or continuous systems. See the simulation part in Section 6.3.2 for more details on this PhD.

6.2. Transparent resource management

6.2.1. Client-side cloud broker.

Integrating the 'pay-as-you-go' pricing model commonly used in IaaS clouds is an important question which profoundly changes the assumptions for job scheduling. From the observation that in most commercial solutions the price of a CPU cycle is identical, be the CPU a fast or slow one, several schedulings may be derived for a same price but with different makespans. Hence, in a context where resources can be started on-demand, scheduling strategies must include a decision process regarding the scaling (number of resources used) of the platform and the types of resources rented over time. In [24], we have studied the impact of these two factors on classic job scheduling strategies applied to bag-of-tasks workloads. The results show that shorter makespans can be achieved through scaling at no extra cost, while using quicker CPUs largely increases the price of the computations. More importantly, we show the difficulty to predict the outcomes of such decisions, which requires to design new provisioning approaches.

6.3. Experimental Methodologies

6.3.1. Overall improvement of SimGrid

2012 was the last year of the USS-SimGrid project granted by the ANR. We thus capitalized the results of the first project by properly releasing them in the public releases. Parallel simulation is now stable enough to be used in practice by our users. In addition, the framework is now able to simulate millions of processes without any particular settings in C. The java bindings were also improved to simulate several hundred thousand processes out of the box [25].

This year was also the first year of the SONGS project, also funded by the ANR. This project is much larger that the previous one, both in funding and targets. In surface, SONGS aims at increasing the scope of the SimGrid simulation framework by enabling the Cloud and HPC scenarios in addition to the existing Grid and P2P ones. Under the hood, it aims at providing new models specifically designed for these use cases, and also provide the necessary internal hooks so that users can modify the used models by themselves.

This project is well started, with three plenary meetings and a user conference organized over the year, but no new publication resulted of this work yet. The first work toward increasing the simulation versatility, initiated last year, was published this year[14]

6.3.2. Dynamic verification of liveness properties in SimGrid

A full featured model-checker is integrated to SimGrid since a few years, but it was limited to the verification of safety properties. We worked toward the verification of liveness properties in this framework. The key challenge is to quantify the state equality at state level, adding and leveraging introspection abilities to arbitrary C programs.

This constitutes the core of the PhD thesis of M. Guthmuller, started last year. A working prototype was developed during this year, described in an initial publication [20].

6.3.3. Grid'5000 and related projects

We continued to play a key role in the Grid'5000 testbed in 2012. Lucas Nussbaum, being delegated by the executive committee to follow the work of the technical team, was heavily involved in the recent evolutions of the testbed (network weathermaps, storage management, etc.) and in other activities such as the preparation of the Grid'5000 winter school. We were also involved in a publication [33] which is a follow-up to the workshop on *Supporting Experimental Computer Science* held during SC'11, and in another publication [32] describing the recent advances on the Grid'5000 testbed in order to support experiments involving virtualization at large scale.

More specifically, our involvement in the *OpenCloudWare* project led us to design several tools that ease the deployment of Cloud stacks on Grid'5000 for experimental purposes. Those tools were also used during an internship that was co-advised with the *Harmonic Pharma* start-up, exploring how complex bio-informatics workflows could be ported to the Cloud.

On the institutional side, we will also play a central role in the *Groupement d'Intérêt Scientifique* that is currently being set up, since Lucas Nussbaum is a member of both the *bureau* and of the *comité d'architectes*.

6.3.4. Distem – DISTributed systems EMulator

In the context of ADT Solfége, we continued our work on Distem. Three releases were made over the year, with several improvements and bug fixes, including support for variable CPU and network emulation parameters during an experiment. See http://distem.gforge.inria.fr/ for more information, or our paper accepted at PDP'2013 [26].

6.3.5. Kadeploy3 – scalable cluster deployment solution

Thanks to the support of ADT Kadeploy3, many efforts were carried out on Kadeploy3. Two releases were made, including many new features (many improvements to the handling of parallel commands and to the inner automaton for more fault-tolerant deployments; use of Kexec for faster deployments) as well as bug fixes.

Kadeploy3 was featured during several events (*journée 2RCE*, *SuperComputing 2012*), and in two publications: one unsuccessfully submitted to LISA'2012 [35], one accepted in USENIX *;login:* [13].

Finally, Kadeploy3 was also the basis of submissions to the *SCALE challenge held with CCGrid'2012*, of which we were finalists, and of the winner challenge entry at *Grid'5000 winter school 2012*.

6.3.6. Business workflows for the description and control of experiments

We are exploring the use of Business Process Modelling and Management for the description and the control of complex experiments. In [28], we outlined the required features for an experiment control framework, and described how business workflows could be used to address this issue. In [27] and [15], we described our early implementation of XPFlow, a experiment control engine relying on business workflows paradigms.

6.3.7. Towards Open Science for distributed systems research

One of our long term goal on experimental methodologies would be the advance of an Open Science in the research domain of Distributed Systems. Scientific tools would be sufficiently assessed and easily combined when necessary, and scientific experiments would be perfectly reproducible. These objectives are still very ambitious for the researches targeting distributed systems.

In order to precisely evaluate the path remaining toward these goals, and try addressing some of the challenges that they pose, we currently host Maximiliano Geier as an Inria intern. While most researchers try to answer brilliant scientific questions with simple scientific methodologies, he is asked to answer a simple question (on the adaptation of the BitTorrent protocol to high bandwidth networks) using an advanced scientific methodology. We are also surveying the experimental methodology used in top tier conferences to gain further insight on this topic.

In addition, we are organizing Realis, an event aiming at testing the experimental reproducibility of papers submitted to Compas'2013. Associated to the Compas'13 conference, this workshop aims at providing a place to discuss the reproducibility of the experiments underlying the publications submitted to the main conference. We hope that this kind of venue will motivate the researchers to further detail their experimental methodology, ultimately allowing others to reproduce their experiments.

AVALON Team

6. New Results

6.1. HPC Component Model

Participants: Zhengxiong Hou, Vincent Pichon, Christian Pérez.

6.1.1. L2C: A Low Level Component Model

We have proposed a low level component model (L^2C) that supports directly native connectors for typical scenarios of high performance computing, such as MPI, shared memory and method invocation [10]. We have applied it to a typical example of stencil computation, i.e., a 2-D Jacobi application with domain decomposition. The experimental results have shown that L^2C can achieve the equivalent performance as native implementations, while gaining benefits such as performance portability on the basis of the software component model.

6.1.2. Auto-tuning of Stencil Based Applications

We started modeling the performance of stencil applications on multi-core clusters. We focused in particular on a 2D Jacobi benchmark application and the NEMO application as well as memory bandwidth performance. We derived a tuning approach including data partitioning within one node, the selection of the number of threads within a multi-core node, a data partitioning for multi nodes, and the number of nodes for a multi-core cluster. This model is based on a set of experiments on machines of GRID'5000 and on Curie and Juqueen supercomputers. A paper presenting these results is in preparation.

6.2. Cooperative Resource Managers

Participants: Eddy Caron, Cristian Klein, Christian Pérez, Noua Toukourou.

6.2.1. Integration of SALOME with CooRM

We have continued the validation works of the CooRM RMS architecture [52]. To this end, we focused on the SALOME numerical simulation platform developed and used jointly by EDF and CEA. In 2012, we have mostly started the integration of CooRMv1 concepts in SALOME. CooRMv1 targets moldable applications and allows them to efficiently employ their custom resource selection algorithms. We have done the necessary changes in SALOME, thus obtaining a working prototype implementation. Thanks to this, SALOME applications could be published with a custom launcher (implementing a resource selection algorithm) so as to transparently launch applications efficiently, instead of having to leave this burden to the user.

6.2.2. A Distributed Resource Management Architecture for Moldable Applications

In 2011, we have proposed CooRMv1 [52], a centralized RMS architecture to efficiently support moldable applications. Having a centralized architecture is however undesirable for geographically-distributed resources such as Grids or multiple Clouds. For example, if there is a network failure, some users will not be able to access any resources, not even those that are located on their side of the bisection.

To this end, we extended CooRMv1 and proposed a distributed version of it, distCooRM, in collaboration with the Myriads team. It allows moldable applications to efficiently co-allocated resource managed by independent agents. Simulation results show that the approach is feasible and scales well for a reasonable number of applications. In other words, it presents good strong scalability, but not weak scalability, which we intend to address in future work.

6.2.3. A Resource Management Architecture for Fair Scheduling of Optional Computations

In collaboration with two teams from IRIT, we have identified a use-case that is currently badly supported. Some applications, such as Monte-Carlo simulations, contain optional computations: These are not critical, but completing them would improve the results. When executing these application on HPC resources, most resource managers, such as batch schedulers, require the user to submit a predefined number of computing requests. If the user submits too many requests, the platform might become overloaded, whereas if the user submits too few requests, then resources might be left idle.

To solve this issue, we proposed a resource management architecture, called DIET-ethic [42], which auto-tunes the number of optional requests. It improves user happiness, fairness and the number of completed requests, when compared to a system which does not support optional computations.

6.3. Large-Scale Data Management and Processing

Participants: José Saray, Bing Tang, Gilles Fedak, Anthony Simonet.

6.3.1. Data Management on Hybrid Distributed Infrastructure

The BITDEW framework addresses the issue of how to design a programmable environment for automatic and transparent data management on Grids, Clouds and Desktop Grids. BITDEW 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.

In collaboration with Mohamed Labidi, University of Sfax (Tunisia), we have developed a data-aware and parallel version of Magik, an application for Arabic writing recognition using the BITDEW 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. [20].

In 2012, we have also surveyed P2P strategies (replication, erasure code, replica repair, hybrid storage), which provide reliable and durable storage on top of hybrid distributed infrastructures composed of volatile and stable storage. Following these 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 [25].

6.3.2. MapReduce Programing Model for Desktop Grid

MapReduce is an emerging programming model for data-intense applications 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 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 achieved linear speedup on the classical WordCount benchmark. Several scenarios involving lagger hosts and host crashes demonstrated that the prototype is able to cope with an experimental context similar to real-world Internet [9].

In collaboration with the Huazhong University of Science & Technology (China), we have developed an emulation framework to assess MapReduce on Internet Desktop Grid. We have made extensive comparison on BitDew-MapReduce and Hadoop using GRID'5000 which show that our approach has all the properties desirable to cope with an Internet deployment, whereas Hadoop fails on several tests [22].

We have published a joint work in collaboration with Virginia Tech (USA), which is a presentation of two alternative implementations of MapReduce for Desktop Grids : Moon and Bitdew [37].

6.4. Computing on Hybrid Distributed Infrastructure

Participants: Simon Delamare, Gilles Fedak, José Saray, Anthony Simonet.

6.4.1. 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 EGI, 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 EGI applications (e.g., the fusion community's applications) that can tolerate such a long latency most of the user communities want much shorter deadlines.

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. In 2012, 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 [14], [21], [35]. The software have now been delivered to the partners and run in production in the European Desktop Grid Infrastructure.

6.4.2. Scheduling on Hybrid Distributed Computing Infrastructures

In collaboration with the Mircea Moca, from the Babes-Bolyai University of Cluj-Napoca (Romania), we have investigated new scheduling algorithms for pull-based scheduler, which relies on Promethee method. We have shown that these heuristics perform efficiently on three different kinds of infrastructures, namely Grids, Clouds and Desktop Grids [23].

6.5. Energy Efficiency in Large Scale Systems

Participants: Ghislain Landry Tsafack, Mohammed El Mehdi Diouri, Olivier Glück, Laurent Lefevre.

6.5.1. Energy Efficiency in HPC Systems

Modern high performance computing subsystems (HPC) – including processor, network, memory, and I/O — are provided with power management mechanisms. These include dynamic speed scaling and dynamic resource sleeping. Understanding the behavioral patterns of high performance computing systems at runtime can lead to a multitude of optimization opportunities including controlling and limiting their energy usage. We have proposed a general purpose methodology for optimizing energy performance of HPC systems considering processor, disk and network. We have relied on the concept of execution vector along with a partial phase recognition technique for on-the-fly dynamic management without any a priori knowledge of the workload. We have demonstrated the effectiveness of our management policy under two real-life workloads. Experimental results have shown that our management policy in comparison with baseline unmanaged execution saves up to 24% of energy with less than 4% performance overhead for our real-life workloads [28], [27], [26]. This work is done under the Large Scale Initiative Hemera project (Joint PhD between Avalon and IRIT (Toulouse) with J.-M. Pierson, P. Stolf and G. Da Costa).

6.5.2. Energy Considerations in Checkpointing and Fault Tolerance Protocols

Two key points should be taken into account in future exascale systems: fault tolerance and energy consumption. To address these challenges, we evaluated checkpointing and existing fault tolerance protocols from an energy point of view. We measured on a real testbed the power consumption of the main atomic operations found in these protocols: checkpointing, message logging and coordination. The results [16], [51] show that process coordination and RAM logging consume more power than checkpointing and HDD logging. However, the results we presented in Joules per Bytes for I/O operations, emphasize that checkpointing and HDD logging consume more energy than RAM logging because of the logging duration which is much more higher on HDD than on RAM . We have also shown that for identical nodes performing the same operation, the extra power cost due to this operation is the same. In general, we have learned that the power consumption of a node during a given operation remains constant during this operation. The power consumption of such a node is equal to its idle power consumption to which we add the extra power consumption due to the operation it is performing. Finally, we proposed to consider energy consumption as a criterion for the choice of fault tolerance protocols. In terms of energy consumption, we should promote message logging for applications exchanging small volumes of data and coordination for applications involving few processes. This work is a joint work with F. Cappello (Inria-UIUC-NCSA Joint Laboratory for Petascale Computing).

6.5.3. Towards a Smart and Energy-Aware Service-Oriented Manager for Extreme-Scale Applications

To address the issue of energy efficiency for exascale supercomputers, we proposed a smart and energy-aware service-oriented manager for exascale applications: SEASOMES [17]. This framework aggregates the various energy-efficient solutions to "consume less" energy and to "consume better". It involves both internal and external interactions with the various actors interfering directly or indirectly with the supercomputer. On the one hand, we recommended a more fine-grained collaboration between application and hardware resources in order to reduce energy consumption and provide sustainable exascale services. On the other hand, we suggested a cooperation between the user, the administrator, the resource manager and the energy supplier for the purpose of "consuming better".

6.6. Green-IT Innovation Analysis

Participant: Laurent Lefevre.

Green IT has recently appeared as a mandatory approach to take into account of energy efficiency in Information Technology. This research investigates the Green IT area and its opportunities for innovation. Main motivations for Green IT have been analyzed and we have proposed new definition of Green IT including social, environmental and economic concerns. We have proposed a new model of a virtuous circle that appears in Green IT: while Green IT has its own motivations, resulting research feeds other research field in a virtuous circle. Innovation in this particular sector paves the way for further innovation by means of original research not foreseen at first thoughts.

This analysis is joint work with IRIT (Toulouse - C. Herozog, J.-M. Pierson) [19].

6.7. Workflow Scheduling

Participants: Eddy Caron, Frédéric Desprez, Cristian Klein, Vincent Lanore, Sylvain Gault, Christian Pérez, Adrian Muresan, Frédéric Suter.

6.7.1. High-Level Waste Application Scheduling

Brought forward by EDF, a partner in the ANR COOP project, High-Level Waste is a multi-level application: It is composed of many moldable tasks, part of which are initially known. Some of these tasks may, with a certain probability, launch other tasks, which usually take longer. We have proposed several scheduling algorithms to optimize the performance of such applications, which are little studied in current literature. Experiments with simulations showed that considerable gains can be made, not only in terms of performance, but also performance portability. This work will be published in 2013 [31].

6.7.2. Elastic Scheduling for Functional Workflows

As a recent research direction we have focused on the development of an allocation strategy for budget-constrained workflow applications that target IaaS Cloud platforms. The workflow abstraction is very common amongst scientific applications. It is easy to find examples in any field from bioinformatics to geography. The reasons for the proliferation of workflow applications in science are various, from the building of applications on top of legacy code to modeling of applications that have an inherent workflow structure. The first workflow applications were composed of sequential tasks, but as computational units became more and more parallel, workflow applications have also evolved and are now formed of parallel tasks and, occasionally, parallel moldable tasks. The classic DAG structure of workflow applications has also changed as some applications need to perform refinement iteration, creating loop-like constructs.

We have considered a general model of workflow applications that permit non-deterministic transitions. We have elaborated two budget-constrained allocation strategies for this type of workflow. The problem is a bi-criteria optimization problem as we are optimizing both budget and workflow makespan [12].

For a practical validation of the work, we are currently working on the implementation of the budget-constrained scheduler as part of the Nimbus open source cloud platform. This is being tested with a cosmological simulation workflow application called *Ramses* (see Section 4.4). This is a parallel MPI application that, as part of this work, has been ported for execution on dynamic virtual platforms. This work has been done in the form of a two month internship at the Argonne National Laboratory, USA, under the guidance of Kate Keahey and has been accepted for poster presentation in the XSEDE 2012 conference.

6.7.3. Self-Healing of Operational Workflow Incidents on Distributed Computing Infrastructures

Distributed computing infrastructures are commonly used through scientific gateways, but operating these gateways requires important human intervention to handle operational incidents. We have designed a self-healing process that quantifies incident degrees of workflow activities from metrics measuring long-tail effect, application efficiency, data transfer issues, and site-specific problems. These metrics are simple enough to be computed online and they make little assumptions on the application or resource characteristics. From their degree, incidents are classified in levels and associated to sets of healing actions that are selected based on association rules modeling correlations between incident levels. We specifically study the long-tail effect issue, and propose a new algorithm to control task replication. The healing process is parametrized on real application traces acquired in production on the European Grid Infrastructure. Experimental results obtained in the Virtual Imaging Platform show that the proposed method speeds up execution up to a factor of 4, consumes up to 26% less resource time than a control execution and properly detects unrecoverable errors.

This work is done in collaboration with Tristan Glatard and Rafael Ferreira Da Silva from CREATIS (UMR5220).

6.7.4. Scheduling for MapReduce Based Applications

We have worked on scheduling algorithms for MapReduce applications in Grids and Clouds as we aim at providing resource-efficient and time-efficient scheduling algorithms. This work is mainly done within the scope of the Map-Reduce ANR project.

A deliverable presenting the heuristics for scheduling data transfers derived from a previous work by Berlinska and Drozdowsky has been written [50]. A section of a collaborative paper has been written and the paper has been presented at the ICA CON conference [9], [4]. The results of the aforementioned heuristics that has been previously implemented in a visualization / simulation tool, has been summarized in a paper accepted for RenPar. Moreover, these algorithms and heuristics have been implemented in the MapReduce framework HoMR.

6.8. Performance Evaluation and Modeling

Participants: Eddy Caron, Frédéric Desprez, Matthieu Imbert, Georges Markomanolis, Jonathan Rouzaud-Cornabas, Frédéric Suter.

6.8.1. Time-Independent Log Format

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 computed 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 improved the accuracy of the performance's prediction by applying different instrumentation configurations according to the requirements of our framework. Some performance issues of the executed benchmarks were taken under consideration for more accurate predictions. Also the simulator was reimplemented in order to have more accurate results and take advantage of the last SimGrid's simulation techniques. 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. For the extreme cases where we used many nodes by mapping a lot of processes per core, some issues were indicated that we are trying to solve in order to be able to apply our methodology with less overhead. Also we plan to predict the performance of more benchmarks.

6.8.2. Dynamic Network Forecasting

In distributed systems the knowledge of the network is mandatory to know the available connections and their performance. Indeed, to be able to efficiently schedule network transfers on computing platforms such as clusters, grids or clouds, accurate and timely predictions of network transfers completion times are needed. We designed a new metrology and performance prediction framework called Pilgrim which offers a service predicting the completion times of current and concurrent TCP transfers. This service uses SimGrid to simulate the network transfers. Ongoing work is to obtain experimental results comparing the predictions obtained from Pilgrim to the real transfer completion times.

6.8.3. Amazon EC2 simulation

During this year, we have developed an extension of SimGrid to simulate multi-platforms Clouds: SimGrid Cloud Broker (SGCB). It simulates the suite of services provided by Amazon AWS: EC2 for virtual machines, S3 for key-value storage and EBS for block storage. SGCB allows to easily evaluate different resource selection policy but also to simulate an entire application running on a set of resources that come from multiple Clouds. As the billing mechanism is a crucial feature of the Clouds, SGCB is able to simulate it. For this, we extended SimGrid in order to do the accounting of all virtual resources used. With this accounting, we are able to simulate the process of billing as Amazon does it. We are working to increase the accuracy of our performance models, and therefore the validity of the results for different use cases.

6.9. Cloud Resource Management

Participants: Eddy Caron, Frédéric Desprez, Arnaud Lefray, Jonathan Rouzaud-Cornabas, Julien Carpentier, Jean-Patrick Gelas, Laurent Lefevre, Maxime Morel, Olivier Mornard, Francois Rossigneux.

6.9.1. Resource Provisioning for Federations of Clouds

Since the visit of Jose Luis Lucas Simarro, we have established a collaboration with the Distributed Systems Architecture Research Group at Complutense University of Madrid (Spain) on resource brokering strategies for multiples Clouds. The purpose is to design new strategies that are able to migrate services from a Cloud to another one. VM migration is done to save money when the price of running a given VM change. Indeed, in modern Clouds such as Amazon EC2, Spot Instances have dynamic prices that change based on the law of supply and demand. Most of the current solutions only take into account the cost of computation when migrating services between Clouds. However, when a service is migrated, we need to pay network traffic between the two Clouds and the storage of the Virtual Machine image in both Clouds during the migration. We are studying trough simulations different resource selection algorithms that take into account the cost of all resources: compute, storage, and network.

6.9.2. Energy Efficient Clouds

Within the projects CompatibleOne (Open Source Cloud Broker) and XLcloud (Energy Efficiency in OpenStack based clouds), we explore the design of energy aware and energy efficient cloud infrastructures. Monitoring of physical and virtual resources is injected into cloud frameworks. Systems based on such metrics are designed in order to benefit from energy usage knowledge in virtual machines mapping and precise accounting [13].

6.9.3. User Isolation

Inter-VM and virtual network isolation is weak in terms of both security and performance. Accordingly, it can not guarantee performance, security and privacy requirements. This is a serious issue as most of clouds are multi-tenant and users do not trust each other. By improving the resource allocation process, we show how these issue can be solved and thus the overall security of the clouds improved. Moreover, we show how a Cloud Service Provider (CSP) can let the users express their security requirements. We show that isolation requirements have a cost for the Cloud Service Providers but they can bill requirements as an additional service. By doing so, they will have a new resource of income and the users trust in their platforms will increase as they can express security requirements.

6.9.4. Cloud Security

Mandatory Access Control is really poorly supported by Cloud environments. Our work proposes extensions of the OpenNebula Cloud in order to provide an advanced MAC protection of the virtual machines hosted by the different nodes of the Cloud. Thus, unique SELinx security labels are associated with the virtual machines and their resources. The instantiations and migrations of the virtual machines maintain those unique security labels. Moreover, PIGA-Virt provides a unified way to control the information flows within a virtual machine but also between multiple virtual machines. SELinux controls the direct flows. PIGA-Virt adds advanced controls. Thus, a PIGA protection rule can control several direct and indirect flows. The benchmarks of PIGA-Virt show that our Trusted OpenNebula Cloud is efficient regarding the quality of the protection.

This work is done in collaboration with Christian Toinard from LIFO/ENSI de Bourges.

6.10. Virtualizing Home Gateways at Large Scale

Participants: Jean-Patrick Gelas, Laurent Lefevre.

About 80-90% of the energy in today's wireline networks is consumed in the access network, with about 10 W per user being dissipated mostly by the customer premises equipment (CPE). Home gateway is a popular equipment deployed at the end of networks and supporting a set of heterogeneous services (from network to multimedia services). These gateways are difficult to manage for network operators and consume a lot of energy. This research explores the possibility to reduce the complexity of such equipment by moving services to some external dedicated and shared equipments. When combined to quasi passive CPE, this approach can reduce the energy consumption of wired networks infrastructures. This research is done within the GreenTouch initiative which aims to increase network energy efficiency by a factor of 1000 from current levels by 2015.

This work is done with collaboration with Addis Abeba University (Ethiopia) (M. Mulugeta and T. Assefa) [18].

6.11. Self-Adaptive Deployment

Participants: Eddy Caron, Maurice-Djibril Faye, Jonathan Rouzaud-Cornabas.

Software systems are increasingly expected to be self-adaptive. Such software systems have the capability to autonomously modify their behavior at run-time in response to changes in their environment. This capability may be included in the software systems at design time or later by external mechanisms. Therefore, along their development process multiple adaptation concerns must be considered, such as the response to changes in the utilization patterns, the need for alternative algorithms for implementing a function, or the diversity of the infrastructure. We have designed an architecture which aims to add self-adaptive capabilities to an existing middleware so that its deployment becomes self-adaptive. The framework uses external mechanisms for that purpose since this capability was not a native feature.

CEPAGE Project-Team

6. New Results

6.1. Resource allocation and Scheduling

6.1.1. 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 [22], 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. (Joint work with Loris Marchal from ENS Lyon)

6.1.2. Scheduling for Distributed Continuous Integration

Participants: Olivier Beaumont, Nicolas Bonichon, Ludovic Courtès.

In [21], we consider the problem of schedul- ing a special kind of mixed data-parallel applications arising in the context of continuous integration. Continuous integration (CI) is a software engineering technique, which consists in re- building and testing interdependent software components as soon as developers modify them. The CI tool is able to provide quick feedback to the developers, which allows them to fix the bug soon after it has been introduced. The CI process can be described as a DAG where nodes represent package build tasks, and edges represent dependencies among these packages; build tasks themselves can in turn be run in parallel. Thus, CI can be viewed as a mixed data-parallel application. A crucial point for a successful CI process is its ability to provide quick feedback. Thus, makespan minimization is the main goal. Our contribution is twofold. First, we provide and analyze a large dataset corresponding to a build DAG. Second, we compare the performance of several scheduling heuristics on this dataset.

6.1.3. Resource Allocation in Clouds

Participants: Olivier Beaumont, Lionel Eyraud-Dubois, Hejer Rejeb.

In [14], we consider the problem of assigning a set of clients with demands to a set of servers with capacities and degree constraints. The goal is to find an allocation such that the number of clients assigned to a server is smaller than the server's degree and their overall demand is smaller than the server's capacity, while maximizing the overall throughput. This problem has several natural applications in the context of independent tasks scheduling or virtual machines allocation. We consider both the *offline* (when clients are known beforehand) and the *online* (when clients can join and leave the system at any time) versions of the problem. We first show that the degree constraint on the maximal number of clients that a server can handle is realistic in many contexts. Then, our main contribution is to prove that even if it makes the allocation problem more difficult (NP-Complete), a very small additive resource augmentation on the servers degree is enough to find in polynomial time a solution that achieves at least the optimal throughput. After a set of theoretical results on the complexity of the offline and online versions of the problem, we propose several other greedy heuristics to solve the online problem and we compare the *performance* (in terms of throughput) and the *cost* (in terms of disconnections and reconnections) of all proposed algorithms through a set of extensive simulation results. (Joint work with Christopher Thraves-Caros, University of Madrid)

6.1.4. Non Linear Divisible Load Scheduling

Participants: Olivier Beaumont, Hubert Larchevêque.

Divisible Load Theory (DLT) has received a lot of attention in the past decade. A divisible load is a perfect parallel task, that can be split arbitrarily and executed in parallel on a set of possibly heterogeneous resources. The success of DLT is strongly related to the existence of many optimal resource allocation and scheduling algorithms, what strongly differs from general scheduling theory. Moreover, recently, close relationships have been underlined between DLT, that provides a fruitful theoretical framework for scheduling jobs on heterogeneous platforms, and MapReduce, that provides a simple and efficient programming framework to deploy applications on large scale distributed platforms. The success of both have suggested to extend their framework to non-linear complexity tasks. In [24], we show that both DLT and MapReduce are better suited to workloads with linear complexity. In particular, we prove that divisible load theory cannot directly be applied to quadratic workloads, such as it has been proposed recently. We precisely state the limits for classical DLT studies and we review and propose solutions based on a careful preparation of the dataset and clever data partitioning algorithms. In particular, through simulations, we show the possible impact of this approach on the volume of communications generated by MapReduce, in the context of Matrix Multiplication and Outer Product algorithms. (Joint work with Loris Marchal from ENS Lyon)

6.1.5. Reliable Service Allocation in Clouds

Participants: Olivier Beaumont, Lionel Eyraud-Dubois, Hubert Larchevêque.

In [23], we consider several reliability problems that arise when allocating applications to processing resources in a Cloud computing platform. More specifically, we assume on the one hand that each computing resource is associated to a capacity constraint and to a probability of failure. On the other hand, we assume that each service runs as a set of independent instances of identical Virtual Machines, and that the Service Level Agreement between the Cloud provider and the client states that a minimal number of instances of the service should run with a given probability. In this context, given the capacity and failure probabilities of the machines, and the capacity and reliability demands of the services, the question for the cloud provider is to find an allocation of the instances of the services (possibly using replication) onto machines satisfying all types of constraints during a given time period. The goal of this work is to assess the impact of the reliability constraint on the complexity of resource allocation problems. We consider several variants of this problem, depending on the number of services and whether their reliability demand is individual or global. We prove several fundamental complexity results (#P' and NP-completeness results) and we provide several optimal and approximation algorithms. In particular, we prove that a basic randomized allocation algorithm, that is easy to implement, provides optimal or quasi-optimal results in several contexts, and we show through simulations that it also achieves very good results in more general settings.

6.1.6. Optimizing Resource allocation while handling SLA violations in Cloud Computing platforms

Participants: Lionel Eyraud-Dubois, Hubert Larchevêque.

In [29], we study a resource allocation problem in the context of Cloud Computing, where a set of Virtual Machines (VM) has to be placed on a set of Physical Machines (PM). Each VM has a given demand (e.g. CPU demand), and each PM has a capacity. However, each VM only uses a fraction of its demand. The aim is to exploit the difference between the demand of the VM and its real utilization of the resources, to exploit the capacities of the PMs as much as possible. Moreover, the real consumption of the VMs can change over time (while staying under its original demand), implying sometimes expensive "SLA violations", corresponding to some VM's consumption not satisfied because of overloaded PMs. Thus, while optimizing the global resource utilization of the PMs, it is necessary to ensure that at any moment a VM's need evolves, a few number of migrations (moving a VM from PM to PM) is sufficient to find a new configuration in which all the VMs' consumptions are satisfied. We modelize this problem using a fully dynamic bin packing approach and we present an algorithm ensuring a global utilization of the resources of 66%. Moreover, each time a PM is overloaded at most one migration is necessary to fall back in a configuration with no overloaded PM, and only

3 different PMs are concerned by required migrations that may occur to keep the global resource utilization correct. This allows the platform to be highly resilient to a great number of changes.

6.2. Compact Routing

6.2.1. Compact routing with forbidden-set in planar graphs

Participant: Cyril Gavoille.

In [20], we consider fully dynamic $(1+\varepsilon)$ distance oracles and $(1+\varepsilon)$ forbidden-set labeling schemes for planar graphs. For a given n-vertex planar graph G with edge weights drawn from $[1, M]$ and parameter $\varepsilon > 0$, our forbidden-set labeling scheme uses labels of length $\lambda = O(\varepsilon^{-1} \log^2 n \log(nM) \cdot \max logn)$. Given the labels of two vertices s and t and of a set F of faulty vertices/edges, our scheme approximates the distance between s and t in $G \setminus F$ with stretch $(1+\varepsilon)$, in $O(|F|^2\lambda)$ time.

We then present a general method to transform $(1+\varepsilon)$ forbidden-set labeling schemas into a fully dynamic $(1+\varepsilon)$ distance oracle. Our fully dynamic $(1+\varepsilon)$ distance oracle is of size $O(n \log n \cdot \max logn)$ and has $\widetilde{O}(n^{1/2})$ query and update time, both the query and the update time are worst case. This improves on the best previously known $(1+\varepsilon)$ dynamic distance oracle for planar graphs, which has worst case query time $\widetilde{O}(n^{2/3})$ and amortized update time of $\widetilde{O}(n^{2/3})$.

Our $(1+\varepsilon)$ forbidden-set labeling scheme can also be extended into a forbidden-set labeled routing scheme with stretch $(1+\varepsilon)$.

6.2.2. Planar Spanner of geometric graphs

Participants: Nicolas Bonichon, Cyril Gavoille, Nicolas Hanusse.

In [26], we determine the stretch factor of L_1-Delaunay and L_∞-Delaunay triangulations, and we show that this stretch is $\sqrt{4+2\sqrt{2}} \approx 2.61$. Between any two points x, y of such triangulations, we construct a path whose length is no more than $\sqrt{4+2\sqrt{2}}$ times the Euclidean distance between x and y, and this bound is best possible. This definitively improves the 25-year old bound of $\sqrt{10}$ by Chew (SoCG '86).

To the best of our knowledge, this is the first time the stretch factor of the well-studied L_p-Delaunay triangulations, for any real $p \geq 1$, is determined exactly.

6.3. Mobile Agents

6.3.1. More efficient periodic traversal in anonymous undirected graphs

Participants: David Ilcinkas, Ralf Klasing.

In [15], we consider the problem of *periodic graph exploration* in which a mobile entity with constant memory, *an agent*, has to visit all n nodes of an input simple, connected, undirected graph in a periodic manner. Graphs are assumed to be anonymous, that is, nodes are unlabeled. While visiting a node, the agent may distinguish between the edges incident to it; for each node v, the endpoints of the edges incident to v are uniquely identified by different integer labels called *port numbers*. We are interested in algorithms for assigning the port numbers together with traversal algorithms for agents using these port numbers to obtain short traversal periods.

Periodic graph exploration is unsolvable if the port numbers are set arbitrarily; see Budach (1978). However, surprisingly small periods can be achieved by carefully assigning the port numbers. Dobrev *et al.* (2005) described an algorithm for assigning port numbers and an oblivious agent (i.e., an agent with no memory) using it, such that the agent explores any graph with n nodes within the period $10n$. When the agent has access to a constant number of memory bits, the optimal length of the period was proved in Gasieniec *et al.* (2008) to be no more than $3.75n - 2$ (using a different assignment of the port numbers and a different traversal algorithm). In our work, we improve both these bounds. More precisely, we show how to achieve a period length of at most $(4 + \frac{1}{3})n - 4$ for oblivious agents and a period length of at most $3.5n - 2$ for agents with constant memory. To obtain our results, we introduce a new, fast graph decomposition technique called a *three-layer partition* that may also be useful for solving other graph problems in the future. Finally, we present the first non-trivial lower bound, $2.8n - 2$, on the period length for the oblivious case.

6.3.2. Gathering of Robots on Anonymous Grids without Multiplicity Detection

Participant: Ralf Klasing.

In [28] , we study the gathering problem on grid networks. A team of robots placed at different nodes of a grid have to meet at some node and remain there. Robots operate in Look-Compute-Move cycles; in one cycle, a robot perceives the current configuration in terms of occupied nodes (Look), decides whether to move towards one of its neighbors (Compute), and in the positive case makes the computed move instantaneously (Move). Cycles are performed asynchronously for each robot. The problem has been deeply studied for the case of ring networks. However, the known techniques used on rings cannot be directly extended to grids. Moreover, on rings, another assumption concerning the so-called *multiplicity detection* capability was required in order to accomplish the gathering task. That is, a robot is able to detect during its Look operation whether a node is empty, or occupied by one robot, or occupied by an undefined number of robots greater than one.

In our work, we provide a full characterization about gatherable configurations for grids. In particular, we show that in this case, the multiplicity detection is not required. Very interestingly, sometimes the problem appears trivial, as it is for the case of grids with both odd sides, while sometimes the involved techniques require new insights with respect to the well-studied ring case. Moreover, our results reveal the importance of a structure like the grid that allows to overcome the multiplicity detection with respect to the ring case.

GRAND-LARGE Project-Team

5. New Results

5.1. Communication avoiding algorithms for linear algebra

Participants: Laura Grigori, 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.

Our 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) [18] [8]. Since then, we continue designing communication avoiding algorithm for other operations in both dense and sparse linear algebra. 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 2012, our research [43] has focused on the design of the LU decomposition with panel rank revealing pivoting (LU_PRRP), an LU factorization algorithm based on strong rank revealing QR panel factorization. LU_PRRP is more stable than Gaussian elimination with partial pivoting (GEPP), with a theoretical upper bound of the growth factor of $(1+\tau b)^{(n/b)-1}$, where b is the size of the panel used during the block factorization, τ is a parameter of the strong rank revealing QR factorization, and n is the number of columns of the matrix. For example, if the size of the panel is $b = 64$, and $\tau = 2$, then $(1+2b)^{(n/b)-1} = (1.079)^{n-1} \ll 2^{n-1}$, where 2^{n-1} is the upper bound of the growth factor of GEPP. Our extensive numerical experiments show that the new factorization scheme is as numerically stable as GEPP in practice, but it is more resistant to pathological cases. The LU_PRRP factorization does only $O(n^2 b)$ additional floating point operations compared to GEPP. We have also introduced CALU_PRRP, a communication avoiding version of LU_PRRP that minimizes communication. CALU_PRRP is based on tournament pivoting, with the selection of the pivots at each step of the tournament being performed via strong rank revealing QR factorization. CALU_PRRP is more stable than CALU, the communication avoiding version of GEPP, with a theoretical upper bound of the growth factor of $(1+\tau b)^{\frac{n}{b}(H+1)-1}$, where H is the height of the reduction tree used during tournament pivoting. The upper bound of the growth factor of CALU is $2^{n(H+1)-1}$. CALU_PRRP is also more stable in practice and is resistant to pathological cases on which GEPP and CALU fail.

Our work has also focused on designing algorithms that are optimal over multiple levels of memory hierarchy and parallelism. In [32] we present an algorithm for performing the LU factorization of dense matrices that is suitable for computer systems with two levels of parallelism. This algorithm is able to minimize both the volume of communication and the number of messages transferred at every level of the two-level hierarchy of parallelism. We present its implementation for a cluster of multicore processors based on MPI and Pthreads. We show that this implementation leads to a better performance than routines implementing the LU factorization in well-known numerical libraries. For matrices that are tall and skinny, that is they have many more rows than columns, our algorithm outperforms the corresponding algorithm from ScaLAPACK by a factor of 4.5 on a cluster of 32 nodes, each node having two quad-core Intel Xeon EMT64 processors.

5.2. 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 preconditoner is an incomplete decomposition that satisfies the filtering property [13]. 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 unsructured 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 [13]. But the main result is the development of a generalized method [10], [11] 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. In addition to finalizing this work, our research also focused on extending the block filtering factorization to include other approximation techniques that allowed us to introduce a parameter whose tuning permits to solve very difficult problems.

5.3. MIcrowave Data Analysis for petaScale computers

Participants: Laura Grigori, Mikolaj Szydlarski, Meisam Shariffy.

Generalized least square problems with non-diagonal weights arise frequently in an estimation of two dimensional images from data of cosmological as well as astro- or geo- physical observations. As the observational data sets keep growing at Moore's rate, with their volumes exceeding tens and hundreds billions of samples, the need for fast and efficiently parallelizable iterative solvers is generally recognized.

In this work [36] we propose a new iterative algorithm for solving generalized least square systems with weights given by a block-diagonal matrix with Toeplitz blocks. Such cases are physically well motivated and correspond to measurement noise being piece-wise stationary – a common occurrence in many actual observations. Our iterative algorithm is based on the conjugate gradient method and includes a parallel two-level preconditioner (2lvl-PCG) constructed from a limited number of sparse vectors estimated from the coefficients of the initial linear system.

Our prototypical application is the map-making problem in the Cosmic Microwave Background data analysis. We show experimentally that our parallel implementation of 2lvl-PCG outperforms by a factor of up to 6 the standard one-level PCG in terms of both the convergence rate and the time to solution on up to $12,228$ cores of NERSC's Cray XE6 (Hopper) system displaying nearly perfect strong and weak scaling behavior in this regime.

5.4. 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 [14].

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 [16]. 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 [54]& [25]. 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.

5.5. 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.

5.6. 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.

5.7. Non-self-stabilizing and self-stabilizing gathering in networks of mobile agents–the notion of speed

Participants: Joffroy Beauquier, Janna Burman, Julien Clment, 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.

5.8. 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.

5.9. 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.

5.10. 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.

5.11. 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.

5.12. 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 salf-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.

5.13. Large Scale Peer to Peer Performance Evaluations

Participant: Serge Petiton.

5.13.1. 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 [82]. We adapted [63] 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.

5.13.2. 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 [93] 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.

5.13.3. 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 eigensolver [64]. 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.

5.14. 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).

5.15. Emulation of Volatile Systems

Participants: Thomas Largillier, Benjamin Quetier, Sylvain Peyronnet, Thomas Hérault, Franck Cappello.

In the process of developping 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 experimation 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 developping 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.

5.16. 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 became 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 adress 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.

HIEPACS Project-Team

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/). We have used the StarPU runtime system to ensure the portability of our algorithms and codes. We have also investigated the case of partial pivoting LU factorization. The pivot selection induces a large number of low granularity tasks which are a potential bottleneck when handled with a runtime system; we have thus designed methods which aim at limiting the number of tasks.

6.1.2. Task-based Conjugate-Gradient for multi-GPUs platforms

Whereas most today parallel High Performance Computing (HPC) software is written as highly tuned code taking care of low-level details, the advent of the manycore area forces the community to consider modular programming paradigms and delegate part of the work to a third party software. That latter approach has been shown to be very productive and efficient with regular algorithms, such as dense linear algebra solvers. In this paper we show that such a model can be efficiently applied to a much more irregular and less compute intensive algorithm. We illustrate our discussion with the standard unpreconditioned Conjugate Gradient (CG) that we carefully express as a task-based algorithm. We use the StarPU runtime system to assess the efficiency of the approach on a computational platform consisting of three NVIDIA Fermi GPUs. We show that almost optimum speed up (up to 2.89) may be reached (relatively to a mono-GPU execution) when processing large matrices and that the performance is portable when changing the low-level memory transfer mechanism. This work is developed in the framework of the PhD of Stojce Nakov.

6.1.3. Resilience in numerical simulations

Various interpolations strategies to handle restarting Krylov subspace methods in case of core faults have been investigated. The underlying idea is to recover fault entries of the iterate via interpolation from existing values available on neighbor cores. In particular, we design a scheme that enables to preserve the key property of GMRES that is the residual norm monotonicity of the iterates even when failures occur. This work is developed in the framework of Mawussi Zounon's PhD funded by the ANR-RESCUE. Notice that theses activities are also part of our contribution to the G8-ECS (Enabling Climate Simulation at extreme scale).

6.1.4. Block GMRES method with inexact breakdowns and deflated restarting

We have considered the solution of large linear systems with multiple right-hand sides using a block GMRES approach. We designed a new algorithm that effectively handles the situation of almost rank deficient block generated by the block Arnoldi procedure and that enables the recycling of spectral information at restart. The first feature is inherited from an algorithm introduced by Robbé and Sadkane [M. Robbé and M. Sadkane. Exact and inexact breakdowns in the block gmres method. Linear Algebra and its Applications, 419: 265-285, 2006.], while the second one is obtained by extending the deflated restarting strategy proposed by Morgan [R. B. Morgan. Restarted block GMRES with deflation of eigenvalues. Applied Numerical Mathematics, 54(2): 222-236, 2005.]. Through numerical experiments, we have shown that the new algorithm combines the attractive numerical features of its two parents that it outperforms. This work was developed in the framework of the post-doc position of Yan-Fei Jing.

6.1.5. Scalable numerical schemes for scientific applications

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. Perfectly Matched Layers has been designed to cope with the designed numerical scheme and a software prototype for 2D simulation has been implemented.

The calculation of acoustic modes in combustion chambers is a challenging calculation for large 3D geometries. It requires the parallel calculation of a few of the smallest eigenpairs of large unsymmetric matrices in a nonlinear iterative scheme. Various numerical techniques have been considered to attempt recycling spectral information from one nonlinear step to the next that includes Jacobi-Davidson, Krylov-Schur and block Krylov-Schur algorithms. This is part of the PhD research activity of Pablo Salas.

6.1.6. Fast Multipole Methods

Concerning the Fast Multipole Method, our prototype called ScalFMM was completely rewritten in order to easily add new features. There is two main parts: the management of the octree and the parallelization of the method and kernels. This new architecture allow us to easily add new FMM algorithm or kernels and new paradigm of parallelization. The limitation of the classical FMM was that we need all operators (P2M, M2M, M2L, L2L, L2P) on the multipole expansions if we want to add a new kernel. To overcome this and in the context of associated team FastLA, we introduced the black-box FMM algorithm that allow us to be now kernel independent.

6.1.6.1. Optimizations for the M2L operator of the Chebyshev Fast Multipole Method

Most Fast Multipole Methods (FMM) have been developed and optimized for specific kernel functions. Our goal is to improve the efficiency of an FMM that is kernel function independent. The formulation is based on a Chebyshev interpolation scheme and has been studied for asymptotically smooth kernel functions G(x,y) and also for oscillatory ones, such as K(x,y) = G(x,y) exp(ik|x-y|). Two weak points of this formulation are the expensive precomputation of the M2L operators and the higher computational intensity compared to other FMMs. We focused our recent research on these issues. We have come up with a set of optimizations that exploit symmetries far-field interactions and blocking schemes that pave the road for highly optimized matrix-matrix product implementations. Recall, the scope of the FMM as an algorithm to perform fast matrix-vector products (Ax = y) may be twofold: on one hand the result (y) and on the other hand the solution (x). A fast precomputation is crucial in the first and fast running times in the second case. We proposed optimizations that provide more than 1000 times faster precomputation, much less memory requirement and much faster running times than before. All these results are submitted in Journal of computational Physics [27].

6.1.6.2. Pipelining the Chebyshev Fast Multipole Method over a runtime system

Fast Multipole Method are a fundamental operation for the simulation of many physical problems. The high performance design of such methods usually requires to carefully tune the algorithm for both the targeted physics and the hardware. For the Chebyshev Fast Multipole Method (black-box FMM) we have proposed a new approach that achieves high performance across heterogeneous architectures. Our method consists of expressing the Fast Multipole Method algorithm as a task flow and employing a state-of-the-art runtime system, StarPU, in order to process the tasks on the different processing units. We carefully design the task flow, the mathematical operators, their Central Processing Unit (CPU) and Graphics Processing Unit (GPU) implementations, as well as scheduling schemes. We compute potentials and forces of 200 million particles in 48.7 seconds on a homogeneous 160 cores SGI Altix UV 100 and of 30 million particles in 10.9 seconds on a heterogeneous 12 cores Intel Nehalem processor enhanced with 3 Nvidia M2090 Fermi GPUs. All these results are available in [24].

6.2. Efficient algorithmics for code coupling in complex simulations

Dynamic load balancing is an important step conditioning the performance of parallel adaptive codes whose load evolution is difficult to predict. Most of the studies which answer this problem perform well, but are limited to an initially fixed number of processors which is not modified at runtime. These approaches can be very inefficient, especially in terms of resource consumption, as demonstrated by Iqbal et al. As computation progresses, the global workload may increase drastically, exceeding memory limit for instance. In such a case, we argue it should be relevant to adjust the number of processors while maintaining the load balanced. However, this is still an open question that we currently focus on.

To overcome this issue, we propose a new graph repartitioning algorithm, which accepts a variable number of processors, assuming the load is already balanced. We call this problem the $M \times N$ graph repartitioning problem, with M the number of former parts and N the number of newer parts. Our algorithm minimizes both data communication (i.e., cut size) and data migration overheads, while maintaining the computational load balance in parallel. This algorithm is based on a theoretical result, that constructs optimal communication matrices with both a minimum migration volume and a minimum number of communications. It uses recent graph/hypergraph partitioning techniques with fixed vertices in a similar way than the one used in Zoltan for dynamic load-balancing of adaptive simulations. We validate this work for a large variety of real-life graphs (i.e., university of Florida sparse matrix collection), comparing it against state-of-the-art partitioners (Metis, Scotch, Zoltan).

We are considering several perspectives to our work. First, we focus on graph repartitioning in the more general case where both the load and the number of processors vary. We expect this work to be really suitable for next generation of adaptive codes. Finally, to be useful in real-life applications, our algorithm needs to work in parallel, that mainly requires to use a direct *k-way* parallel partitioning software that handle fixed vertices, like *Scotch*. This should allow us to partition much larger graph in larger part number. As another perspective, this approach can be relevant in the context of code coupling: e.g., if one code becomes more computationally intensive relatively to the other, it could be valuable to dynamically migrate some processor resources to the other code, and thus to equilibrate the whole coupled application. This work is currently conducted in the framework of Clément Vuchener PhD thesis and should be defended in september 2013.

6.3. Distributed Shared Memory approach for the steering of parallel simulations

As a different approach of EPSN, we recently propose in the thesis of J. Soumagne *an in-situ visualization approach for parallel coupling and steering of simulations through distributed shared memory files (DSM)*. Indeed, as simulation codes become more powerful and more interactive, it is desirable to monitor a simulation in-situ, performing not only visualization but also analysis of the incoming data as it is generated. Monitoring or post-processing simulation data in-situ has obvious advantage over the conventional approach of saving to – and reloading data from – the file system; the time and space it takes to write and then read the data from disk is a significant bottleneck for both the simulation and subsequent post-processing steps. Furthermore, the simulation may be stopped, modified, or potentially steered, thus conserving CPU resources.

In this thesis, we propose a loosely coupled approach that enables a simulation to transfer data to a visualization server via the use of in-memory files. We show in this study how the interface, implemented on top of a widely used hierarchical data format (HDF5), allows us to efficiently decrease the I/O bottleneck by using efficient communication and data mapping strategies. For steering, we present an interface that allows not only simple parameter changes but also complete re-meshing of grids or operations involving regeneration of field values over the entire computational domain to be carried out. This approach is generic enough so that no particular knowledge of the underlying model is required and a user can therefore plug any simulation to this framework without any re-compilation work.

A scalability study have demonstrated the performance of this solution up to 2048 cores on a Cray machine. Finally, the environment has been validated on two industrial test cases: the first one is developed by Ecole Centrale de Nantes and HydrOcean and an object placed into a wave maker is dynamically modified and steered, thereby making use of the re-meshing capabilities introduced by the framework; and the other is

developed by Ecole Centrale de Lyon and ANDRITZ HYDRO, a Pelton turbine is dynamically controlled and results are analyzed in-situ.

This thesis has been defended by J. Soumagne in december 2012. This work was supported by NextMuSE project receiving funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 225967. It has been realized in collaboration with the Swiss National Supercomputing Centre (CSCS). J.

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 bellow.

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, the code called *fast*, is released of the inria's gforge.

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.

these results were presented in [29], [30].

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 focused on the definition of efficient cache aware data structure to manage points and segments. All the algorithms have been adapted to this structure and we have stared the development of the OPTIDIS prototype. This prototype has been parallelized with OpenMP model. More physics will be added by our partners that will give us the capability to grow our simulation and run some meaningful benchmark.

We will work in three directions. Firstly, we will investigate how to adapt our fast multipole method to compute constraints and then forces in the context of FastLA associated team. Secondly, we will improved the displacement of the segments and the way to treat collision in parallel. Finally, we will move on hybrid parallelism for our prototype.

KERDATA Project-Team

6. New Results

6.1. Optimizing MapReduce processing

6.1.1. Hybrid infrastructures

Participants: Alexandru Costan, Bharath Vissapragada, Gabriel Antoniu.

As Map-Reduce emerges as a leading programming paradigm for data-intensive computing, today's frameworks which support it still have substantial shortcomings that limit its potential scalability. At the core of Map-Reduce frameworks stays a key component with a huge impact on their performance: the storage layer. To enable scalable parallel data processing, this layer must meet a series of specific requirements. An important challenge regards the target execution infrastructures. While the Map-Reduce programming model has become very visible in the cloud computing area, it is also subject to active research efforts on other kinds of large-scale infrastructures, such as desktop grids. We claim that it is worth investigating how such efforts (currently done in parallel) could converge, in a context where large-scale distributed platforms become more and more connected together.

In 2012 we investigated several directions where there is room for such progress: they concern storage efficiency under massive data access concurrency, scheduling, volatility and fault-tolerance. We placed our discussion in the perspective of the current evolution towards an increasing integration of large-scale distributed platforms (clouds, cloud federations, enterprise desktop grids, etc.) ([16]). We proposed an approach which aims to overcome the current limitations of existing Map-Reduce frameworks, in order to achieve scalable, concurrency-optimized, fault-tolerant Map-Reduce data processing on hybrid infrastructures. We are designing and implementing our approach through an original architecture for scalable data processing: it combines two approaches, BlobSeer and BitDew, which have shown their benefits separately (on clouds and desktop grids respectively) into a unified system. The global goal is to improve the behavior of Map-Reduce-based applications on the target large-scale infrastructures. The internship of Bharath Vissapragada was dedicated to this topic.

This approach will be evaluated with real-life bio-informatics applications on existing Nimbus-powered cloud testbeds interconnected with desktop grids.

6.1.2. Scheduling: Maestro

Participants: Shadi Ibrahim, Gabriel Antoniu.

As data-intensive applications became popular in the cloud, data-intensive cloud systems call for empirical evaluations and technical innovations. We have investigated some performance limits in current MapReduce frameworks (Hadoop in particular). Our studies reveal that the current Hadoop's scheduler for map tasks is inadequate, as it disregards replicas distributions. It causes performance degradation due to a high number of non-local map tasks, which in turn causes too many needless speculative map tasks and leads to imbalanced execution of map tasks among data nodes. We addressed these problems by developing a new map task scheduler called Maestro.

In [19], we developed a scheduling algorithm (Maestro) to alleviate the nonlocal map tasks executions problem of MapReduce. Maestro is conducive to improving the locality of map tasks executions efficiency by virtue of the finer-grained replica aware execution of map tasks, thereby having one additional factor for the chunk hosting status: the expected number of map tasks executions to be launched. Maestro keeps track of the chunks' locations along with their replicas' locations and the number of other chunks hosted by each node. In doing so, Maestro can efficiently schedule the map task to the node with minimal impacts on other nodes' local map tasks executions. Maestro schedules the map tasks in two waves: first, it fills the empty slots of each data node based on the number of hosted map tasks and on the replication scheme for their input data; second, runtime

scheduling takes into account the probability of scheduling a map task on a given machine depending on the replicas of the task's input data. These two waves lead to a higher locality in the execution of map tasks and to a more balanced intermediate data distribution for the shuffling phase.

We evaluated Maestro through a set of experiments on the Grid'5000 [35] testbed. Preliminary results [19] show the efficiency and scalability of our proposals, as well as additional benefits brought forward by our approach.

6.1.3. Fault tolerance

Participants: Bunjamin Memishi, Shadi Ibrahim, Gabriel Antoniu.

The simple philosophy of MapReduce has made huge community interest for its exploration, especially in environments where data-intensive applications are primary concern. Fault tolerance is one of the key features of the MapReduce system. MapReduce tasks are re-executed in case of failure, and a potential failure of a single master causes an additional bottleneck. It is observed that the detection of the failed worker tasks in Hadoop have a certain delay, yet not solved. Willing to improve the applications performance and optimal resource utilization, both of this concerns were more than a motivation so that we show in [36] that a little attention has been devoted to the failure detection in Hadoop's MapReduce which currently uses a timeout based mechanism for detecting failed tasks.

We have performed an in-depth analysis of MapReduce's failure detection, and these preliminary studies have revealed that the current static timeout value (600 seconds) is not adequate and demonstrate significant variations in the application's response time with different timeout value. Moreover, in the presence of single machine failure, the applications latencies vary not only in accordance to the occupancy time of the failure, similar to [33], but also vary with the job length (short or long).

Based on our aforementioned micro-analysis of failure detection in MapReduce, we are currently investigating an adaptive failure detection mechanism for Hadoop, which basically addresses the timeout adjustment in real-time for different jobs and applications, so that finally to adjust this model into a Shared Hadoop Cluster. Another work should discuss in details different failures types in MapReduce system and survey the different mechanisms used in MapReduce for detecting, handling and recovering from these failures and their inherited pros and cons; additionally, to a particular interest will be the analyzing of different execution environments including Cluster, Cloud and Desktop Grid on the efficiency of fault-tolerance in MapReduce. This work will soon be published.

6.2. A-Brain and TomusBlobs

6.2.1. TomusBlobs

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Enabling high-throughput massive data processing on cloud data becomes a critical issue, as it impacts the overall application performance. In the framework of the MSR-Inria A-Brain co-led by Gabriel Antoniu (KerData) and Bertrand Thirion (PARIETAL), the TomusBlobs[22] system was designed and implemented by KerData to address such challenges at the level of the cloud storage. The system we introduce is a concurrency-optimized data storage system which federates the virtual disks associated to VMs. As TomusBlobs does not require modifications to the cloud middleware, it can serve as a high-throughput globally-shared data storage for the cloud applications that require data passing among computation nodes.

We leveraged the performance of this solution to enable efficient data-intensive processing on commercial clouds by building an optimized prototype MapReduce framework for Azure. The system, deployed on 350 cores in Azure, was used to execute a real-life application, A-Brain with the goal of searching for significant associations between brain locations and genes.

The achieved throughput increased with an order of 2 for reading, respectively 3 for writing compared to the remote storage. With our approach for MapReduce data processing, the computation time is reduced to 50 % compared to the existing solutions, while the cost is reduced up to 30 %.

6.2.2. Iterative MapReduce

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu, Louis-Claude Canon.

While MapReduce has arisen as a major programming model for data analysis on clouds, there are many scientific applications that require processing patterns different from this paradigm. As such, reduce-intensive algorithms are becoming increasingly useful in applications such as data clustering, classification and mining. These algorithms have a common pattern: data are processed iteratively and aggregated into a single final result. While in the initial MapReduce proposal the reduce phase was a simple aggregation function, recently an increasing number of applications relying on MapReduce exhibit a reduce-intensive pattern, that is, an important part of the computations are done during the reduce phase. However, platforms like MapReduce or Dryad lack built-in support for reduce-intensive workloads.

To overcome these issues, we introduced MapIterativeReduce [23], a framework which: 1) extends the MapReduce programming model to better support reduce-intensive applications by exploiting the inherent parallelism of the reduce tasks which have an associative and/or commutative operation; and 2) substantially improves their efficiency by eliminating the implicit barrier between the Map and the Reduce phase. We showed how to leverage this architecture for scientific applications by enhancing the fault tolerance support in Azure and TomusBlobs, the underlying storage system, with a light checkpointing scheme and without any centralized control.

We evaluated MapIterativeReduce on the Microsoft Azure cloud with synthetic benchmarks and with a real-life application. Compared to state-of-art solutions, our approach enables faster data processing, by reducing the execution times by up to 75 %.

6.2.3. Adaptive file management for clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Recently, there is an increasing interest to execute general data processing schemas in clouds, as it would allow many scientific applications to migrate to this computing infrastructures. The natural way to do this is to designe and adopt Workflow Processing engines built for clouds. Such workflow processing in clouds would involve data propagation on the computation nodes based on well defined data access patterns. Having an efficient file management backend for a workflow engines is thus essential as we move to the world of BigData.

We proposed a new approach for a transfer-optimized file management in clouds On the one hand, our solution manages files within the deployment leveraging data locality. On the other hand, we envision an adaptive system that adopts the transfer method most suited based on the data transfer context.

The performance evaluation showed significant gains in terms of transfer throughput and computation time. File transfer times are reduced up to a factor of 5 with respect to the remote storage, while the timespan of running applications is reduced by more than 25% compared with other frameworks like Hadoop on Azure. This work was done in the context of a 3-month internship of Radu Tudoran hosted by the Advance Technology Lab from Microsoft Europe, Germany, Aachen.

6.3. Autonomic Cloud data storage management

Participants: Gabriel Antoniu, Alexandru Costan.

Providing the users with the possibility to store and process data on externalized, virtual resources from the cloud requires simultaneously investigating important aspects related to security, efficiency and quality of service. To this purpose, it clearly becomes necessary to create mechanisms able to provide feedback about the state of the storage system along with the underlying physical infrastructure. This information thus monitored, can further be fed back into the storage system and used by self-managing engines, in order to enable an autonomic behavior, possibly with several goals such as self-configuration, self-optimization, or self-healing. Within the DataCloud@work Associate Team in partnership with Politehnica University of Bucharest, our goal was to bring substantial contributions in this direction by leveraging previous efforts materialized through the BlobSeer data-sharing platform and several large-scale applications.

6.3.1. Evaluating BlobSeer for sharing application data on IaaS cloud infrastructures

. We showed how several types of large scale applications (e.g. scientific data aggregation, context-aware data management, video and image processing) rely on BlobSeer's support for high concurrency and increased data access throughput in order to achieve their goals. Several building blocks were implemented to address all the applications' requirements (new meta-data management, extended clients). An illustrative class of applications is represented by the context-aware ones. Our goal was to provide a cloud-based storage layer for sensitive context data, collected from a vast amount of sources: from smartphones to sensors located in the environment. We developed a layer on top of BlobSeer to allow two major things: efficient access to data based on meta-information (a catalogue of context data), and the support from mobility in the form of distributed caches able to support the movement of people and give support for fast access to real-time event of interest (dissemination of events of interest). The system as a whole was evaluated in extensive experiments, involving thousands of simulated clients, and the results proved its valuable contribution to advance the current state-of-the-art in the area of interested (middlewares to support context-aware apps).

6.3.2. Fault-tolerant VM management in Clouds, using BlobSeer

. We were also concerned about the fault tolerance support for the aforementioned applications on the cloud. A first step towards this goal consisted in exploring ways to deploy, boot and terminate VMs very quickly, enabling cloud users to exploit elasticity to find the optimal trade-off between the computational needs (number of resources, usage time) and budget constraints. We built a VM management system based on the FUSE interface leveraging the high throughput under increased concurrency of BlobSeer. We integrated it within the Nimbus cloud to allow fast VM deployment / snapshotting/ live migration. An adaptive prefetching mechanism is used to reduce the time required to simultaneously boot a large number of VM instances on clouds from the same initial VM image (multi-deployment). This proposal does not require any foreknowledge of the exact access pattern. It dynamically adapts to it at run time, enabling the slower instances to learn from the experience of the faster ones. Since all booting instances typically access only a small part of the virtual image along almost the same pattern, the required data can be pre-fetched in the background. In parallel, we investigated ways to ensure the anonimity of the data management layer, a requirement for HPC applications deployed into the clouds.

6.4. Advanced techniques for scalable cloud storage

6.4.1. Adaptive consistency

Participants: Houssem-Eddine Chihoub, Shadi Ibrahim, Gabriel Antoniu.

In just a few years cloud computing has become a very popular paradigm and a business success story, with storage being one of the key features. To achieve high data availability, cloud storage services rely on replication. In this context, one major challenge is data consistency. In contrast to traditional approaches that are mostly based on strong consistency, many cloud storage services opt for weaker consistency models in order to achieve better availability and performance. This comes at the cost of a high probability of stale data being read, as the replicas involved in the reads may not always have the most recent write. In [17], we propose a novel approach, named Harmony, which adaptively tunes the consistency level at run-time according to the application requirements. The key idea behind Harmony is an intelligent estimation model of stale reads, allowing to elastically scale up or down the number of replicas involved in read operations to maintain a low (possibly zero) tolerable fraction of stale reads. As a result, Harmony can meet the desired consistency of the applications while achieving good performance. We have implemented Harmony and performed extensive evaluations with the Cassandra cloud storage on Grid'5000 testbed and on Amazon EC2. The results show that Harmony can achieve good performance without exceeding the tolerated number of stale reads. For instance, in contrast to the static eventual consistency used in Cassandra, Harmony reduces the stale data being read by almost 80%. Meanwhile, it improves the throughput of the system by 45% while maintaining the desired consistency requirements of the applications when compared to the strong consistency model in Cassandra.

While most optimizations efforts for consistency management in the cloud focus on how to provide adequate trade-offs between consistency guarantees and performance, a little work has been investigating the impact of consistency on monetary cost. However, and since strict strong consistency is not always required for large class of applications, in [25] we argue that monetary cost should be taken into consideration when evaluating or selecting a consistency level in the cloud. Accordingly, we define a new metric called consistency-cost efficiency. Based on this metric, we present a simple, yet efficient economical consistency model, called Bismar, that adaptively tunes the consistency level at run-time in order to reduce the monetary cost while simultaneously maintaining a low fraction of stale reads. Experimental evaluations with the Cassandra cloud storage on a Grid'5000 testbed show the validity of the metric and demonstrate the effectiveness of the proposed consistency model allowing up to 31 % of money saving while tolerating a very small fraction of stale reads.

6.4.2. In-memory data management

Participants: Viet-Trung Tran, Gabriel Antoniu, Luc Bougé.

As a result of continuous innovation in hardware technology, computers are made more and more powerful than their prior models. Modern servers nowadays can possess large main memory capability that can size up to 1 Terabytes (TB) and more. As memory accesses are at least 100 times faster than disk, keeping data in main memory becomes an interesting design principle to increase the performance of data management systems. We design DStore [27], a document-oriented store residing in main memory to fully exploit high-speed memory accesses for high performance. DStore is able to scale up by increasing memory capability and the number of CPU-cores rather than scaling horizontally as in distributed data-management systems. This design decision favors DStore in supporting fast and atomic complex transactions, while maintaining high throughput for analytical processing (read-only accesses). This goal is (to our best knowledge) not easy to achieve with high performance in distributed environments.

To achieve its goals, DStore is built with several design principles. DStore follows a single threaded execution model to execute update transactions sequentially by one *master thread* while relying on a versioning concurrency control to enable multiple *reader threads* running simultaneously. DStore builds indexes for fast document lookups. Those indexes are built using the *delta-indexing* and *bulk updating* mechanisms for faster indexes maintenance and for atomicity guarantees of complex queries. Moreover, DStore is designed to favor stale reads that only need to access isolated snapshots of the indexes. Thus, it can eliminate interference between transactional processing and analytical processing.

We conducted multiple synthetic benchmarks on the Grid'5000 to evaluate the DStore prototype. Our preliminary results demonstrated that DStore achieved high performance even in scenarios where *Read*, *Insert* and *Delete* queries were performed simultaneously. In fact, the processing rate measured was about 600,000 operations per second for each concurrent process.

6.4.3. Scalable geographically distributed storage systems

Participants: Viet-Trung Tran, Gabriel Antoniu, Luc Bougé.

To build a globally scalable distributed file system that spreads over a wide area network (WAN), we propose an integrated architecture for a storage system relying on a distributed metadata-management system and BlobSeer, a large-scale data-management service. Since BlobSeer was initially designed to run on cluster environments, it is necessary to extend BlobSeer in order to take into account the latency hierarchy on geographically distributed environments.

We proposed BlobSeer-WAN, an extension of BlobSeer optimized for geographically distributed environments. First, in order to keep metadata I/O local to each site as much as possible, we proposed an asynchronous metadata replication scheme at the level of metadata providers. As metadata replication is asynchronous, we guarantee a minimal impact on the writing clients that generate metadata. Second, we introduced a distributed version management in BlobSeer-WAN by leveraging an implementation of multiple version managers and using vector clocks for detection and resolution of collision. This extension to BlobSeer keeps BLOBs consistent while they are globally shared among distributed sites under high concurrency.

Several experiments were performed on the Grid'5000 testbed demonstrated that BlobSeer-WAN can offer scalable aggregated throughput when concurrent clients append to one BLOB. The aggregated throughput reached to 1400 MB/s for 20 concurrent clients. We also compared BlobSeer-WAN and the original BlobSeer in local site accesses. The experiments shown that the overhead of the multiple version managers implementation and the metadata replication scheme in BlobSeer-WAN is minimal, thanks to our asynchronous replication scheme.

6.5. Scalable I/O for HPC

6.5.1. Damaris and HPC visualization

Participants: Matthieu Dorier, Gabriel Antoniu.

In the context of the Joint Inria/UIUC/ANL Laboratory for Petascale computing (JLCP), have proposed the Damaris approach to enable efficient I/O, data analysis and visualization at ver large scale from SMP machines. The I/O bottlenecks already present on current petascale systems as well as the amount of data written by HPC applications force to consider new approaches to get insights from running simulations. Trying to bypass the storage or drastically reducing the amount of data generated will be of outmost importance for exascale. In-situ visualization has therefor been proposed to run analysis and visualization tasks closer to the simulation, as it runs.

The first results obtained with Damaris in achieving scalable, jitter-free I/O, were published this year [18]. In order to achieve efficient in-situ visualization at extreme scale, we investigated the limitations of existing in-situ visualization software and proposed to fill the gaps of these software by providing in-situ visualization support to Damaris. The use of Damaris on top of existing visualization packages allows us to:

- Reduce code instrumentation to a minimum in existing simulations,
- Gather the capabilities of several visualization tools to offer adaptability under a unified data management interface,
- Use dedicated cores to hide the run time impact of in-situ visualization and
- Efficiently use memory through a shared-memory-based communication model.

Experiments are now being conducted on BlueWaters (Cray XK6 at NCSA), Intrepid (BlueGene/P at ANL) and Grid5000 with representative visualization scenarios for the CM1 [31] atmospheric simulation and the Nek5000 [34] CFD solver.

Results will be submitted to a conference in early 2013. We plan to further investigate the role that Damaris can take in performing efficient and self-adaptive data analysis in HPC simulations.

6.5.2. Advanced I/O and Storage

Participants: Matthieu Dorier, Alexandru Costan, Gabriel Antoniu.

The recent extension of the JLPC to Argonne National Lab (ANL) has opened new research directions in the field of advanced I/O and storage for HPC, in collaboration with Robert Ross's team at ANL's Mathematics and Computer Science Division (MCS). A founding from the FACCTS program (France And Chicago CollaboraTing in Science) allowed multiple visits (see Section 8.4) of students and researchers from both sides to initiate this new collaboration and explore potential research directions.

One outcome of these visits has been the adaptation of Damaris to work on BlueGene/P and BlueGene/Q machines installed at ANL. Several exchanges led to the design of new I/O scheduling algorithms leveraging Damaris for efficient asynchronous I/O and storage. These algorithms are currently being evaluated, and expected to be published in early 2013.

During these exchanges we also investigated new storage architectures for Exascale systems leveraging BLOB-based large-scale storage able to cope with complex data models. We will explore how we can combine the benefits of the approaches to Big Data storage currently developed by the partners: the BlobSeer approach (KerData), which provides support for multi- versioning and efficient fine-grain access to huge data under heavy concurrency and the Triton approach (ANL), which introduces new object storage semantics. The final goal of the resulting architecture will be to propose efficient solutions to data-related bottlenecks in Exascale HPC systems.

MESCAL Project-Team

6. New Results

6.1. Analysis and Control of Large Stochastic Systems

Perfect sampling is a very efficient technique that uses coupling arguments to provide a sample from the stationary distribution of a Markov chain in a finite time without ever computing the distribution. Even though, the general (non-monotone) case needs to consider the whole state space, we developed a new approach for the general case that only needs to consider two trajectories, an approach which is particularly effective when the state space can be partitioned into pieces where envelopes can be easily computed [8]. Importantly, we also showed that perfect sampling is possible in Jackson networks, even though the underlying Markov chain has a large or even infinite state space and illustrated the efficiency of our approach via numerical simulations [17]. In a similar vein, given that the analysis of a system's dynamics relies on the collection and the description of events, we developed in [37] a new approach to reduce the descriptional complexity of a system by aggregating events' properties, such as their Shannon entropy, entropy gain, divergence etc. These measures were applied to the evaluation of geographic aggregations in the context of news analysis and they allowed us to determine which abstractions one should prefer depending on the task to perform.

In the study of Markov decision processes composed of a large number of objects, we showed that the optimal reward satisfies a Bellman equation, which converges to the solution of a continuous Hamilton-Jacobi-Bellman (HJB) equation based on the mean field approximation of the Markov decision process [10]. We also gave bounds on the difference of the rewards and an algorithm for deriving an approximating solution to the Markov decision process from a solution of the HJB equations. Furthermore, we also studied deterministic limits of Markov processes with discontinuous drifts and showed that under mild assumptions, the stochastic system is a constant-step stochastic approximation algorithm which converges to a differential inclusion obtained by convexifying the rescaled drift of the Markov chain [9].

Finally, in terms of performance evaluation and its applications, we also studied resource-aware business process models by defining a new framework that allows the generation of analytical models. We showed that the analysis of the generated SAN model provides several performance indices we showed that these indices can be easily calculated by a business specialist with no skills in stochastic modeling [7].

6.2. Game Theory and Applications

As far as results in pure game theory are concerned, we studied in [12] a general framework of systems wherein there exists a Pareto optimal allocation that is Pareto superior to an inefficient Nash equilibrium and defined a 'Nash proportionately fair' Pareto optima. In this context, we provided conditions for the existence of a Pareto-optimal allocation that is, truly or most closely, proportional to a Nash equilibrium – an approach with applications in non-cooperative flow-control problems in communication networks.

In a learning context, we also explored what happens beyond the standard first-order framework of continuous time game dynamics and introduced in [42] a class of higher order game dynamics, extending all first order imitative dynamics, and, in particular, the replicator dynamics to higher orders. In stark contrast to the first order case, we showed that weakly dominated strategies become eliminated in all n-th order payoff-monotonic dynamics for all $n > 1$ and strictly dominated strategies become extinct in n-th order dynamics n orders as fast as in first order. Finally, we also established a higher order analogue of the folk theorem of evolutionary game theory which shows that higher order accelerate the rate of convergence to equilibria in games.

In terms of applications, we also examined the distribution of traffic in networks whose users try to minimise their delays by adhering to a simple learning scheme inspired by the replicator dynamics of evolutionary game theory. A major challenge occurs in this context when the users' delays fluctuate unpredictably due to random external factors, but we showed that if users are not too greedy in their learning scheme, then the long-term averages of the users' traffic flows converge to the vicinity of an equilibrium [43].

6.3. Wireless networks

Power and energy considerations in wireless networks have brought to the forefront the need for efficient power allocation and handover policies.

In [13], we analyze the power allocation problem for orthogonal multiple access channels by means of a non-cooperative potential game in which each user distributes his power over the channels available to him. When the channels are static, we show that this game possesses a unique optimum point; moreover, if the network's users follow a distributed learning scheme based on the replicator dynamics of evolutionary game theory, then they converge to this optimum exponentially fast.

On the other hand, in case the network users have access to multiple-antenna technologies (as most smarphone users do nowadays, we also analyze the problem of finding the optimal signal covariance matrix for MIMO multiple access channels by using an approach based on "exponential learning" – a novel optimization method which applies more generally to (quasi-)convex problems defined over sets of positive-definite matrices (with or without trace constraints) [24]. Furthermore, by using a Riemannian-geometric approach, we devise a distributed optimization algorithm which converges to the optimum signal distribution exponentially fast: users attain an ϵ-neighborhood of the system's optimum configuration in time which is at most $\mathcal{O}(\log(1/\epsilon))$ (and, in practice, within only a few iterations, even for large numbers of users) [25].

In the context of heterogeneous wireless networks where vertical handovers are allowed, we also studied a control problem for a new joint admission and resource allocation controller. To account for multi-objective optimization, we considered the maximization of an objective subject to a set of constraints, and we turned this constrained problem into an unconstrained one that we solved numerically using the Semi-Markovian Decision Process (SMDP) framework [19].

6.4. Scheduling

The parallel computing platforms available today are increasingly larger, so it is necessary to develop efficient strategies providing safe and reliable completion for parallel applications. In [6], we proposed a performance model that expresses formally the checkpoint scheduling problem by exhibiting the tradeoff between the impact of the checkpoints operations and the lost computation due to failures. In particular, we proved that the checkpoint scheduling problem is NP-hard even in the simple case of uniform failure distribution and also presented a dynamic programming scheme for determining the optimal checkpointing times in all variants of the problem. On a similar issue, we proposed in [35] a fair scheduling algorithm that handles the problem of fair scheduling by adopting processor fair-share as a strategy for user fairness. Our approach showed that a parallel machine can give a similar type of performance guarantee as a round-robin scheduler, without requiring job preemption been required.

From a network calculus perspective, we presented in [16] a new formalism for data packetization in networks, the "packet curves". Indeed, a more precise knowledge of the packet characteristics can be efficiently exploited to get tighter performance bounds, for example for aggregation of flows, packet-based service policies and shared buffers; finally, we also gave a model for a wormhole switch and showed how our results can be used to get efficient delay bounds.

6.5. Multi-Core Systems

Modern multi-core platforms feature complex topologies with different cache levels and hierarchical memory subsystems, so thread and data placement become crucial to achieve good performance. In [14], we evaluate CPU and memory affinity strategies for numerical scientific multithreaded benchmarks on multi-core platforms and analyzed hardware performance event counters in order to acquire a better understanding of such impact. Likewise, thread mapping is an appealing approach to efficiently exploit the potential of modern chip-multiprocessors, so we proposed in [18] a dynamic thread mapping approach to automatically infer a suitable thread mapping strategy for transactional memory applications composed of multiple execution phases with potentially different transactional behavior in each phase. Our results showed that the proposed dynamic

approach presents performance improvements up to 31% compared to the best static solution. esp From an optimization perspective, the asymmetry in memory access latencies may reduce the overall performance of the system. Therefore, to achieve scalable performance in this environment, we exploited in [28] the machine architecture while taking into account the application communication patterns. Specifically, we introduced a topology-aware asymptotically optimal load balancing algorithm named HwTopoLB which combines the machine topology characteristics with the communication patterns of the application to equalize the application load on the available cores while reducing latencies. We also introduced in [27] a topology-aware load balancer called NucoLB that focuses on redistributing work while reducing communication costs among and within compute nodes, thus leading to performance improvements of up to 20% when compared to state-of-the-art load balancers.

6.6. Cloud Computing

Even though a new era of Cloud Computing has emerged, the characteristics of Cloud load in data centers is not perfectly clear. In [20], we characterized the job/task load and host load in a real-world production data center at Google Inc. by using a detailed trace of over 25 million tasks across over 12,500 hosts. We found that the Google data center exhibits finer resource allocation with respect to CPU and memory than that of Grid/HPC systems and Google jobs are always submitted with much higher frequency and they are much shorter than Grid jobs, leading to higher variance and noise. Moreover, as far as prediction is concerned, we designed in [21] a Bayes model to predict the mean load over a long-term time interval, as well as the mean load in consecutive future time intervals. Real-world experiments showed that our Bayes method achieved high accuracy with a mean squared error of 0.0014 and that it improves the load prediction accuracy by 5.6-50% compared to other state-of-the-art methods based on moving averages, auto-regression, and/or noise filters.

In a similar vein, the exploitation of Best Effort Distributed Computing Infrastructures (BE-DCIs) allows operators to maximize the utilization of the infrastructures, and users to access the unused resources at relatively low cost. Profiling the execution of Bag-of-Tasks (BoT) applications on several kinds of BE-DCIs demonstrates that their task completion rate drops near the end of the execution. In [33], we presented the SpeQuloS service which enhances the QoS of BoT applications executed on BE-DCIs by reducing the execution time, improving its stability, and reporting to users a predicted completion time. We presented the design and development of the framework and several strategies to decide when and how Cloud resources should be provisioned; moreover, performance evaluation using simulations showed that SpeQuloS fulfill its objectives in speeding up the execution of BoTs, in the best cases by a factor greater than 2, while offloading less than 2.5% of the workload to the Cloud. These topics were also further explored in the book chapter [30].

6.7. Experimentation and Visualization in Large Systems

Despite a widespread belief regarding the simulation of large-scale computing systems, we showed in [15] that achieving high scalability does not necessarily require to resort to overly simple models and ignore important phenomena. In fact, by relying on a modular and hierarchical platform representation while taking advantage of regularity when possible, we were able to model systems such as data and computing centers, peer-to-peer networks, grids, or clouds in a scalable way. Finally, in [34], we examined the ability to conduct consistent, controlled, and repeatable large-scale experiments in areas of computer science where availability, repeatability, and open sharing of electronic products are still difficult to achieve.

We also discussed in [22] the concept of the reconstructability of software environments and we proposed a tool for dealing with this problem. In a similar vein, we developed Expo [41], a tool for conducting experiments on distributed platforms. Our experiments confirmed that Expo is a promising tool to help the user with two primary concerns: how to perform a large scale experiment efficiently and easily, together with its reproducilibity.

The exponential number of processes that are executed in high performance applications and the very detailed behavior that we can record about them lead to a trace size explosion both in space and time dimensions. Thus, if the amount of data is not properly treated for visualization, the analysis may give the wrong idea

about the behavior registered in the traces. We dealt with this issue in [38] in two ways: first, by detailing data aggregation techniques that are fully configurable by the user to control the level of details in both space and time dimensions, and second, by presenting two visualization techniques that take advantage of the aggregated data to scale.

Furthermore, given that the performance of parallel and distributed applications is highly dependent on the characteristics of the execution environment, the network topology and characteristics directly impact data locality and movements as well as contention. Unfortunately few visualization available to the analyst are capable of accounting for such phenomena, so we proposed in [39] an interactive topology-based visualization technique based on data aggregation that enables to correlate network characteristics, such as bandwidth and topology, with application performance traces. Such visualization techniques enable us to explore and understand non-trivial behaviors that are impossible to grasp otherwise and the combination of multi-scale aggregation and dynamic graph layout allows us to scale the visualization seamlessly to large distributed systems.

MOAIS Project-Team

6. New Results

6.1. Work Stealing inside GPU

Graphics Processing units (GPU) have become a valuable support for High Performance Computing (HPC) applications. However, despite the many improvements of General Purpose GPUs, the current programming paradigms available, such as NVIDIA?s CUDA, are still low-level and require strong programming effort, especially for irregular applications where dynamic load balancing is a key point to reach high performances. We have introduced a new hybrid programming scheme for general purpose graphics processors using two levels of parallelism. In the upper level, a program creates, in a lazy fashion, tasks to be scheduled on the different Streaming Multiprocessors (MP), as defined in the NVIDIA?s architecture. We have embedded inside GPU a well-known work stealing algorithm to dynamically balance the workload. At lower level, tasks exploit each Streaming Processor (SP) following a data-parallel approach. Preliminary comparisons on data-parallel iteration over vectors show that this approach is competitive on regular workload over the standard CUDA library Thrust, based on a static scheduling. Nevertheless, our approach outperforms Thrust-based scheduling on irregular workloads.

6.2. XKaapi on top of Multi-CPU Multi-GPU

Most recent HPC platforms have heterogeneous nodes composed of a combination of multi-core CPUs and accelerators, like GPUs. Programming such nodes is typically based on a combination of OpenMP and CUDA/OpenCL codes; scheduling relies on a static partitioning and cost model. We have experiment XKaapi runtime system for multi-CPU and multi-GPU architectures, which supports a data-flow task model and a locality-aware work stealing scheduler. The XKaapi enables task multi-implementation on CPU or GPU and multi-level parallelism with different grain sizes. We demonstrate performance results on two dense linear algebra kernels, matrix product (GEMM) and Cholesky factorization (POTRF), to evaluate XKaapi on a heterogeneous architecture composed of two hexa-core CPUs and eight NVIDIA Fermi GPUs. Our conclusion is two-fold: First, fine grained parallelism and online scheduling achieve performance results as good as static strategies, and in most cases outperform them. This is due to an improved work stealing strategy that includes locality information; to a very light implementation of the tasks in XKaapi; and to an optimized search for ready tasks. Next, our XKaapi Cholesky is highly efficient on multi-CPU/multi- GPU due to its multi-level parallelism. Using eight NVIDIA Fermi GPUs and four CPUs, we measure up to 2.43 TFlop/s on double precision matrix product and 1.79 TFlop/s on Cholesky factorization; and respectively 5.09 TFlop/s and 3.92 TFlop/s in single precision. This is the first time that such a performance is obtained with more than four GPUs.

6.3. Formalizing the concept of cooperation

We study how to optimize scheduling problems for a large number of objectives, when multiple users are competing for common resources, with some appropriate notion of fairness between users. Formalizing the concept of cooperation in relation with multi-objective optimization, we can refine the classical methods in combinatorial optimization (that usually optimize one centralized objective) by introducing extra features (adding more objectives or constraints). The PhD thesis of Daniel Cordeiro [2] proposed various ways for handling this problem: multi-organization scheduling and its relaxed variant, impact of selfishness. In the same context, we investigated the field of Game Theory through the existence of Nash equilibria in some situations.

6.4. Fault-tolerance for large parallel systems

This PhD thesis of Slim Bouguerra [1] studied fault-tolerance issues for large parallel systems. We revisited, via a formal proof, the old well-known result which states that the optimal policy for exponential failure law is to put the check-points at periodic moments. We proposed new algorithms to handle check-points for any law in the input and variable check-point costs (JPDC paper).

ROMA Team

5. New Results

5.1. Unified model for assessing checkpointing protocols at extreme-scale

In this work [38], we defined a unified model for several well-known checkpoint/restart protocols. The proposed model is generic enough to encompass both extremes of the checkpoint/restart space, from coordinated approaches to a variety of uncoordinated checkpoint strategies (with message logging). We identified a set of crucial parameters, instantiated them and compared the expected efficiency of the fault tolerant protocols, for a given application/platform pair. We then proposed a detailed analysis of several scenarios, including some of the most powerful currently available HPC platforms, as well as anticipated Exascale designs. The results of this analytical comparison are corroborated by a comprehensive set of simulations. Altogether, they outlined comparative behaviors of checkpoint strategies at very large scale, thereby providing insight that is hardly accessible to direct experimentation.

5.2. Impact of fault prediction on checkpointing strategies

We dealt [34] with the impact of fault prediction techniques on checkpointing strategies. We extended the classical analysis of Young and Daly in the presence of a fault prediction system, which is characterized by its recall and its precision, and which provides either exact or window-based time predictions. We succeeded in deriving the optimal value of the checkpointing period (thereby minimizing the waste of resource usage due to checkpoint overhead) in all scenarios. These results allow to analytically assess the key parameters that impact the performance of fault predictors at very large scale. In addition, the results of this analytical evaluation were nicely corroborated by a comprehensive set of simulations, thereby demonstrating the validity of the model and the accuracy of the results.

5.3. Combining process replication and checkpointing for resilience on exascale systems

Processor failures in post-petascale settings are common occurrences. The traditional fault-tolerance solution, checkpoint-rollback, severely limits parallel efficiency. One solution is to replicate application processes so that a processor failure does not necessarily imply an application failure. Process replication, combined with checkpoint-rollback, has been recently advocated by Ferreira et al. [52]. We first identified [41] an incorrect analogy made in their work between process replication and the birthday problem, and derived correct values for the Mean Number of Failures To Interruption and Mean Time To Interruption for exponentially distributed failures. We then extended these results to arbitrary failure distributions, including closed-form solutions for Weibull distributions. Finally, we evaluated process replication using both synthetic and real-world failure traces. Our main findings are: (i) replication is less beneficial than claimed by Ferreira et al.; (ii) although the choice of the checkpointing period can have a high impact on application execution in the no-replication case, with process replication this choice is no longer critical.

5.4. On the complexity of scheduling checkpoints for computational workflows

This work [22] dealt with the complexity of scheduling computational workflows in the presence of Exponential failures. When such a failure occurs, rollback and recovery is used so that the execution can resume from the last checkpointed state. The goal is to minimize the expected execution time, and we have to decide in which order to execute the tasks, and whether to checkpoint or not after the completion of each given task. We showed that this scheduling problem is strongly NP-complete, and proposed a (polynomial-time) dynamic programming algorithm for the case where the application graph is a linear chain. These results laid the theoretical foundations of the problem, and constituted a prerequisite before discussing scheduling strategies for arbitrary DAGS of moldable tasks subject to general failure distributions.

5.5. Scheduling tree-shaped task graphs to minimize memory and makespan

We [44] investigated the execution of tree-shaped task graphs using multiple processors. Each edge of such a tree represents a large IO file. A task can only be executed if all input and output files fit into memory, and a file can only be removed from memory after it has been consumed. Such trees arise, for instance, in the multifrontal method of sparse matrix factorization. The maximum amount of memory needed depends on the execution order of the tasks. With one processor the objective of the tree traversal is to minimize the required memory. This problem was well studied and optimal polynomial algorithms were proposed. We extended the problem by considering multiple processors, which is of obvious interest in the application area of matrix factorization. With the multiple processors comes the additional objective to minimize the time needed to traverse the tree, i.e., to minimize the makespan. Not surprisingly, this problem proves to be much harder than the sequential one. We studied the computational complexity of this problem and provided an inapproximability result even for unit weight trees. We proposed several heuristics, each with a different optimization focus, and we analyzed them in an extensive experimental evaluation using realistic trees.

5.6. Memory allocation for different classes of DAGs

In this work, we studied the complexity of traversing workflows whose tasks require large I/O files. Such workflows arise in many scientific fields, such as image processing, genomics or geophysical simulations. They usually exhibit some regularity, and most of them can be modeled as Series-Parallel Graph. We target a classical two-level memory system, where the main memory is faster but smaller than the secondary memory. A task in the workflow can be processed if all its predecessors have been processed, and if its input and output files fit in the currently available main memory. The amount of available memory at a given time depends upon the ordering in which the tasks are executed. We focus on the problem of minimizing the amount of main memory needed to process the whole DAG.

We first concentrate on the parallel composition of task chains, or fork-join graphs. We adapt an algorithm designed for trees by Liu [54]. We prove that an optimal schedule for fork-join can be split in two optimal tree schedules, which are obtained using Liu's algorithm. We then move to Series-Parallel graphs and propose a recursive adaptation of the previous algorithm, which consists in serializing every parallel compositions, starting from the innermost, using the fork-join algorithm. Simulations show that this algorithm always reach the optimal performance, and we provide a sketch of the optimality proof. We also study compositions of complete bipartite graphs, which are another important class of DAGs arising in scientific workflows. We propose an optimal algorithm for a class of compositions which we name tower of complete bipartite graphs.

5.7. Scheduling non-linear divisible loads

Divisible Load Theory (DLT) has received a lot of attention in the past decade. A divisible load is a perfect parallel task, that can be split arbitrarily and executed in parallel on a set of possibly heterogeneous resources. The success of DLT is strongly related to the existence of many optimal resource allocation and scheduling algorithms, what strongly differs from general scheduling theory. Moreover, recently, close relationships have been underlined between DLT, that provides a fruitful theoretical framework for scheduling jobs on heterogeneous platforms, and MapReduce, that provides a simple and efficient programming framework to deploy applications on large scale distributed platforms.

The success of both have suggested to extend their framework to non-linear complexity tasks. We show [35] that both DLT and MapReduce are better suited to workloads with linear complexity. In particular, we prove that divisible load theory cannot directly be applied to quadratic workloads, such as it has been proposed recently. We precisely state the limits for classical DLT studies and we review and propose solutions based on a careful preparation of the dataset and clever data partitioning algorithms. In particular, through simulations, we show the possible impact of this approach on the volume of communications generated by MapReduce, in the context of Matrix Multiplication and Outer Product algorithms.

5.8. Energy-aware scheduling under reliability and makespan constraints

We consider [13] a task graph mapped on a set of homogeneous processors. We aim at minimizing the energy consumption while enforcing two constraints: a prescribed bound on the execution time (or makespan), and a reliability threshold. Dynamic voltage and frequency scaling (DVFS) is an approach frequently used to reduce the energy consumption of a schedule, but slowing down the execution of a task to save energy is decreasing the reliability of the execution.

In this work, to improve the reliability of a schedule while reducing the energy consumption, we allow for the re-execution of some tasks. We assess the complexity of the tri-criteria scheduling problem (makespan, reliability, energy) of deciding which task to re-execute, and at which speed each execution of a task should be done, with two different speed models: either processors can have arbitrary speeds (continuous model), or a processor can run at a finite number of different speeds and change its speed during a computation (VDD model). We propose several novel tri-criteria scheduling heuristics under the continuous speed model, and we evaluate them through a set of simulations. The two best heuristics turn out to be very efficient and complementary.

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

We consider [32] the problem of scheduling an application on a parallel computational platform. The application is a particular task graph, either a linear chain of tasks, or a set of independent tasks. The platform is made of identical processors, whose speed can be dynamically modified. It is also subject to failures: if a processor is slowed down to decrease the energy consumption, it has a higher chance to fail. Therefore, the scheduling problem requires to re-execute or replicate tasks (i.e., execute twice a same task, either on the same processor, or on two distinct processors), in order to increase the reliability. It is a tri-criteria problem: the goal is to minimize the energy consumption, while enforcing a bound on the total execution time (the makespan), and a constraint on the reliability of each task.

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

5.10. Optimal algorithms and approximation algorithms for replica placement with distance constraints in tree networks

We study [16] the problem of replica placement in tree networks subject to server capacity and distance constraints. The client requests are known beforehand, while the number and location of the servers are to be determined. The Single policy enforces that all requests of a client are served by a single server in the tree, while in the Multiple policy, the requests of a given client can be processed by multiple servers, thus distributing the processing of requests over the platform. For the Single policy, we prove that all instances of the problem are NP-hard, and we propose approximation algorithms. The problem with the Multiple policy was known to be NP-hard with distance constraints, but we provide a polynomial time optimal algorithm to solve the problem in the particular case of binary trees when no request exceeds the server capacity.

5.11. Throughput optimization for pipeline workflow scheduling with setup times

We tackle [15] pipeline workflow applications that are executed on a distributed platform with setup times. In such applications, several computation stages are interconnected as a linear application graph, and each stage holds a buffer of limited size where intermediate results are stored and a processor setup time occurs when passing from one stage to another. The considered stage/processor mapping strategy is based on interval

mappings, where an interval of consecutive stages is performed by the same processor and the objective is the throughput optimization. Typical examples for this kind of applications are streaming applications such as audio and video coding or decoding, image processing using co-processing devices as FPGA. Even when neglecting setup times, the problem is NP-hard on heterogeneous platforms and we therefore restrict to homogeneous resources. We provide an optimal algorithm for constellations with identical buffer capacities. When buffer sizes are not fixed, we deal with the problem of allocating the buffers in shared memory and present a $b/(b+1)$-approximation algorithm.

5.12. Semi-matching algorithms for scheduling parallel tasks under resource constraints

We study [37] the problem of minimum makespan scheduling when tasks are restricted to subsets of the processors (resource constraints), and require either one or multiple distinct processors to be executed (parallel tasks). This problem is related to the minimum makespan scheduling problem on unrelated machines, as well as to the concurrent job shop problem, and it amounts to finding a semi-matching in bipartite graphs or hypergraphs. While the problem was known to be NP-complete for bipartite graphs, but solvable in polynomial time for unweighted graphs (i.e., unit tasks), we prove that the problem is NP-complete for hypergraphs even in the unweighted case. We design several greedy algorithms of low complexity to solve two versions of the problem, and assess their performance through a set of exhaustive simulations. Even though there is no approximation guarantee on these linear algorithms, they return solutions close to the optimal (or a known lower bound) in average.

5.13. A Symmetry preserving algorithm for matrix scaling

We present an iterative algorithm which asymptotically scales the ∞-norm of each row and each column of a matrix to one. This scaling algorithm preserves symmetry of the original matrix and shows fast linear convergence with an asymptotic rate of $1/2$. We discuss extensions of the algorithm to the one-norm, and by inference to other norms. For the 1-norm case, we show again that convergence is linear, with the rate dependent on the spectrum of the scaled matrix. We demonstrate experimentally that the scaling algorithm improves the conditioning of the matrix and that it helps direct solvers by reducing the need for pivoting. In particular, for symmetric matrices the theoretical and experimental results highlight the potential of the proposed algorithm over existing alternatives. This work resulted in an improved version [43] of an earlier technical report [55].

5.14. On shared-memory parallelization of a sparse matrix scaling algorithm

We discuss [25] efficient shared memory parallelization of sparse matrix computations whose main traits resemble to those of the sparse matrix-vector multiply operation. Such computations are difficult to parallelize because of the relatively small computational granularity characterized by small number of operations per each data access. Our main application is a sparse matrix scaling algorithm which is more memory bound than the sparse matrix vector multiplication operation. We take the application and parallelize it using the standard OpenMP programming principles. Apart from the common race condition avoiding constructs, we do not reorganize the algorithm. Rather, we identify associated performance metrics and describe models to optimize them. By using these models, we implement parallel matrix scaling algorithms for two well-known sparse matrix storage formats. Experimental results show that simple parallelization attempts which leave data/work partitioning to the runtime scheduler can suffer from the overhead of avoiding race conditions especially when the number of threads increases. The proposed algorithms perform better than these algorithms by optimizing the identified performance metrics and reducing the overhead.

5.15. Investigations on push-relabel based algorithms for the maximum transversal problem

In a technical report [42], we investigate the push-relabel algorithm for solving the problem of finding a maximum cardinality matching in a bipartite graph in the context of the maximum transversal problem. We describe in detail an optimized yet easy-to-implement version of the algorithm and fine-tune its parameters. We also introduce new performance-enhancing techniques. On a wide range of real-world instances, we compare the push-relabel algorithm with state-of-the-art augmenting path-based algorithms and the recently proposed pseudoflow approach. We conclude that a carefully tuned push-relabel algorithm is competitive with all known augmenting path-based algorithms, and superior to the pseudoflow-based ones. We finalized this work by reporting the most important results in a journal article [9].

5.16. On optimal and balanced sparse matrix partitioning problems

We investigate [20] one dimensional partitioning of sparse matrices under a given ordering of the rows/columns. The partitioning constraint is to have load balance across processors when different parts are assigned to different processors. The load is defined as the number of rows, or columns, or the nonzeros assigned to a processor. The partitioning objective is to optimize different functions, including the well-known total communication volume arising in a distributed memory implementation of parallel sparse matrix-vector multiplication operations. The difference between our problem in this work and the general sparse matrix partitioning problem is that the parts should correspond to disjoint intervals of the given order. Whereas the partitioning problem without the interval constraint corresponds to the NP-complete hypergraph partitioning problem, the restricted problem corresponds to a polynomial-time solvable variant of the hypergraph partitioning problem. We adapt an existing dynamic programming algorithm designed for graphs to solve two related partitioning problems in graphs. We then propose graph models for a given hypergraph and a partitioning objective function so that the standard cutsize definition in the graph model exactly corresponds to the hypergraph partitioning objective function. In extensive experiments, we show that our proposed algorithm is helpful in practice. It even demonstrates performance superior to the standard hypergraph partitioners when the number of parts is high.

5.17. 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 [50], [51]. The construction algorithm has a worst-case time complexity of $\Theta(mn)$ for an $n \times n$ unsymmetric matrix having m off-diagonal nonzeros. We proposed [53] another algorithm that has a worst-case time complexity of $\mathcal{O}(m \log n)$. During this reporting period, we compared the two algorithms experimentally and showed that both algorithms are efficient in general. The known algorithm [51] is faster in many practical cases, yet there are instances in which there is a significant difference between the running time of the two algorithms in favor of the proposed one.

5.18. Introduction of shared memory parallelism in a distributed-memory sparse multifrontal solver

We study the adaptation of a parallel distributed-memory solver, MUMPS, into a shared-memory code, targetting multicore architectures. An advantage of adapting the code rather than starting with a new design is to fully benefit from its numerical kernels and functionalities. We show how one can take advantage of OpenMP directives and of existing libraries optimized for shared-memory environments, in our case BLAS libraries [48]. We have also started to study approaches that take advantage of the specificities of NUMA architectures.

5.19. Improving multifrontal methods by means of low-Rank representations

Matrices coming from elliptic PDEs have been shown to have a low-rank property. Although the dense internal datastructures involved in a multifrontal method, the so-called frontal matrices or fronts, are full-rank, their off-diagonal blocks can then be approximated by low-rank products. We have studied a low-rank format called Block Low Rank and explained how it can be used to reduce the memory footprint and complexity of both the factorization and solve phases, depending on the way variables are grouped. The proposed approach can be used either to accelerate the factorization and solution phases or to build a preconditioner [47]. We have started the development of a version of MUMPS that exploits such properties. This work is in collaboration with EDF (contract funding for the Ph.D. thesis of C. Weisbecker at INPT) and C. Ashcraft (LSTC).

5.20. Parallel computation of inverse entries of a sparse matrix

We have worked on the parallel computation of several entries [31] of the inverse of a large sparse matrix. We assume that the matrix has already been factorized by a direct method and that the factors are distributed. Entries are efficiently computed by exploiting sparsity of the right-hand sides and the solution vectors in the triangular solution phase. We demonstrate that in this setting, parallelism and computational efficiency are two contrasting objectives. We develop an efficient approach and show its efficacy by runs using the MUMPS code that implements a parallel multifrontal method.

5.21. Robust memory-aware mappings for parallel multifrontal factorization

We have studied the memory scalability of the parallel multifrontal factorization of sparse matrices. In particular, we are interested in controlling the active memory specific to the multifrontal factorization. We illustrate why commonly used mapping strategies (e.g. proportional mapping) cannot achieve a high memory efficiency. We propose a class of "memory-aware" algorithms that aim at maximizing performance under given memory constraints, and explain why they provide reliable memory estimates, thus a more robust solver. We study these issues in the context of the MUMPS solver, in which new experimental static scheduling strategies have been implemented and experimented on large matrices [46].

RUNTIME Project-Team

6. New Results

6.1. Mastering Heterogeneous Platforms

Participants: Cedric Augonnet, Olivier Aumage, Nicolas Collin, Ludovic Courtès, Nathalie Furmento, Sylvain Henry, Andra Hugo, Raymond Namyst, Cyril Roelandt, Corentin Rossignon, Ludovic Stordeur, Samuel Thibault, Pierre-André Wacrenier.

- We continued our work on extending StarPU to master exploitation of Heterogeneous Platforms.
- We have released version 1.0.0 of StarPU, now really considered a stable project that a lot of collaborators can base their work on.
- We have extended our lightweight DSM over MPI to support caching data [17], which dramatically reduces data transfers for classical applications.
- We have extended the StarPU scheduler to let the application provide several implementations of a function for the same architecture, implementation choice being performed by the scheduler according to actually measured performance, energy consumption, etc.
- We have collaborated with Computer Graphics research team in the MediaGPU project to make it possible to directly graphically render results from StarPU computations.
- Work has been initiated to integrate StarPU and SimGrid for the SONGS project, which will allow to simulate application execution on heterogeneous architectures, and thus easily experiment with scheduling strategies.
- We have extended StarPU with a protocol that permits to make it run with a master-slave model, which allowed to easily port it to the Intel SCC and Intel Xeon Phi processors, and will allow an easy load balancing support over MPI.
- We have extended StarPU to allow multiple parallel codes to run concurrently with minimal interference. Such parallel codes run within *scheduling contexts* that provide confined execution environments which can be used to partition computing resources. Scheduling contexts can be dynamically resized to optimize the allocation of computing resources among concurrently running libraries. We introduced a *hypervisor* that automatically expands or shrinks contexts using feedback from the runtime system (e.g. resource utilization).

  We demonstrated the relevance of our approach using benchmarks invoking multiple high performance linear algebra kernels simultaneously on top of heterogeneous multicore machines. We showed that our mechanism can dramatically improve the overall application run time (-34%), most notably by reducing the average cache miss ratio (-50%).
- We have improved [15] the OpenCL implementation on top of StarPU (SOCL) to allow applications to use StarPU's scheduling contexts through OpenCL's contexts and to explicitly schedule some kernels to enhance performance. Moreover, SOCL fully supports the OpenCL ICD extension and can now be dynamically selected amongst other available platforms which makes it easier to use.
- We have continued collaborations on applications on top of StarPU with the University of Mons [14], the University of Vienna [20], the University of Linköping, the University of Tsukuba, TOTAL, the CEA INAC in Grenoble and the BRGM French public institution in Earth science applications.
- In a joint work with French SME company CAPS entreprise, as part of the ANR ProHMPT project, we have demonstrated a proof of concept framework enabling three kinds of pieces of applicative code — a native StarPU code, a Magma/StarPU code and a HMPP/StarPU code annotated with HMPP's directives — to integrate and cooperate together on a computation as a single coherent application.

- As part of the HPC-GA project, we initiated a preliminary study with University of Rio Grande do Sul (UFRGS), Brazil, to cooperate on the modeling of common computing kernel tasks and potentially making use of kernel models designed at UFRGS within the StarPU's task cost evaluation framework.
- As part of the partnership with Total, and in relationship with StarPU's task scheduling work, we have explored solutions to semi-automatically adapt the grain of elementary tasks to the available computing resources.

6.2. High-Performance Intra-node Collective Operations

Participant: Brice Goglin.

- KNEM is known to improve the performance of point-to-point intra-node MPI communication significantly [13].
- 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.
- 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.

6.3. Process Placement and Topology-Aware Computing

Participants: Emmanuel Jeannot, Guillaume Mercier, François Tessier.

- TREEMATCH's limitations have been addressed. In particular, it is now able to handle unbalanced physical topologies.
- TREEMATCH has been compared to various competitors. We carried out various experiments that showed that TREEMATCH outperforms other solutions based on graph partitioning or graph embedding. These experiments also showed the limitations of some existing solutions (Scotch for instance).
- TREEMATCH has been integrated into several major parallel programming environments. It is implemented as a load-balancer in Charm++ (François TESSIER made severals at UI Urbana Champaign) and is used to enhance topology management routines in Open MPI and MPICH2. It is indeed employed to allow rank reordering in functions such as `MPI_Dist_graph_create` for instance. This work started with a visit at UTK by Guillaume MERCIER.
- We set-up several collaborations: besides the collaboration with the Open MPI group, we also work with the CERFACS in order to speed-up existing CFD parallel applications developed by this group.

6.4. Thread placement and memory allocation on NUMA machines

Participant: Emmanuel Jeannot.

We have worked on optimizing the tiled Cholesky factorization on NUMA machine. We have designed a new symbolic technnique for allocating task and tiles at the same time called SMA (Symbolic Mapping and Allocation). SMA provide an optimal allocation in terms of point-to-point communication for the Cholesky factorization. We have studied some performance issues regarding the way threads are grouped and tiles are allocated in the memory. We have shown how to optimize thread placement and data placement in order to achieve performance gain up to 50% compared to state-of-the-art libraries such as Plasma or MKL. This work has been published in PAAP 2012 [25].

6.5. Scheduling for System On Chip

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 and HEVC) 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 STHORM and aimed at meeting the latest codecs' requirements.

This year, we focused on the memory constraints embedded systems are subject to. As video coding is rather demanding in terms of storage capacity, we have proposed a method aimed at introducing the notion of memory into a class of widespread scheduling heuristics that exhibit both good performance and low complexity. Thanks to this technique, we achieved speedups over 20%.

The next step is to formalize an execution model on top of which a runtime software will be built. This implies specifying both the application requirements and modeling precisely the target platform, namely STHORM.

6.6. High-Performance Point-to-Point Communications

Participants: Alexandre Denis, Sébastien Barascou, 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 a communication protocol [21] for InfiniBand that amortizes the cost of checksums as used by fault-tolerant MPI implementations. We have modeled the behavior of the network and proposed auto-tuning mechanisms to adapt the protocol to the hardware properties.
- This work was initiated in the context of the FP3C collaboration with the University of Tokyo.

DANTE Team

6. New Results

6.1. Use of wireless sensor network for Assessing Interactions between Healthcare Workers and Patients under Airborne Precautions

Direct observation has been widely used to assess interactions between healthcare workers (HCWs) and patients but is time-consuming and feasible only over short periods. We used a Radio Frequency Identification Device (RFID) system to automatically measure HCW-patient interactions [14]. The RFID was well accepted by HCWs. This original technique holds promise for accurately and continuously measuring interactions between HCWs and patients, as a less resource-consuming substitute for direct observation. The results could be used to model the transmission of significant pathogens. HCW perceptions of interactions with patients accurately reflected reality.

6.2. Psychological Aspects of Social Communities

Social Network Analysis has often focused on the structure of the network without taking into account the characteristics of the individual involved. In this work [28], [8], we aim at identifying how individual differences in psychological traits affect the community structure of social networks. Instead of choosing to study only either structural or psychological properties of an individual, our aim is to exhibit in which way the psychological attributes of interacting individuals impacts the social network topology. Using psychological data from the myPersonality application and social data from Facebook, we confront the personality traits of the subjects to metrics obtained after applying the C3 community detection algorithm [41] to the social neighborhood of the subjects. We observe that introverts tend to have less communities and hide into large communities, whereas extroverts tend to act as bridges between more communities, which are on average smaller and of varying cohesion.

6.3. Community detection: dynamic, overlapping, fuzzy

Community, a notion transversal to all areas of Social Network Analysis, has drawn tremendous amount of attention across the sciences in the past decades. Numerous attempts to characterize both the sociological embodiment of the concept as well as its observable structural manifestation in the social network have to this date only converged in spirit. No formal consensus has been reached on the quantiffable aspects of community, despite it being deeply linked to topological and dynamic aspects of the underlying social network.

The DANTE team proceeded results on several aspects of community detection is large scale networks.

- Presenting a fresh approach to the evaluation of communities, we introduces and builds upon the cohesion [8], a novel metric which captures the intrinsic quality, as a community, of a set of nodes in a network. The cohesion, defined in terms of social triads, was found to be highly correlated to the subjective perception of communitiness through the use of a large-scale online experiment in which users were able to compute and rate the quality of their social groups on Facebook. The use of the cohesion proves invaluable in that it offers non- trivial insights on the network structure and its relation to the associated semantic. The use of the cohesion was use for example in order to study Agreement Groups in the United States Senate [35].
- Overlapping community detection is a popular topic in complex networks. As compared to disjoint community structure, overlapping community structure is more suitable to describe networks at a macroscopic level. Overlaps shared by communities play an important role in combining different communities. In this paper, two methods are proposed to detect overlapping community structure. One is called clique optimization, and the other is named fuzzy detection. Clique optimization aims at detecting granular overlaps. The clique optimization method is a fine grain scale approach. Each

granular overlap is a node connected to distinct communities and it is highly connected to each community. Fuzzy detection is at a coarser grain scale and aims at identifying modular overlaps. Modular overlaps represent groups of nodes that have high community membership degrees with several communities. A modular overlap is itself a possible cluster/sub-community [7], [38].

6.4. Structure of Changes in Dynamic Contact Networks

We present a methodology to investigate the structure of dynamic networks in terms of concentration of changes in the network. We handle dynamic networks as series of graphs on a set of nodes and consider the changes occurring between two consecutive graphs in the series. We apply our methodology to various dynamic contact networks coming from different contexts and we show that changes in these networks exhibit a non-trivial structure: they are not spread all over the network but are instead concentrated around a small fraction of nodes. We compare our observations on real-world networks to three classical dynamic network models and show that they do not capture this key property [31].

6.5. Dynamic Resource Management in Clouds: A Probabilistic Approach

Dynamic resource management has become an active area of research in the Cloud Computing paradigm. Cost of resources varies significantly depending on configuration for using them. Hence efficient management of resources is of prime interest to both Cloud Providers and Cloud Users. In this work we suggest a probabilistic resource provisioning approach that can be exploited as the input of a dynamic resource management scheme. Using a Video on Demand use case to justify our claims, we propose an analytical model inspired from standard models developed for epidemiology spreading, to represent sudden and intense workload variations. We show that the resulting model verifies a Large Deviation Principle that statistically characterises extreme rare events, such as the ones produced by ?buzz/flash crowd effects? that may cause workload overflow in the VoD context. This analysis provides valuable insight on expectable abnormal behaviours of systems. We exploit the information obtained using the Large Deviation Principle for the proposed Video on Demand use-case for defining policies (Service Level Agreements). We believe these policies for elastic resource provisioning and usage may be of some interest to all stakeholders in the emerging context of cloud networking [4], [24].

6.6. Classification of Content and Users in BitTorrent by Semi-supervised Learning Methods

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 the present work is to provide a framework for sub-classification of P2P traffic generated by the BitTorrent protocol. The general intuition is that the users with similar interests download similar contents. This intuition can be rigorously formalised with the help of graph based semi- supervised learning approach. We have chosen to work with PageRank based semi-supervised learning method, which scales well with very large volumes of data. We provide recommendations for the choice of parameters in the PageRank based semi-supervised learning method. In particular, we show that it is advantageous to choose labelled points with large PageRank score.

This work was awarded best paper at the 3rd International Workshop on Traffic Analysis and Classification (in conjunction with the 8th International Wireless Communications and Mobile Computing Conference, 2012) [21] and led to a companion paper [22].

6.7. Large deviations estimates for the multiscale analysis of heart rate variability

In the realm of multiscale signal analysis, multifractal analysis provides with a natural and rich framework to measure the roughness of a time series. As such, it has drawn special attention of both mathematicians and practitioners, and led them to characterize relevant physiological factors impacting the heart rate variability. Notwithstanding these considerable progresses, multi- fractal analysis almost exclusively developed around the concept of Legendre singularity spectrum, for which efficient and elaborate estimators exist, but which are structurally blind to subtle features like non-concavity or, to a certain extent, non scaling of the distributions. Large deviations theory al- lows bypassing these limitations but it is only very recently that performing estimators were proposed to reliably compute the corresponding large devia- tions singularity spectrum. In this article, we illustrate the relevance of this approach, on both theoretical objects and on human heart rate signals from the Physionet public database. As conjectured, we verify that large devia- tions principles reveal significant information that otherwise remains hidden with classical approaches, and which can be reminiscent of some physiolog- ical characteristics. In particular we quantify the presence/absence of scale invariance of RR signals.

These results gather most achievements we carried out within the ANR project DMASC.

6.8. An Inexpensive Packet Capture Solution with Robust and Accurate Timestamping

The availability of inexpensive and reliable packet capture solutions is highly desirable for the management of future Internet infrastructures and practices. Currently, available solutions are either 1) based on GPS antennas and dedicated hardware and thus are expensive and difficult to deploy, or 2) based on commodity hardware and standard synchronization protocols and thus have inaccurate timestamps and cannot handle monitoring at high rate. In a series of ongoing works in collaboration with the Melbourne University (Australia), we proposed an architecture for a packet monitoring solution which combines inexpensive network cards capable of hardware timestamping, with RAD- clock, an open source software clock. In different papers, we presented the first implementation and evaluation of our approach, demonstrating a good compromise between affordability and accuracy [33], [36].

6.9. KBAC: Knowledge-Based Admission Control

Many methods have been proposed in the literature to perform admission control in order to provide a sufficient level of Quality of Service (QoS) to accepted flows. In this work, we introduce a novel data-driven method based on a timevarying model that we refer to as Knowledge-Based Admission Control solution (KBAC). Our KBAC solution consists of three main stages: (i) collect measurements on the on-going traffic over the communication link; (ii) maintain an up-to-date broad view of the link behavior, and feed it to a Knowledge Plane; (iii) model the observed link behavior by a mono-server queue whose parameters are set automatically and which predicts the expected QoS if a flow requesting admission were to be accepted. Our KBAC solution provides a probabilistic guarantee whose admission threshold is either expressed, as a bounded delay or as a bounded loss rate. We run extensive simulations to assess the behavior of our KBAC solution in the case of a delay threshold. The results show that our KBAC solution leads to a good trade-off between flow performance and resource utilization. This ability stems from the quick and automatic adjustment of its admission policy according to the actual variations on the traffic conditions [19].

6.10. Substitution Networks: Performance Collapse due to Overhead in Communication Times

A substitution network is a wireless solution whose purpose is to bring back connectivity or to provide additional bandwidth capacity to a network that just suffered a failure or a dramatic surge in its workload. We analyze the performance of the simplest possible multihop topology for a substitution network, i.e., the

multihop chain subject to traffic transmitted in both directions. Clearly, the potential capacity of a substitution network, whose technology should be embedded in mobile routers, is very likely to be far much smaller than the prior base network. We investigate the actual performance attained by such a substitution network under various conditions of the chain length and the carrier sensing range. Our results show that the capacity, viz. its maximum attainable throughput, reaches a peak at a given workload and then, for larger values of workload, decreases towards an asymptote which value can be drastically lower than the peak value. We give insights into this performance collapse and show the need for a suitable admission control [18].

6.11. Characterisation and Application of Idle Period Durations in IEEE 802.11 DCF-based Multihop Wireless Networks

Multihop wireless networks are used to provide internet connectivity to the users and the level of performance and quality expected by these users are increasing. In order to meet these performance and quality requirements, wireless communications should be enhanced. Previous works from the literature show that the performance and quality provided by an IEEE 802.11-based multihop wireless network are far from optimal and that there exist different ways to increase the efficiency and the quality of service of such a network. Some studies show that using the medium state as a parameter to tune the behaviour of an IEEE 802.11-based multihop network is an appropriate way to proceed. A station in a IEEE 802.11-based multihop wireless network senses the medium either busy or idle. The durations of idle periods and busy periods and their distributions have a clear impact on the network and nodes performance. The understanding of the relationship between these indicators, namely idle and busy periods, the network topology and the traffic, would give new insights to enhance the performance and quality of multihop wireless networks. Due to its multihop and distributed nature, the characterisation of idle period durations is difficult in such a network. This work explores the characterisation of idle period distribution by proposing a new analytical model and provides an application of this characterisation with the design of an adaptive backoff algorithm based on idle periods [30].

DIONYSOS Project-Team

5. New Results

5.1. Quality of Experience

Participants: Gerardo Rubino, Adlen Ksentini, Yassine Hadjadj-Aoul, Sofiene Jelassi, Sebastián Basterrech.

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 efficiently, allowing if necessary its use 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 [96] 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) [93] 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) [95] and *Echo State Networks* (ESN) [93] and constitute today one of the basic paradigms for Recurrent Neural Networks modeling [94]. 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 [38] we propose a new learning tool which merges the capabilities of Random Neural Networks (RNNs) with those of Reservoir Computing Models (RCMs). We keep some of the nice features of RNNs with the ability of RCMs in predicting time series values. Our tool is called Echo State Queueing Network. In the paper, we illustrate its performances in predicting, in particular, Internet traffic. We also worked on the bottleneck of the PSQA building process, from the time consuming point of view, the subjective test sessions. We proposed in [49] and [48] new PSQA modules for VoIP and SVC video, respectively. In [49], we used PESQ for replacing the subjective test in the training step of PSQA. This module is dedicated to iLBC and Speex codecs. Whereas in [48], we used VQM tool to evaluate the SVC video sequences to train PSQA.

In [31], a general presentation of our approach in Dionysos was given, together with some guidelines in looking for extensions able to deal with the evaluation of generic applications or services over the Internet.

We presented a tutorial on Quality of Experience in Qest'2012 [69], based on our past research results in evaluating the perceptual quality in voice or video applications, and on the current work performed in the QuEEN project.

Our perceptual quality work is being extended to investigate the quality of user experience including a large scope that involves human and technology factors. This work is conducted in the context of the Celtic-QuEEN project where a complete QoE monitoring platform is being designed. In Qest'2012 [69], we presented a tutorial on Quality of Experience based on our past research results in evaluating the perceptual quality in voice or video applications, and on the current work performed in QuEEN.

On the other hand, we continue our study of quality of temporally interrupted VoIP service frequently observed over wireless and data networks. A flagship paper regarding the perception of interruptions in the context of VoIP service is published in [53]. In [21] we presented a detailed state-or-the-art in the area.

5.2. Network Economics

Participants: 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* (see for instance [83]). 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 equilibria due competition among providers in different contexts, iii) looking at the economics of applications, for example adword auctions for search engines, iv) studying the network neutrality issue, and v) the not so considered problem of search-neutrality.

On the first item, we have studied in [78] 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 follows 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 ending this year. 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 [77]. A general model of competition in loss networks is described and analyzed in [25] as a two-levels game: at the smallest time scale, users' demand is split among providers according to Wardrop principle, depending on the access price and available QoS (depending itself on the level of demand at the provider); 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. In [59], we present a similar model of competition on prices between two telecommunication service providers sharing an access resource, which can for example be a single WiFi spectrum. We again obtain a two-level game corresponding to two time scales of decisions: at the smallest time scale, users play an association game by choosing their provider (or none) depending on price, provider reputation and congestion level; at the largest time scale, providers compete on prices. We show that the association game always has an equilibrium, but that several equilibria can exist. The pricing game is then solved by assuming that providers are risk-averse and try to maximize the minimal revenue they can get at a user equilibrium. We illustrate what can be the outcome of this game and that there are situations for which providers can co-exist.

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. The strategies of the search engine and advertisers are described and analyzed in [24].

A new issue on which most of our work has focused in 2012 is related to the *network neutrality debate*. This debates comes from the increasing traffic asymmetry between Internet Service Providers (ISPs), mainly due to some prominent and resource consuming content providers (Cps) which are usually connected to a single ISP. Thus the ISPs to whom those CPs are not directly connected have started to wonder why distant CPs should not be charged by them, with the threat of their traffic not being delivered if they do not accept to pay, or their quality of service decreased. In [79], we have described and analyzed the respective arguments of neutrality

proponents and opponents, and we have also participated to Inria's response to the ARCEP consultation on the topic [90]. We have reviewed in [50], [85] the economic transit agreements between ISPs in order to determine their best strategy. We have defined a model with two ISPs, each providing direct connectivity to a fixed proportion of the content and competing in terms of price for end users, who select their ISP based on the price per unit of available content. We have analyzed and compared, thanks to game-theoretic tools, three different situations: the case of peering between the ISPs, the case where ISPs do not share their traffic (exclusivity arrangements), and the case where they fix a transfer price per unit of volume. The impact on the network neutrality debate is then discussed. An analysis with a hierarchy of providers, with separated backbone providers and access providers, is performed in [89]. We also remarked that while there have been many studies discussing the advantages and drawbacks of neutrality, there is no game-theoretical work dealing with the observable situation of competitive ISPs in front of a (quasi-)monopolistic CP. Though, this is a typical situation that is condemned by ISPs and, according to them, another reason of the non-neutrality need. We have developed and analyzed in [40], [84] two different models describing the relations between two competitive ISPs and a single CP, played as a three-level game corresponding to three different time scales. At the largest time scale, side payments (if any) are determined. At a smaller time scale, ISPs decide their (flat-rate) subscription fee (toward users), then the CP chooses the (flat-rate) price to charge users. Users finally select their ISP (if any) using a price-based discrete choice model in [84] or following Wardrop principle in [40], and decide whether to also subscribe to the CP service. The game is analyzed by backward induction. As a conclusion, we obtain among other things that non-neutrality may be beneficial to the CP, and not necessarily to ISPs, unless the side payments are decided by ISPs (through a non-cooperative game). Another specific scenario is studied in [51], where the impact of wholesale prices is examined in a context where the end customer access both free content and pay-per-use content, delivered by two different providers through a common network provider. We formulate and solve the game between the network provider and the pay-per-use content provider, where both use the price they separately charge the end customer with as a leverage to maximize their profits. In the neutral case (the network provider charges equal wholesale prices to the two content providers), the benefits coming from wholesale price reductions are largely retained by the pay-per-use content provider. When the free content provider is charged more than its pay-per-use competitor, both the network provider and the pay-per-use content provider see their profit increase, while the end customer experiences a negligible reduction in the retail price.

If network neutrality has recently attracted a lot of attention, *search neutrality* is also becoming a vivid subject of discussion because a non-neutral search may prevent some relevant content from being accessed by users. We propose in [88] to model two situations of a non-neutral search engine behavior, which can rank the link propositions according to the profit a search can generate for it, instead of just relevance: the case when the search engine owns some content, and the case when it imposes a tax on organic links, a bit similarly to what it does for commercial links. We analyze the particular (and deterministic) situation of a single keyword, and describe the problem for the whole potential set of keywords. In [52], we analyze one behavior that results in search bias: the payment by content providers to the search engine in order to improve the chances to be located (and accessed) by a search engine user. A simple game theory-based model is presented, where both a search engine and a content provider interact strategically, and the aggregated behavior of users is modeled by a demand function. The output of each stakeholder when the search engine is engaged in such a non-neutral behavior is compared with the neutral case when no such side payment is present.

5.3. Wireless Networks

Participants: Adlen Ksentini, Yassine Hadjadj-Aoul, Bruno Sericola.

Long Term Evolution (LTE) represents the next generation of Cellular networks or 4G. It allows increasing the data rate and hence services that can be proposed to users. A notable part of activity in cellular networks and particularly in LTE, is related to increasing the user QoE. Due to their numerous advantages, current trends show a growing number of femtocell deployments. However, femtocells would become less attractive to the general consumers if they cannot keep up with the service quality that the macro cellular network should provide. Given the fact that the quality of mobile services provided at femtocells depends largely on the level

of congestion on the backhaul link, in [71] we introduced a flow mobility/handover admission control method that makes decisions on layer-three handovers from macro network to femtocell network and/or on entire or partial flow mobility between the two networks based on predicted QoS taking into account metrics such as network load/congestion indications and based on predicted QoE metrics. In [70], we proposed a complete framework that anticipates QoS/QoE (Quality of Experience) degradation and proactively defines policies for LTE-connected cars (UEs) to select the most adequate radio access out of WiFi and LTE. For a particular application, the proposed framework considers the application type, the mobility feature (e.g., speed, user mobility entire/partial path, user final/intermediate destination), and the traffic dynamics over the backhauls of both LTE and WiFi networks in order to predict and allow the UE to select the best network that maximizes user QoE throughout the mobility path.

In [33],[23] we considered LTE networks as candidates for hosting the Machine to Machine communication (or Machine Type Communication in the 3GPP jargon). One of the most important problems posed by this kind of traffic is congestion. Congestion concerns all the parts of the network, both the radio and the core networks impacting both the user data and the control planes. In these works, we proposed a congestion aware admission control solution that selectively rejects signaling messages from MTC devices at the radio access network following a probability that is set based on a proportional integrative derivative (PID) controller (from control theory) reflecting the congestion level of a relevant core network node.

Another part of our activities in wireless network are related to energy saving. Indeed, one of the biggest problem today in the wireless world is that wireless devices are battery driven, which reduce their operating lifetime. We addressed the energy issue in wireless network for two different contexts: (i) rich media (such as VoIP) delivery in Wireless LAN; (ii) Wireless Sensor Network (WSN).

In WLAN, mobile stations conserve energy by maximizing the sleep mode periods of the wireless interfaces. Despite of its efficiency, this mode is incompatible with real-time applications and media streaming, like VoIP. In fact, maximizing the sleep mode periods is directly translated into an increased delay, which induces packets losses when exceeding certain thresholds (e.g. buffer overflow and late packet loss), and may severely degrade the perceived user's QoE. We first review a clear state of the art on energy saving for mobiles communication [22]. Then, in [56], we showed the relation between user QoE and the sleep period in the context of Voice over Wireless Lan (VoWLAN). The system was modeled and controlled using a PID controller, which computes the sleep period enabling to reach a QoE reference value. Thus, we achieved the trade-off between energy consumption and QoE.

On the other hand, Wireless Sensor Networks (WSN) protocols focus primarily on power conservation, because of the limited capacity of the sensor nodes' batteries. In [64] we addressed the case of using radio diversity in WSN (more than one antenna). In this work, we proposed a scheme for radio diversity that can balance, depending on the traffic nature in the network, between minimizing the energy consumption or minimizing the end-to-end delay. The proposed scheme combines the benefit of two metrics, which aim separately to minimize the energy consumption, and to minimize delay when delivering packets to the end-user. In [57], we worked on the localization problem in WSN by introducing a new way to determine the sensors' residence area. Our new localization algorithm is based on the geometric shape of half-symmetric lens. In [81] we developed a performance analysis of a compression scheme designed to save energy, for specific types of WSN.

In [55], we presented the DVB-T2 simulation module for OPNET. Note that this module is the only available implementation of DVB-T2 in network simulators.

5.4. Information-Centric Networks

Participants: Yassine Hadjadj-Aoul, Gerardo Rubino, Leila Ghazzai.

The rise of popularity of video streaming services has resulted in increased volumes of network traffic, which in turn have created Internet bottlenecks leading to perceived quality degradations. One of the recognized good ways to tackle this type of congestion is to make the contents available inside ISPs' networks. We thus proposed, in [73] a network-friendly content delivery architecture that considers the complex video distribution

chain and its associated business models. This comprehensive architecture allows a network operator to fully engineer video traffic distribution in order to both alleviate peering links' workload and improve delivered QoS. This proposal is fully compatible with Adaptive Bitrate Streaming (ABS) architectures, which are currently used to distribute video in the Internet.

The Content providers are increasingly becoming interested in evaluating the performance of such streaming protocol from the final users' perspective. Indeed, more importance is being attached to the quality as perceived by the final users, or Quality of Experience (QoE), as compared to just Quality of Service. Thus, we addressed in [68] the problem of estimating the QoE of video streaming in TCP/IP networks. As a solution, we designed an automatic no-reference QoE estimation module for HTTP video streaming using TCP and H.264 video codec. The proposed approach is different from the existing ones as it addresses the problem of measuring QoE in the combined case of adaptive video bitrates and the use of a reliable transport protocol. This is the case of the adaptive streaming over HTTP.

On the other hand, as introduced by ICN's content caching mainly addresses the management of the content in a particular cache, while the content replication consists in disseminating data in its way to the destination. The benefits of contents' replication can be completely cancelled with a bad caching technique. Thus, we proposed, in [75], to analyse the interaction existing between caching strategies and content replication.

5.5. Interoperability assessment and Internet of Things

Participants: César Viho, Nanxing Chen, Anthony Baire.

The Internet of Things (IoT) brings new challenges to interoperability assessment by introducing the necessity to deal with non reliable environments connecting plenty billions of objects widely distributed. In this context, the IETF Constrained Application Protocol (CoAP) has been designed, which is an application-layer protocol on keeping in mind the various issues of constrained environment to realize interoperations with constrained networks and nodes.

As one of the most important protocol for the future Internet of Things, the number of smart objects using CoAP is expected to grow substantially. For CoAP applications to be widely adopted by the industry, interoperability testing is required to ensure that CoAP implementations from different vendors work well together. Therefore, in the recent period, we propose an interoperability testing methodology using a *passive* approach. Contrary to the classical testing method used in conventional interoperability testing events, which is done by actively stimulating the implementations and verifying the outputs, we apply passive testing. It is a technique based only on observation [47]. Its non-intrusive nature makes it appropriate for interoperability testing, especially in the context of IoT. We have also developed a tool that implement this passive method that has been used successfully to test CoAP implementations during the two CoAP Plugtest interoperability sessions organized by ETSI and IPSO Alliance [44]. Our contributions and originality of this work published in [46] are three-fold: *(i)* A new testing method using a passive approach. *(ii)* As IoT implies providing services in lossy networks, we also take into account fundamental CoAP implementations interoperability testing in lossy context. *(iii)* Contrary to manual verification used in conventional interoperability testing events, the verification procedure has been automatized by a test validation tool, which increases the test efficiency while reducing testing time and costs.

5.6. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Gerardo Rubino, Laura Aspirot, Romaric Ludinard.

In [92] and [13], 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.

In [14], we present an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, and in particular of peer-to-peer systems. When designing such systems, two major issues need to be faced. First, population of these systems evolves continuously (nodes can join and leave the system as often as they wish without any central authority in charge of their control), and second, these systems being open, one needs to defend against the presence of malicious nodes that try to subvert the system. Given robust operations and adversarial strategies, we propose an analytical model of the local behavior of clusters, based on Markov chains. This local model provides an evaluation of the impact of malicious behaviors on the correctness of the system. Moreover, this local model is used to evaluate analytically the performance of the global system, allowing to characterize its global behavior with respect to its dynamics and to the presence of malicious nodes, and then to validate our approach.

Monitoring a system is the ability of collecting and analyzing relevant information provided by the monitored devices so as to be continuously aware of the system's state. However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. The usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. In [34], we propose to address this issue by pushing the monitoring task at the edge of the network. We present a peer-to-peer-based architecture, which enables nodes to self-organize according to their "health" indicators. By exploiting both temporal and spatial correlations that exist between a device and its vicinity, our approach guarantees that only isolated anomalies (an anomaly is isolated if it impacts solely a monitored device) are reported on the fly to the network operator. We show that the end-to-end detection process, i.e., from the local detection to the management operator reporting, requires a logarithmic number of messages in the size of the network. This work led to the patent [91] with Technicolor.

In [66] we continued previous efforts in the design of peer-to-peer networks for transmitting video content. In the past, we develop tools allowing a perceptual quality-based design tool. In [66], we explore an architectural idea where the video stream is decomposed in sequential sets of chunks that we call "windows". The paper explores some aspects of the performance of such a transmission scheme. The techniques used are Markovian models which are simulated, and deterministic dynamical systems that allow for some equilibrium analysis.

5.7. Monte Carlo

Participants: Bruno Tuffin, Gerardo Rubino, Pablo Sartor.

We maintain a research activity in different areas related to dependability, performability and vulnerability analysis of communication systems, using both the Monte Carlo and the Quasi-Monte Carlo approaches to evaluate the relevant metrics. Monte Carlo (and Quasi-Monte Carlo) methods often represent the only tool able to solve complex problems of these types. However, when the events of interest are rare, simulation requires a special attention, to accelerate the occurrence of the event and get unbiased estimators of the event of interest with a sufficiently small relative variance. This is the main problem in the area. Dionysos' work focuses then in dealing with the rare event situation.

In [72] we have overviewed how the zero-variance importance sampling can be approximated in classical reliability problems. In general, we look for estimators such that the relative accuracy of the output is "controlled" when the rarity is getting more and more critical. Different robustness properties of estimators have been defined in the literature. However, these properties are not adapted to estimators coming from a parametric family for which the optimal parameter is random due to a learning algorithm. These estimators have random accuracy. For this reason, we motivate in [65] the need to define probabilistic robustness properties. We especially focus on the so-called probabilistic bounded relative error property. We additionally provide sufficient conditions, both in general and in Markov settings, to satisfy such a property, and hope that it will foster discussions and new works in the area.

In [43] and [18] we present results concerning the evaluation using Monte Carlo techniques, of a specific reliability metric for communication networks, based not only on connectivity properties, as in the classical network reliability measure, but also in the lengths of the paths. In [43], we propose bounds of the metric that

can be used to derive a variance reduction technique. In [18], we describe techniques to analyze what could be called performability aspects of networks also based on the number of hops between sources and terminals. Let us also mention here our publication [16], where we discuss the exact computation of these new types of metrics, and [29], where other related combinatorial problems are discussed (here, optimization problems also based on connectivity properties, from the design point of view). In [17], we propose a new version of the RVR principle, leading to a variance reduction technique for the classic network reliability problem. Paper [28] proposes a splitting algorithm for the same problem. The approach is quite straightforward, after the static problem is transformed into a dynamic one using the well known Creation Process. In [42] we explore a very general conditioning-based approach, including as a particular case the family of splitting procedures. We explore this idea through the analysis of dependability properties of complex systems using Markov models.

When looking specifically at static network reliability models, as described in the previous paragraph, it is often typically assumed that the failures of their components are independent. This assumption allows for the design of efficient Monte Carlo algorithms that can estimate the network reliability in settings where it is a rare-event probability. Despite this computational benefit, independent component failures is frequently not a realistic modeling assumption for real-life networks. In [39] we show how the splitting methods for rare-event simulation can be used to estimate the reliability of a network model that incorporates a realistic dependence structure via the Marshal-Olkin copula.

In [15], we present a versatile Monte Carlo method for estimating multidimensional integrals, with applications to rare-event probability estimation. The method fuses two distinct and popular Monte Carlo simulation methods, Markov chain Monte Carlo and importance sampling, into a single algorithm. We show that for some applied numerical examples the proposed Markov Chain importance sampling algorithm performs better than methods based solely on importance sampling or MCMC.

Finally, in two presentations [67] and [32] we discuss the main problems when analyzing rare events using Monte Carlo methods, focusing on robustness properties of the corresponding estimators.

5.8. Analytic models

Participants: Bruno Sericola, Gerardo Rubino, Raymond Marie, Laura Aspirot.

Fluid models are powerful tools for evaluating the performance of packet telecommunication networks. By masking the complexity of discrete packet based systems, fluid models are in general easier to analyze and yield simple dimensioning formulas. Among fluid queuing systems, those with arrival rates modulated by Markov chains are very efficient to capture the burst structure of packet arrivals, notably in the Internet because of bulk data transfers. By exploiting the Markov property, very efficient numerical algorithms can be designed to estimate performance metrics such the overflow probability, the delay of a fluid particle or the duration of a busy period. In [76], we analyze the transient behaviour 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.

In [36], another type of fluid model is considered. We present preliminary results on the analysis of a Machine Repairman Model when the number of machines goes to infinity. The analysis is based on identifying appropriate fluid limits of the associated stochastic processes. We are currently working on the analysis of the speed of the convergence of these stochastic processes towards their fluid limits.

In [19], we present an approximate method for the transient analysis of stiff CTMC. The origin of the method is due to S. M. Ross who proposed to approximate the transient probability at a deterministic time t by the value of the transient probability at a random time X where X is an Erlang random variable having expectation t. The major contributions of the paper are the use of new numerical techniques to solve the basic equations of the original method and the extension of the method to reward measures. We also conduct an experimental evaluation of the resulting errors using non-trivial examples.

In [86], we presented an extension of ROBDDs that is able to accommodate certain dependencies among their (Boolean) variables. In particular, this extension shows evidence of being applicable to evaluating the dependability (reliability, availability) of systems whose structures are representable by a Boolean function. This extension consists of three main parts. The first part is the notion of a phratry with its associated new definitions and constraints. The second part consists of the adaptation and complementation of the original rules used in the construction of ROBDDs. The final part concerns additional custom-made steps needed to determine the functional valuations that are specific to solving measure in question.

The survivability of a system being its ability to function during and after a failure, we developed in [63] a model to study the power distribution in smart grids during the (transient) period that starts after a failure till the system fully recovers. The proposed model bridges power flow modeling of reactive power compensation with performability/survivability modeling of automation distribution networks. We use a Markov chain to characterize the phased recovery of the system after a failure. Then, we associate with each state of the Markov chain a set of corresponding rewards to characterize the active and reactive power supplied and demanded in that state. We connect the survivability model with an availability model, to produce a generalization of the System Average Interruption Duration Index (SAIDI) and the Customer Average Interruption Duration Index (CAIDI), which are two of the most important power grid reliability metrics. The survivability model allows us to obtain closed form expressions for the SAIDI and related metrics.

In [62], we consider the case of important systems located on operational sites far away from logistic support forces, either because the operational site is in an inhospitality place, or because it is not profitable to maintain a dedicated team on the operational site. Due to the importance of the systems, some service level agreement has been signed, including conditional financial clauses. To take into account such a situation, a preventive maintenance is realized according to projected calendars. The paper shows that, given that the life-times of equipments are supposed to be Erlang-k distributed, it is optimal to realize a preventive maintenance, as long as the ratio of the two intervention costs C_p/C_c is lower than the ratio $(k-1)/k$, C_p being the cost of a preventive maintenance intervention and C_c being the cost of a curative maintenance intervention (because of excessive delay, there is a significant penalty associated with each curative maintenance intervention). The methodology to compute the optimal value of the period T^* and the corresponding optimal cost per time unit are presented, for a given value of the ratio C_p/C_c. An extended version of this work has been accepted for publication in a journal ([26]).

The study [60] focuses on the determination of the probability distributions of two random variables, the asymptotic "up-time" and "down-time" of a system for the sake of potential "Service Level Agreement". In these new generation agreements, penalizations can be enforced for a too long "down-time" or for a too short "up-time". First, we determine the probability distributions of the two random variables "up-time" and "down-time", for a system with a general structural function. Second, we point out the importance of rare events such as the backorders in the contribution of a large tail distribution of the down-time. Respectively, we exhibit the importance of redundant structures and also of sub-system hyper-exponential lifetimes in the existence of short up-times, with respect to the mean up-time value of the system.

The study [61] deals with the determination of spares of systems of systems of the same type (such as fleet of aircraft, fleet of ship). For a multi-site workshop and multi-level of repair organization, we present an optimization algorithm using the criteria of expected number of backorders as local objective. With respect to a previous algorithm based only on the criteria of the global availability of the system, the new algorithm is, for large maintenance systems, very efficient, in terms of execution time and in of data manipulation.

The study [41] concerns the performance evaluation of crisis management systems with respect to the dimensioning of the system. By definition, a crisis has no steady state and the study must be done on the transient behavior. A faithful model was built (in ALTARICA) and solved thanks to simulation. Our own participation was mainly to determine the number of objects to create such that the simulation ends successfully with a high probability, before running out of available objects.

Last, in [54] we continue the exploration of the concept of duality proposed by Anderson, applied to the analysis of the transient behavior of queueing systems. This work analyzes the transient distribution of the

number of customers in a Restart Markovian queue, where together with “typical” customers other signals arrive to the queue having as a consequence the removal of all the customers present in the system.

DISTRIBCOM Project-Team

6. New Results

6.1. Fundamental results and algorithms: distributed planning

Participants: Eric Fabre, Loig 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/marked 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 contribution to the domain was the first optimal factored planning algorithm [47]. 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 [53] 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 2012, we have extended this framework in two directions. The first one considers large planning problems for which the interaction graph of components is not a tree. It is well known that message passing algorithms (also called belief propagation) is optimal on trees. To recover such a situation where distributed optimal planning can be resolved exactly, one therefore has to smartly group components into larger ones in order to recover a tree of larger components. This is done at the expense of the complexity in the resolution of local planning problems (which augments exponentially with the number of assembled components). Alternately, one can also ignore that the graph is not a tree, and thus use the so-called loopy belief propagation, which requires minor adaptations. This results in a new approach to the resolution of planning problems, where approximate solutions are provided: one can check that the computed plans are valid, but their optimality is not guaranteed. We have experimented this turbo-planning idea on a series of random benchmarks, some of them being not accessible to standard planning methods. The results are surprisingly good: distributed plans are found in most cases, and are often close to optimal. However, no theoretical results can yet support this phenomenon [30].

The second extension to distributed planning concerns the multi-agent version of the central A* (A-star) algorithm, which is at the core of numerous planners. By contrast with the previous setting, we do not build all plans here, in a distributed manner, but perform a search for an optimal plan. The centralized version of A* performs a depth-first search of a winning path in a graph, guided by some heuristic function that orients the search towards the goal. In our setting, several path searches must be performed in the graphs of the different components (or local planning problems), under the constraint that the provided paths are compatible, i.e. agree on the execution of the common actions. The resulting local paths must also be jointly optimal, once their costs are added. We have proposed a complete solution to this problem, called A# (A-sharp) [29]. Our efforts now aim at mixing these ideas with the turbo planning approach.

6.2. Fundamental results and algorithms: communication with messages and scenarios

Participants: Loïc Hélouët, Rouwaida Abdallah, Claude Jard, Blaise Genest, Sundararaman Akshay.

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*.

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 satisfiable, 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 ρ of a TC-MSC G is associated with a TC-MSC T_ρ obtained by concatenation of TC-MSC along ρ. 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 with paths of G. Because of inconsistent timing constraints, some path may have no possible realization (i.e $L(T_\rho = \varnothing)$). One can even design a MSC Graph G such that $L(G) = \varnothing$ - such TC-MSC graph is clearly inconsistent-. It has been shown [49] that checking whether $L(G) = \varnothing$ 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_ρ 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. This extends decidability results beyond regular specifications. The second restriction is K-non-zenoness, which imposes that for a fixed K, for every path ρ of G, there exists one realization such that at every date d, at most K events occur between dates 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 [10]. We actually succeeded to use a different technique by symbolically encoding the configuration reached. It allows to remove the K-non-zeno restriction, we don't need the seminal result on timed automata of Alur-Dill 1994, and we have a true partial order algorithm, which does not need to consider different interleavings of the same execution [18].

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 [51] that a simple projection of a HMSC on each of its processes 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 proposed an implementation solution that uses 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). This synthesis algorithm is consistent for a particular syntactic class of scenarios, namely the class of local HMSCs. This work was accepted for publication in [14].

6.3. Fundamental results and algorithms: timed models

Participants: Claude Jard, Aurore Junier, Sundararaman Akshay, Loïc Hélouët.

Our work on that subject mainly concerns Time Petri Nets (TPNs) and their robustness. Robustness of timed systems aims at studying whether infinitesimal perturbations in clock values can result in new discrete behaviors. A model is robust if the set of discrete behaviors is preserved under arbitrarily small (but positive) perturbations. We have tackled this problem for Time Petri Nets (TPNs for short) by considering the model of parametric guard enlargement which allows time-intervals constraining the firing of transitions in TPNs to be enlarged by a (positive) parameter.

We have shown that TPNs are not robust in general and that checking if they are robust with respect to standard properties (such as boundedness, safety) is undecidable. We have also provided two decidable robustly bounded subclasses of TPNs, and shown that one can effectively build a timed automaton which is timed bisimilar even in presence of perturbations. This allowed us to apply existing results for timed automata to these TPNs and show further robustness properties. This work was published in [20].

In a second work, we have considered robustness issues in Time Petri Nets (TPN) under constraints imposed by an external architecture. Our main objective was to check whether a timed specification, given as a TPN behaves as expected when subject to additional time and scheduling constraints. These constraints are given by another TPN that constrains the specification via read arcs. Our robustness property says that the constrained net does not exhibit new timed or untimed behaviors. We show that this property is not always guaranteed but that checking for it is always decidable in 1-safe TPNs. We further show that checking if the set of untimed behaviors of the constrained and specification nets are the same is also decidable. Next we turn to the more powerful case of labeled 1-safe TPNs with silent transitions. We show that checking for the robustness property is undecidable even when restricted to 1-safe TPNs with injective labeling, and exhibit a sub-class of 1-safe TPNs (with silent transitions) for which robustness is guaranteed by construction. This sub-class already lies close to the frontiers of intractability. This work was published in [19].

Finally, in cooperation with IRCCyN in Nantes, we defined a more general model, called "clock transition systems", which generalizes both TPNs and networks of timed automata [32]. This model will allow us to transfer new results on TPNs to the timed automata community.

6.4. Fundamental results and algorithms: dynamic epistemic logic

Participants: Guillaume Aucher, François Schwarzentruber.

Within the research line related to Dynamic Epistemic Logic (DEL), we have addressed two parallel lines of research, which have resulted in two publications [22] and [21]. The first deals with the computational complexity of the model checking problem and the satisfiability problem of DEL and the second deals with providing formal means to reason about the effects of sequences of events on the beliefs of multiple agents when these events are only partially specified. This second line of research is a continuation of the work started last year and was motivated by concerns and problems stemming from the Univerself project of Eric Fabre about IMS network.

1. Although DEL is an influential logical framework for representing and reasoning about information change, little is known about the computational complexity of its associated decision problems. In fact, we only know that for public announcement logic, a fragment of DEL, the satisfiability problem and the model-checking problem are respectively PSPACE-complete and in P. We contributed to fill this gap by proving that for the DEL language with event models, the model-checking problem is, surprisingly, PSPACE-complete. Also, we proved that the satisfiability problem is NEXPTIME-complete. In doing so, we provided a sound and complete tableau method deciding the satisfiability problem.
2. Let us consider a sequence of formulas providing partial information about an initial situation, about a set of events occurring sequentially in this situation, and about the resulting situation after the occurrence of each event. From this whole sequence, we want to infer more information, either about the initial situation, or about one of the events, or about the resulting situation after one of the events. Within the framework of Dynamic Epistemic Logic, we show that these different kinds of problems are all reducible to the problem of inferring what holds in the final situation after the occurrence of

all the events. We then provide a tableau method deciding whether this kind of inference is valid. We implement it in LotrecScheme and show that these inference problems are NEXPTIME-complete. We extend our results to the cases where the accessibility relation is serial and reflexive and illustrate them with the coordinated attack problem.

Parallely to the study of abstract dynamic epistemic logic, we initiate the study of the interaction of argumentation theory and epistemic reasoning [33].

6.5. Fundamental results and algorithms: statistical model checking

Participants: Sean Sedwards, Benoit Boyer, Kevin Corre, 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 modeling 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.

Plasma has been embedded in a tool chain for the design and the verification of Systems of Systems. The tool has also been used in a planing algorithm.

6.6. Fundamental results and algorithms: quantitative model checking and quantitative specification theories

Participants: Ulrich Fahrenberg, Blaise Genest, Axel Legay, Sundararaman Akshay, Louis-Marie Traonouez, Benoit Delahaye.

In 2012 we have successfully widened the applicability of interface and specification theories to systems with quantitative information such as energy usage, time constraints, or hybrid variables. Building on work done in 2011, we have introduced general quantitative specification theories. These provide a framework for reasoning about a wide range of different specification theories for different quantitative settings. We have provide one particularly important instantiation of the framework, which allows quantitative reasoning about real-time specifications.

Work on timed specifications theory has been continued in 2012 around the tool ECDAR. New case studies have been tested using the tool. These results, published in STTT, demonstrate the interest of the compositional approach for analyzing large systems. Besides the theory of robust specifications has been extended to allow a parametric estimation of the robustness. These results have been implemented in a new tool PyECDAR.

In 2012, we also successfully pursued our work on probabilistic specification theories by enhancing the framework of Abstract Probabilistic Automata, that we introduced in 2010, with several new operators. We first introduced a notion of satisfaction for stuttering implementations and showed how this new notion fits in the framework of APAs. Stuttering implementations are Probabilistic Automata that allow "silent" transitions by using local variables that are invisible to the specification. In this context, we also introduced a new logic, called ML-(A)PA that allows specifying properties of APA specifications and stuttering PA implementations. Our next contribution was to introduce a new difference operator. Given two specification APAs, their difference is a new APA that represents all implementations satisfying the one but not the other. This novel operator brings a new light to the well-known domain of counter-example generation.

Concerning Markov Chains, we have developed a new logic, LTL-I, which can only reason about fixed intervals instead of point values. We developed ϵ under and over approximation of formulas of this logics in [17], with associated algorithms. In all but few cases, we know that results of these algorithms are exact answers, while we didn't need to compute precisely and explicitly every probability involved. Another line of research is to consider very large Markov chain represented by Dynamic Bayesian Network. In [15], we compute only approximated results, as the size of the underlying Markov Chain is too big. However, evaluation of the algorithm shows small errors of our algorithm compared with the exact value.

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 [55].

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 the consumer of a given service. To account for variability and uncertainty in QoS, we proposed in previous work 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, when a called service improves its performance, then so does the overall orchestration.

Last 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 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 developed a general algebra to capture how QoS parameters are transformed while traversing the orchestration and we extended our study of monotonicity. Finally, we have developed corresponding contract composition procedures. This year, John Thywissen (from UT Austin) and Ajay Kattepur have prototyped a toolbox for Orc to support QoS-management. A journal paper is submitted.

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. An implementation on Orc is under development and a publication is in preparation.

Besides this main line of work, the additional topic of *Negotiation Strategies for Probabilistic Contracts in Web Services Orchestrations* has been addressed by Ajay Kattepur as part of his thesis, see [31]. Service Level Agreements (SLAs) have been proposed in the context of web services to maintain acceptable quality of service (QoS) performance. This is specially crucial for composite service orchestrations that can invoke many atomic services to render functionality. A consequence of SLA management entails efficient negotiation proto- cols among orchestrations and invoked services. In composite services where data and QoS (modeled in a probabilistic setting) interact, it is difficult to pick an individual atomic service to negotiate with. A superior improvement in one negotiated domain (eg. latency) might mean deterioration in another domain (eg. cost). In this work, we propose an integer programming formulation based on first order stochastic dom- inance as a strategy for re-negotiation over multiple services. A consequence of this is better end-to-end performance of the orchestration compared to random strategies for re-negotiation. We also demonstrate this optimal strategy can be applied to negotiation protocols specified in languages such as Orc. Such strategies are necessary for composite services where QoS contributions from individual atomic services vary significantly.

6.8. Specific studies: active documents and web services

Participants: Albert Benveniste, Loïc Hélouët, Sundararaman Akshay.

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.

Last year, we have developed 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. This year, we have started an experiment with a case study. We have proposed a distributed chess service palteform; the main idea is to use choreographies to provide solutions for chess problems, relying on an orchestration of specialized services for different phases of a game (opening, end of game, or collecting positions databases. We expect preliminary results in 2013.

Last year, 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: a commercial site, for instance, needs to manage several interactions at a given time. The current trend, as in BPEL, is to associate a unique identifier with each session. Modeling realistic sessions 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 formalism has the expressive power of reset Petri nets). The strength of this formalism is to allow designing systems that use sessions without the obligation to provide identifiers. Its drawback is that it only allows for the design of systems with a fixed number of agents. This year, we have continued extending last year's work with Ph. Darondeau from the S4 Team, and with M. Mukund from the Chennai Mathematical Institute to allow design of systems with sessions and allowing for an arbitrary number of agents.

6.9. 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 (one for shutdown of the equipment, one for restart). This work has been patented in June 2012 jointly with Alcatel-Lucent, and a publication on the topic was accepted at IM'2013.

6.10. 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 cooperation 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.

Since 2011, an important effort was dedicated to a key and difficult part of this approach: the definition of a methodology for self-modeling. 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.

Elements of methodology achieving these goals were proposed in 2011, and further refined in 2012. Besides, we have also worked on the definition of generic Bayesian networks, that could translate into mathematical terms the dependency relations between network resources, in order to reason about them for failure diagnosis. A methodology was then designed to reason on such models. The idea is that one should first consider a subset of network resources (at a given granularity), in order to localize the origin of a given malfunction. The natural start point is the graph of all resources involved in the delivery of the malfunctioning service. As the fault localization is statistical, the model is then progressively expanded to capture more network elements and thus more observations, and thus refine the diagnosis. This model expansion is performed by introducing first the most informative network elements, using information theory criteria. The result is a fault localization algorithm that explores only part of the network, and builds at runtime the necessary part of the model it should use to explain a malfunction [28]. The current efforts aim at extending these ideas to allow for the refinement of the model of some component (multiresolution reasoning).

FUN Team

5. New Results

5.1. Routing in FUN

Participants: Nicolas Gouvy, Xu Li, Nathalie Mitton.

Wireless sensor and actuator/robot networks need some routing mechanisms to ensure that data travel the network to the sink with some guarantees. The FUN research group has investigated different geographic routing paradigms. It first has considered a static network in which the routing either enhances the energy cost [22], [10], balances the load over nodes [21], [8] or respects traffic priorities [18].

A more complex routing paradigm has been proposed in [25] for k-anycasting. In k-anycasting, a sensor wants to report event information to any k sinks in the network. This is important to gain in reliability and efficiency in wireless sensor and actor networks. In this paper, we describe KanGuRou, the first position-based energy efficient k-anycast routing which guarantees the packet delivery to k sinks as long as the connected component that contains s also contains sufficient number of sinks. A node s running KanGuRou first computes a tree including k sinks among the M available ones, with weight as low as possible. If this tree has $m \geq 1$ edges originated at node s, s duplicates the message m times and runs m times KanGuRou over a subset of defined sinks. Simulation results show that KanGuRou allows up to 62% of energy saving compared to plain anycasting.

We then assumed that the sink that collects data is actually mobile and travels the network. Sensor nodes need thus to update the position of the sink in a smart fashion in order to limit the overhead generated by this update. In [9], 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.

When the network is composed of mobile sensors that have the faculty to control their mobility, this property can be exploited to enhance routing performance. In [3], we are interested in energy-aware routing algorithms that explicitly take advantage of 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 results in early termination of information delivery. To mitigate these problems, we propose a family of routing algorithm called CoMNet (Connectivity preservation Mobile routing protocols for actuator and sensor NETworks), that uses local information and modifies the network topology to support resource efficient transmissions. Our extensive simulations show that CoMNet has high energetic performance improvement compared to existing routing algorithms. More importantly, we show that CoMNet guarantees network connectivity and efficient resource consumption.

5.2. Self-organization

Participants: Tony Ducrocq, Xu Li, Nathalie Mitton.

Self-organization encompasses several mechanisms [35]. This year, the FUN research group contributes to some of them such as neighbor discovery, localization, clustering and topology control in FUN.

5.2.1. Neighbor discovery

To perform routing or any specific task, 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 [31], we combine parameters of the neighborhood discovery (sending frequency of hello messages and changes in the neighborhood tables) and transmission range of the nodes. We present two algorithms that adapt transmission range of the sensors in a mobile WSN by still adapting frequency of hello messages in order to save energy and get accurate neighborhood tables. The first solution is based on the knowledge of turnover - change in the number of neighbors in consecutive iterations of the neighborhood discovery - used in conjunction with an adaptation of the message frequency and the transmission range, minimizing overall transmission cost of hello messages. The second solution is based on the computation of optimal range knowing the nodes' speed. Both algorithms are based on theoretical analysis. Results show that they are energy efficient and outperform solutions of the literature by maintaining high accuracy.

5.2.2. 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 [7], 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 [39] routing. We discuss their potential applications and some open problems.

5.2.3. 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 [6], 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.

5.2.4. Clustering

Clustering in wireless sensor networks is an efficient way to structure and organize the network. It aims to identify a subset of nodes within the network and bind it a leader (i.e. cluster-head). This latter becomes in charge of specific additional tasks like gathering data from all nodes in its cluster and sending them by using a longer range communication to a sink. As a consequence, a cluster-head exhausts its battery more quickly than regular nodes. In [14], we present BLAC, a novel Battery-Level Aware Clustering family of schemes. BLAC considers the battery-level combined with another metric to elect the cluster-head. It comes in four variants. The cluster-head role is taken alternately by each node to balance energy consumption. Due to the local nature of the algorithms, keeping the network stable is easier. BLAC aims to maximize the time with all nodes alive to satisfy application requirements. Simulation results show that BLAC improves the full network lifetime 3-time more than traditional clustering schemes by balancing energy consumption over nodes and still delivering high data percentage.

5.3. Self-deployment

Participants: Milan Erdelj, Xu Li, Karen Miranda, Enrico Natalizio, Tahiry Razafindralambo, Dimitris Zorbas.

Robot self-deployment may have different purposes. The FUN research group has addressed four of them that are *(i)* area coverage, *(ii)* barrier coverage, *(iii)* point of interest coverage and *(iv)* deployment for substitution networks.

5.3.1. Area coverage

In [1], with the focus on the self-organizing capabilities of nodes in WSRN, we propose a movement-assisted technique for nodes self-deployment. Specifically, 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).

In [4], we tackle the issue in a different way. We leverage prediction 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 sub-modular 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.

In [34], [13], we leverage some assumptions. One of the main operations in wireless sensor networks is the surveillance of a set of events (targets) that occur in the field. In practice, a node monitors an event accurately when it is located closer to it, while the opposite happens when the node is moving away from the target. This detection accuracy can be represented by a probabilistic distribution. Since the network nodes are usually randomly deployed, some of the events are monitored by a few nodes and others by many nodes. In applications where there is a need of a full coverage and of a minimum allowed detection accuracy, a single node may not be able to sufficiently cover an event by itself. In this case, two or more nodes are needed to collaborate and to cover a single target. Moreover, all the nodes must be connected with a base station that

collects the monitoring data. In this paper we describe the problem of the minimum sampling quality, where an event must be sufficiently detected by the maximum possible amount of time. Since the probability of detecting a single target using randomly deployed static nodes is quite low, we present a localized algorithm based on mobile nodes. Our algorithm sacrifices a part of the energy of the nodes by moving them to a new location in order to satisfy the desired detection accuracy. It divides the monitoring process in rounds to extend the network lifetime, while it ensures connectivity with the base station. Furthermore, since the network lifetime is strongly related to the number of rounds, we propose two redeployment schemes that enhance the performance of our approach by balancing the number of sensors between densely covered areas and areas that are poorly covered. Finally, our evaluation results show an over 10 times improvement on the network lifetime compared to the case where the sensors are static. Our approaches, also, outperform a virtual forces algorithm when connectivity with the base station is required. The redeployment schemes present a good balance between network lifetime and convergence time.

5.3.2. Barrier coverage

Barrier coverage problem in emerging mobile sensor networks has been an interesting research issue. Existing solutions to this problem aim to decide one-time movement for individual sensors to construct as many barriers as possible, which may not work well when there are no sufficient sensors to form a single barrier. In [19], we try to achieve barrier coverage in sensor scarcity case by dynamic sensor patrolling. In specific, we design a periodic monitoring scheduling (PMS) algorithm in which each point along the barrier line is monitored periodically by mobile sensors. Based on the insight from PMS, we then propose a coordinated sensor patrolling (CSP) algorithm to further improve the barrier coverage, where each sensor's current movement strategy is decided based on the past intruder arrival information. By jointly exploiting sensor mobility and intruder arrival information, CSP is able to significantly enhance barrier coverage.We prove that the total distance that the sensors move during each time slot in CSP is the minimum. Considering the decentralized nature of mobile sensor networks, we further introduce two distributed versions of CSP: S-DCSP and G-DCSP. Through extensive simulations, we demonstrate that CSP has a desired barrier coverage performance and S-DCSP and G-DCSP have similar performance as that of CSP.

5.3.3. Point of Interest 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 base station is thus a fundamental issue. This algorithm addresses the problem of autonomous deployment of mobile sensors that need to cover a predefined PoI with a connectivity constraint. In our algorithm [2], each sensor moves toward a PoI but has also to maintain the connectivity with a subset of its neighboring sensors that are part of the Relative Neighborhood Graph (RNG). The Relative Neighborhood Graph reduction is chosen so that global connectivity can be provided locally. Our deployment scheme minimizes the number of sensors used for connectivity thus increasing the number of monitoring sensors. Analytical results, simulation results and practical implementation are provided to show the efficiency of our algorithm.

We then extended this coverage to multiple points of interest in [15], [16]. 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 this work, we have extended [2] to multiple PoI coverage and combined it to and environment exploration in order to capture the dynamics of the monitored area. We examine the performance of our scheme through extensive simulation campaigns.

5.3.4. Substitution networks

A substitution network is a temporary network that will be deployed to support a base network in trouble and help it to provide best service. [11], [24] present how the mobility of routers impacts the performance of a wireless substitution network. To that end, we simulate a scenario where a wireless router moves between three static nodes, a source and two destinations of UDP traffic. Specifically, our goal is to deploy or redeploy the mobile relays so that application-level requirements, such as data delivery or latency, are met. Our proposal for a mobile relay achieves these goals by using an adaptive approach to self-adjust their position based

on local information. We obtain results on the performance of end-to-end delay, jitter, loss percentage, and throughput under such mobility pattern for the mobile relay. We show how the proposed solution is able to adapt to topology changes and to the evolution of the network characteristics through the usage of limited neighborhood knowledge.

5.4. MAC layer

Participant: Tahiry Razafindralambo.

Multihop wireless networks are used to provide Internet connectivity to the users and the level of performance and quality expected by these users are increasing. In order to meet these performance and quality requirements, wireless communications should be enhanced. Previous works from the literature show that the performance and quality provided by an IEEE 802.11-based multihop wireless network are far from optimal and that there exist different ways to increase the efficiency and the quality of service of such a network. Some studies show that using the medium state as a parameter to tune the behavior of an IEEE 802.11-based multihop network is an appropriate way to proceed. A station in a IEEE 802.11-based multihop wireless network senses the medium either busy or idle. The durations of idle periods and busy periods and their distributions have a clear impact on the network and nodes performance. The understanding of the relationship between these indicators, namely idle and busy periods, the network topology and the traffic, would give new insights to enhance the performance and quality of multihop wireless networks. Due to its multihop and distributed nature, the characterization of idle period durations is difficult in such a network. In [27], [26], we explore the characterization of idle period distribution by proposing a new analytical model and provides an application of this characterization with the design of an adaptive backoff algorithm based on idle periods.

5.5. Servicing

Participants: Xu Li, Kalypso Magklara, Nathalie Mitton, Tahiry Razafindralambo, Dimitris Zorbas.

Servicing wireless sensor networks include many primitives. It can range from cloud connection [12] to mobile IPv6 management [29] going through energy prediction [20] and launching mobile robots on request of a specific demand [5] or to reload sensors [23], [17].

5.5.1. Node reloading

A critical problem of wireless sensor networks is the network lifetime, due to the device's limited battery lifetime. The nodes are randomly deployed in the field and the system has no previous knowledge of their position. To tackle this problem, in [23], we use a mobile robot, that discovers the nodes around it and replaces the active nodes, whose energy is drained, by fully charged inactive nodes. We propose two localized algorithms, that can run on the robot and that decide, which nodes to replace. We simulate our algorithms and our findings show that all nodes that fail are replaced in a short period of time.

In [17] we focus on an emerging kind of cooperative networking system in which a small team of robotic agents lies at a base station. Their mission is to service an already-deployed WSN by periodically replacing all damaged sensors in the field with passive, spare ones so as to preserve the existing network coverage. This novel application scenario is here baptized as "multiple-carrier coverage repair" (MC2R) and modeled as a new generalization of the vehicle routing problem. A hybrid metaheuristic algorithm is put forward to derive nearly-optimal sensor replacement trajectories for the robotic fleet in a short running time. The composite scheme relies on a swarm of artificial fireflies in which each individual follows the exploratory principles featured by Harmony Search. Infeasible candidate solutions are gradually driven into feasibility under the influence of a weak Pareto dominance relationship. A repair heuristic is finally applied to yield a full-blown solution. To the best of our knowledge, our scheme is the first one in literature that tackles MC2R instances. Empirical results indicate that promising solutions can be achieved in a limited time span.

5.5.2. Energy prediction

One way to improve energy supply for sensor nodes is through ambient energy harvesting from solar, thermal or vibration energy sources coupled with rechargeable energy storage. Wireless sensors have to adapt to the stochastic nature of the energy harvesting sources. We are convinced that predicting the temporal availability of ambient energy resources is vital to plan the harvesting efficiency, optimum resource utilization and energy conservation within sensor nodes. In [20] we propose a novel two stage Autoregressive Weather conditioned Solar Energy Prediction (AWSEP) model which is characterized by low computational complexity and is used to accurately estimate the amount of solar energy that will be harvested in the near future in a particular region. Our algorithm re-learns the model parameters during the prediction processing situations where the prediction error becomes larger than a predefined prediction error threshold mainly because of the unreliable nature of outdoor solar energy sources caused by changes in weather conditions. The proposed AWSEP model performance is evaluated by varying energy harvesting source prediction intervals, sampling rates, trade-offs in prediction accuracy and computational costs using real solar datasets. We concluded that AWSEP algorithm is more accurate, has reduced computational complexity and memory utilization than other prediction schemes in literature. Our proposed algorithm can assist a node to automatically adapt to the changing weather conditions for effective power management and sensing task scheduling.

5.5.3. Servicing sensor nodes

Due to the robots' potential to unleash a wider set of networking means and thus augment the network performance, WSRNs have rapidly become a hot research area. In [5], we elaborate on WSRNs from two unique standpoints: robot task allocation and robot task fulfillment. The former deals with robots cooperatively deciding on the set of tasks to be individually carried out to achieve a desired goal; the latter enables robots to fulfill the assigned tasks through intelligent mobility scheduling.

5.6. Experimenting

Participants: Nathalie Mitton, Julien Vandaele.

One of the goal of the FUN research group is to validate through experimentations and to provide tools for this purpose. Therefore, the FIT platform is deployed, together with a set of tutorials [37]. Nevertheless, we are aware that using testbed platforms for validation is already a great step but it can not satisfy all needs. This is why we also investigate alternatives as emulation. In [28], [32] for instance, we propose a specifically designed experimental setup using a relatively small number of nodes forming a real one-hop neighborhood used to emulate any real WSN. The source node is a fixed sensor, and all other sensors are candidate forwarding neighbors towards a virtual destination. The source node achieves one forwarding step, then the virtual destination position and neighborhood are adjusted. The same source is used again to repeat the process. The main novelty is to spread available nodes regularly following a hexagonal pattern around the central node, used as the source, and selectively use subsets of the surrounding nodes at each step of the routing process to provide the desired density and achieve changes in configurations. Compared to real testbeds, our proposition has the advantages of emulating networks with any desired node distribution and densities, which may not be possible in a small scale implementation, and of unbounded scalability since we can emulate networks with an arbitrary number of nodes. Finally, our approach can emulate networks of various shapes, possibly with holes and obstacles. It can also emulate recovery mode in geographic routing, which appears impossible with any existing approach.

5.7. RFID middleware

Participants: Roberto Quilez, Nathalie Mitton.

The Object Naming System (ONS) is a central lookup service used in the EPCglobal network for retrieving location information about a specific Electronic Product Code (EPC). This centralized solution lacks scalability and fault tolerance and encounters some political issues. In [30], we present the design principles of a fully-distributed multi-root solution for ONS lookup service. In distributed systems, the problem of providing a scalable location service requires a dynamic mechanism to associate identification and location. We design, prototype, and evaluate PRONS, a DHT-based solution for the multi-root problem. We show that PRONS achieves significant performance levels while respecting a number of neutrality requirements.

5.8. VANET

Participants: Enrico Natalizio, Thierry Delot.

Today, thanks to vehicular networks, drivers may receive useful information produced or relayed by neighboring sensors or vehicles (e.g., the location of an available parking space, of a traffic congestion, etc.). In [33], we address the problem of providing assistance to the driver when no recent information has been received on his/her vehicle. Therefore, we present a cooperative scheme to aggregate, store and exchange these events in order to have an history of past events. This scheme is based on a dedicated spatio-temporal aggregation structure using Flajolet-Martin sketches and deployed on each vehicle. Contrary to existing approaches considering data aggregation in vehicular networks, our main goal here is not to save network bandwidth but rather to extract useful knowledge from previous observations. In this paper, we present our aggregation data structure, the associated exchange protocol and a set of experiments showing the effectiveness of our proposal.

In [36], we present a novel vehicular communication protocol, which aims to reduce the effect of broadcast storm problem in VANETs (Vehcular AdHoc NETworks). When the traffic density is above a certain value (e.g., when vehicles are in congested traffic scenarios), one of the most serious problems is the increase of packet collisions and medium contentions among vehicles which attempt to communicate. Our proposed technique, namely Selective Reliable Broadcast protocol (SRB), is intended to limit the number of packet transmissions, by means of opportunistically selecting neighboring nodes, acting as relay nodes. As a result, the number of forwarder vehicles is strongly reduced, while network performance is preserved. SRB belongs to the class of broadcast protocols, and exploits the traditional vehicular partitioning behavior to select forwarders. Each cluster is automatically detected as a zone of interest, whenever a vehicle is approaching, and packets will be forwarded only to selected vehicles, opportunistically elected as cluster-heads. In respect of traditional broadcast approaches, the main strengths of SRB are the efficiency of detecting clusters and selecting forwarders in a fast way, in order to limit the broadcast storm problem. Simulation results have been carried out both in urban and highway scenarios, in order to validate the effectiveness of SRB, in terms of cluster detection and reduction of number of selected forwarders.

GANG Project-Team

4. New Results

4.1. Understanding graph representations

4.1.1. Notions of Connectivity in Overlay Networks

Participants: Yuval Emek, Pierre Fraigniaud, Amos Korman, Shay Kutten, David Peleg.

How well connected is the network? This is one of the most fundamental questions one would ask when facing the challenge of designing a communication network. Three major notions of connectivity have been considered in the literature, but in the context of traditional (single-layer) networks, they turn out to be equivalent. The paper [17], introduces a model for studying the three notions of connectivity in multi-layer networks. Using this model, it is easy to demonstrate that in multi-layer networks the three notions may differ dramatically. Unfortunately, in contrast to the single-layer case, where the values of the three connectivity notions can be computed efficiently, it has been recently shown in the context of WDM networks (results that can be easily translated to our model) that the values of two of these notions of connectivity are hard to compute or even approximate in multi-layer networks. The current paper shed some positive light into the multi-layer connectivity topic: we show that the value of the third connectivity notion can be computed in polynomial time and develop an approximation for the construction of well connected overlay networks.

4.1.2. Connected graph searching

Participants: Lali Barrière, Paola Flocchini, Fedor V. Fomin, Pierre Fraigniaud, Nicolas Nisse, Nicola Santoro, Dimitrios M. Thilikos.

In the graph searching game the opponents are a set of searchers and a fugitive in a graph. The searchers try to capture the fugitive by applying some sequence of moves that include placement, removal, or sliding of a searcher along an edge. The fugitive tries to avoid capture by moving along unguarded paths. The search number of a graph is the minimum number of searchers required to guarantee the capture of the fugitive. In [2], we initiate the study of this game under the natural restriction of connectivity where we demand that in each step of the search the locations of the graph that are clean (i.e. non-accessible to the fugitive) remain connected. We give evidence that many of the standard mathematical tools used so far in classic graph searching fail under the connectivity requirement. We also settle the question on "the price of connectivity", that is, how many searchers more are required for searching a graph when the connectivity demand is imposed. We make estimations of the price of connectivity on general graphs and we provide tight bounds for the case of trees. In particular, for an n-vertex graph the ratio between the connected searching number and the non-connected one is while for trees this ratio is always at most 2. We also conjecture that this constant-ratio upper bound for trees holds also for all graphs. Our combinatorial results imply a complete characterization of connected graph searching on trees. It is based on a forbidden-graph characterization of the connected search number. We prove that the connected search game is monotone for trees, i.e. restricting search strategies to only those where the clean territories increase monotonically does not require more searchers. A consequence of our results is that the connected search number can be computed in polynomial time on trees, moreover, we show how to make this algorithm distributed. Finally, we reveal connections of this parameter to other invariants on trees such as the Horton–Strahler number.

4.1.3. Computing with Large Populations Using Interactions

Participants: Olivier Bournez, Pierre Fraigniaud, Xavier Koegler.

We define in [12], a general model capturing the behavior of a population of anonymous agents that interact in pairs. This model captures some of the main features of opportunistic networks, in which nodes (such as the ones of a mobile ad hoc networks) meet sporadically. For its reminiscence to Population Protocol, we call our model Large-Population Protocol, or LPP. We are interested in the design of LPPs enforcing, for every $\nu \in [0,1]$, a proportion ν of the agents to be in a specific subset of marked states, when the size of the population grows to infinity; In which case, we say that the protocol computes ν. We prove that, for every $\nu \in [0,1]$, ν is computable by a LPP if and only if ν is algebraic. Our positive result is constructive. That is, we show how to construct, for every algebraic number $\nu \in [0,1]$, a protocol which computes ν.

4.1.4. Collaborative Search on the Plane without Communication

Participants: Ofer Feinerman, Zvi Lotker, Amos Korman, Jean-Sébastien Sereni.

In [19], we use distributed computing tools to provide a new perspective on the behavior of cooperative biological ensembles. We introduce the Ants Nearby Treasure Search (ANTS) problem, a generalization of the classical cow-path problem which is relevant for collective foraging in animal groups. In the ANTS problem, k identical (probabilistic) agents, initially placed at some central location, collectively search for a treasure in the two-dimensional plane. The treasure is placed at a target location by an adversary and the goal is to find it as fast as possible as a function of both k and D, where D is the distance between the central location and the target. This is biologically motivated by cooperative, central place foraging, such as performed by ants around their nest. In this type of search there is a strong preference to locate nearby food sources before those that are further away. We focus on trying to find what can be achieved if communication is limited or altogether absent. Indeed, to avoid overlaps agents must be highly dispersed making communication difficult. Furthermore, if the agents do not commence the search in synchrony, then even initial communication is problematic. This holds, in particular, with respect to the question of whether the agents can communicate and conclude their total number, k. It turns out that the knowledge of k by the individual agents is crucial for performance. Indeed, it is a straightforward observation that the time required for finding the treasure is $\Omega(D + D^2/k)$, and we show in this paper that this bound can be matched if the agents have knowledge of k up to some constant approximation. We present a tight bound for the competitive penalty that must be paid, in the running time, if the agents have no information about k. Specifically, this bound is slightly more than logarithmic in the number of agents. In addition, we give a lower bound for the setting in which the agents are given some estimation of k. Informally, our results imply that the agents can potentially perform well without any knowledge of their total number k, however, to further improve, they must use some information regarding k. Finally, we propose a uniform algorithm that is both efficient and extremely simple, suggesting its relevance for actual biological scenarios.

4.1.5. Memory Lower Bounds for Randomized Collaborative Search and Implications for Biology

Participants: Ofer Feinerman, Amos Korman.

Initial knowledge regarding group size can be crucial for collective performance. We study in [18], this relation in the context of the Ants Nearby Treasure Search (ANTS) problem, which models natural cooperative foraging behavior such as that performed by ants around their nest. In this problem, k (probabilistic) agents, initially placed at some central location, collectively search for a treasure on the two-dimensional grid. The treasure is placed at a target location by an adversary and the goal is to find it as fast as possible as a function of both k and D, where D is the (unknown) distance between the central location and the target. It is easy to see that $T = \Omega(D + D^2/k)$ time units are necessary for finding the treasure. Recently, it has been established that $O(T)$ time is sufficient if the agents know their total number k (or a constant approximation of it), and enough memory bits are available at their disposal. In this paper, we establish lower bounds on the agent memory size required for achieving certain running time performances. To the best our knowledge, these bounds are the first non-trivial lower bounds for the memory size of probabilistic searchers. For example, for every given positive constant ϵ, terminating the search by time $O(\log^{1-\epsilon} k \cdot T)$ requires agents to use $\Omega(\log \log k)$ memory bits.

From a high level perspective, we illustrate how methods from distributed computing can be useful in generating lower bounds for cooperative biological ensembles. Indeed, if experiments that comply with our setting reveal that the ants' search is time efficient, then our theoretical lower bounds can provide some insight on the memory they use for this task.

4.1.6. What Can be Computed without Communications?

Participants: Heger Arfaoui, Pierre Fraigniaud.

When playing the boolean game (δ, f), two players, upon reception of respective inputs x and y, must respectively output a and b satisfying $\delta(a, b) = f(x, y)$, in absence of any communication. It is known that, for $\delta(a, b) = a \oplus b$, the ability for the players to use entangled quantum bits (qbits) helps. In [10], we show that, for δ different from the exclusive-or operator, quantum correlations do not help. This result is an invitation to revisit the theory of dis- tributed checking, a.k.a. distributed verification, currently sticked to the usage of decision functions δ based on the and-operator, hence potentially preventing us from using the potential benefit of quantum effects.

4.1.7. Decidability Classes for Mobile Agents Computing modularity

Participants: Andrzej Pelc, Pierre Fraigniaud.

We establish in [21], a classification of decision problems that are to be solved by mobile agents operating in unlabeled graphs, using a deterministic protocol. The classification is with respect to the ability of a team of agents to solve the problem, possibly with the aid of additional information. In particular, our focus is on studying differences between the decidability of a decision problem by agents and its verifiability when a certificate for a positive answer is provided to the agents. Our main result shows that there exists a natural complete problem for mobile agent verification. We also show that, for a single agent, three natural oracles yield a strictly increasing chain of relative decidability classes.

4.1.8. Randomized Distributed Decision

Participants: Pierre Fraigniaud, Amos Korman, Merav Parter, David Peleg.

The paper [20] tackles the power of randomization in the context of locality by analyzing the ability to "boost" the success probability of deciding a distributed language. The main outcome of this analysis is that the distributed computing setting contrasts significantly with the sequential one as far as randomization is concerned. Indeed, we prove that in some cases, the ability to increase the success probability for deciding distributed languages is rather limited.

We focus on the notion of a $(p, q)-$decider for a language L , which is a distributed randomized algorithm that accepts instances in L with probability at least p and rejects instances outside of L with probability at least q. It is known that every hereditary language that can be decided in t rounds by a $(p, q)-$decider, where $p^2 + q > 1$, can be decided deterministically in $O(t)$ rounds. One of our results gives evidence supporting the conjecture that the above statement holds for all distributed languages and not only for hereditary ones, by proving the conjecture for the restricted case of path topologies. For the range below the aforementioned threshold, namely, $p^2 + q \leq 1$, we study the class $B_k(t)$ (for $k \in \mathbb{N}^* \cup \{\infty\}$) of all languages decidable in at most t rounds by a $(p, q)-$decider, where $p^{1+\frac{1}{k}} + q > 1$. Since every language is decidable (in zero rounds) by a $(p, q)-$decider satisfying $p + q = 1$, the hierarchy B_k provides a spectrum of complexity classes between determinism ($k = 1$, under the above conjecture) and complete randomization ($k = \infty$). We prove that all these classes are separated, in a strong sense: for every integer $k \geq 1$, there exists a language L satisfying $L \in B_{k+1}(0)$ but $L \notin B_k(t)$ for any $t = o(n)$. In addition, we show that $B_\infty(t)$ does not contain all languages, for any $t = o(n)$. In other words, we obtain the hierarchy $B_(t) \subset B_2(t) \subset \cdots \subset B_\infty(t) \subset$ All. Finally, we show that if the inputs can be restricted in certain ways, then the ability to boost the success probability becomes almost null, and in particular, derandomization is not possible even beyond the threshold $p^2 + q = 1$.

4.1.9. The Worst Case Behavior of Randomized Gossip

Participants: Hervé Baumann, Pierre Fraigniaud, Hovhannes A. Harutyunyan, Rémi de Verclos.

In [11] we consider the quasi-random rumor spreading model introduced by Doerr, Friedrich, and Sauerwald in [SODA 2008], hereafter referred to as the list-based model. Each node is provided with a cyclic list of all its neighbors, chooses a random position in its list, and from then on calls its neighbors in the order of the list. This model is known to perform asymptotically at least as well as the random phone-call model, for many network classes. Motivated by potential applications of the list-based model to live streaming, we are interested in its worst case behavior.

Our first main result is the design of an $O(m + nlogn)-$time algorithm that, given any n-node m-edge network G, and any source-target pair $s, t \in V(G)$, computes the maximum number of rounds it may take for a rumor to be broadcast from s to t in G, in the list-based model. This algorithm yields an $O(n(m + nlogn))-$time algorithm that, given any network G, computes the maximum number of rounds it may take for a rumor to be broadcast from any source to any target, in the list-based model. Hence, the list-based model is computationally easy to tackle in its basic version.

The situation is radically different when one is considering variants of the model in which nodes are aware of the status of their neighbors, i.e., are aware of whether or not they have already received the rumor, at any point in time. Indeed, our second main result states that, unless P=NP , the worst case behavior of the list-based model with the additional feature that every node is perpetually aware of which of its neighbors have already received the rumor cannot be approximated in polynomial time within a $\left(\frac{1}{n}\right)^{\frac{1}{2}-\epsilon}$ multiplicative factor, for any $\epsilon > 0$. As a byproduct of this latter result, we can show that, unless P=NP , there are no PTAS enabling to approximate the worst case behavior of the list-based model, whenever every node perpetually keeps track of the subset of its neighbors which have sent the rumor to it so far.

4.1.10. Asymptotic modularity

Participants: Fabien de Montgolfier, Mauricio Soto, Laurent Viennot.

Modularity (Newman-Girvan) has been introduced as a quality measure for graph partitioning. 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 the modularity of a variety of graph classes. In [23], we first consider simple graph classes such as tori and hypercubes. We show that these regular graph families have asymptotic modularity 1 (that is the maximum possible). We extend this result to trees with bounded degree, allowing us to give a lower bound of 2 over average degree for graph classes with low maximum degree (included power law graphs for a sufficiently large exponent).

4.1.11. Modeling social networks

Participants: Nidhi Hegde, Laurent Massoulié, Laurent Viennot.

Social networks offer users new means of accessing information, essentially relying on “social filtering”, i.e. propagation and filtering of information by social contacts. The sheer amount of data flowing in these networks, combined with the limited budget of attention of each user, makes it difficult to ensure that social filtering brings relevant content to the interested users. Our motivation in [26] is to measure to what extent self-organization of the social network results in efficient social filtering. To this end we introduce flow games, a simple abstraction that models network formation under selfish user dynamics, featuring user-specific interests and budget of attention. In the context of homogeneous user interests, we show that selfish dynamics converge to a stable network structure (namely a pure Nash equilibrium) with close-to-optimal information dissemination. We show in contrast, for the more realistic case of heterogeneous interests, that convergence, if it occurs, may lead to information dissemination that can be arbitrarily inefficient, as captured by an unbounded “price of anarchy”. Nevertheless the situation differs when users’ interests exhibit a particular structure, captured by a metric space with low doubling dimension. In that case, natural autonomous dynamics converge to a stable configuration. Moreover, users obtain all the information of interest to them in the corresponding dissemination, provided their budget of attention is logarithmic in the size of their interest set.

4.1.12. Additive Spanners and Distance and Routing Labeling Schemes for Hyperbolic Graphs

Participants: Victor Chepoi, Feodor Dragan, Bertrand Estellon, Michel Habib, Yann Vaxès, Yang Xiang.

δ-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 2δ. 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], we study unweighted δ-hyperbolic graphs. Using the Layering Partition technique, we show that every n-vertex δ-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 δ-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 δ-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.

4.1.13. Constructing a Minimum phylogenetic Network from a Dense triplet Set

Participants: Michel Habib, Thu-Hien To.

For a given set $\mathcal{L}$ of species and a set $\mathcal{T}$ of triplets on $\mathcal{L}$, we seek to construct a phylogenetic network which is consistent with $\mathcal{T}$ i.e. which represents all triplets of $\mathcal{T}$. The level of a network is defined as the maximum number of hybrid vertices in its biconnected components. When $\mathcal{T}$ is dense, there exist polynomial time algorithms to construct level-0, 1 and 2 networks (Aho et al., 1981; Jansson, Nguyen and Sung, 2006; Jansson and Sung, 2006; Iersel et al., 2009). For higher levels, partial answers were obtained in the paper by Iersel and Kelk (2008), with a polynomial time algorithm for simple networks. In [9] this paper, we detail the first complete answer for the general case, solving a problem proposed in Jansson and Sung (2006) and Iersel et al. (2009). For any k fixed, it is possible to construct a level-k network having the minimum number of hybrid vertices and consistent with $\mathcal{T}$, if there is any, in time $O\left(|(T)|^{k+1} n^{\lfloor \frac{4k}{3} \rfloor}\right)$.

4.1.14. Algorithms for Some $H-$Join Decompositions

Participants: Michel Habib, Antoine Mamcarz, Fabien de Montgolfier.

A homogeneous pair (also known as a 2-module) of a graph is a pair $\{M_1, M_2\}$ of disjoint vertex subsets such that for every vertex $x \notin (M_1 \cup M_2)$ and $i \in \{1, 2\}$, x is either adjacent to all vertices in M_i or to none of them. First used in the context of perfect graphs [Chvátal and Sbihi 1987], it is a generalization of splits (a.k.a 1-joins) and of modules. The algorithmics to compute them appears quite involved. In [22], we describe an $O(mn^2)-$time algorithm computing (if any) a homogeneous pair, which not only improves a previous bound of $O(mn^3)$ [Everett, Klein and Reed 1997], but also uses a nice structural property of homogenous pairs. Our result can be extended to compute the whole homogeneous pair decomposition tree, within the same complexity. Using similar ideas, we present an $O(nm^2)-$time algorithm to compute a $N-$join decomposition of a graph, improving a previous $O(n^6)$ algorithm [Feder et al. 2005]. These two decompositions are special case of $H-$joins [Bui-Xuan, Telle and Vatshelle 2010] to which our techniques apply.

4.1.15. Detecting 2-joins faster

Participants: Pierre Charbit, Michel Habib, Nicolas Trotignon, Kristina Vušković.

2-joins are edge cutsets that naturally appear in the decomposition of several classes of graphs closed under taking induced subgraphs, such as balanced bipartite graphs, even-hole-free graphs, perfect graphs and claw-free graphs. Their detection is needed in several algorithms, and is the slowest step for some of them. The classical method to detect a 2-join takes $O(n^3 m)$ time where n is the number of vertices of the input graph and m the number of its edges. To detect *non-path* 2-joins (special kinds of 2-joins that are needed in all of the known algorithms that use 2-joins), the fastest known method takes time $O(n^4 m)$. Here, we give an $O(n^2 m)$-time algorithm for both of these problems. A consequence is a speed up of several known algorithms.

4.2. Large Scale Networks Performance and Modeling

4.2.1. Spatial Interactions of Peers and Performance of File Sharing Systems

Participants: François Baccelli, Fabien Mathieu, Ilkka Norros.

We propose in [24] a new model for peer-to-peer networking which takes the network bottlenecks into account beyond the access. This model allows one to cope with key features of P2P networking like degree or locality constraints or the fact that distant peers often have a smaller rate than nearby peers. We show that the spatial point process describing peers in their steady state then exhibits an interesting repulsion phenomenon. We analyze two asymptotic regimes of the peer-to-peer network: the fluid regime and the hard–core regime. We get closed form expressions for the mean (and in some cases the law) of the peer latency and the download rate obtained by a peer as well as for the spatial density of peers in the steady state of each regime, as well as an accurate approximation that holds for all regimes. The analytical results are based on a mix of mathematical analysis and dimensional analysis and have important design implications. The first of them is the existence of a setting where the equilibrium mean latency is a decreasing function of the load, a phenomenon that we call super-scalability.

4.2.2. User Behavior Modeling: Four Months in DailyMotion

Participants: Yannick Carlinet, The Dang Huynh, Bruno Kauffmann, Fabien Mathieu, Ludovic Noirie, Sébastien Tixeuil.

The growth of User-Generated Content (UGC) traffic makes the understanding of its nature a priority for network operators, content providers and equipment suppliers. In [13], we study a four-month dataset that logs all video requests to DailyMotion made by a fixed subset of users. We were able to infer user sessions from raw data, to propose a Markovian model of these sessions, and to study video popularity and its evolution over time. The presented results are a first step for synthesizing an artificial (but realistic) traffic that could be used in simulations or experimental testbeds.

4.2.3. Multi-Carrier Networks: on the Manipulability of Voting Systems

Participants: François Durand, Fabien Mathieu, Ludovic Noirie.

Today, Internet involves many actors who are making revenues on it (operators, companies, service providers,...). It is therefore important to be able to make fair decisions in this large-scale and highly competitive economical ecosystem. One of the main issues is to prevent actors from manipulating the natural outcome of the decision process. For that purpose, game theory is a natural framework. In that context, voting systems represent an interesting alternative that, to our knowledge, has not yet been considered. They allow competing entities to decide among different options. Strong theoretical results showed that all voting systems are susceptible to be manipulated by one single voter, except for some "degenerated" and non-acceptable cases. However, very little is known about how much a voting system is manipulable in practical scenarios. In [25], we investigate empirically the use of voting systems for choosing end-to-end paths in multi-carrier networks, analyzing their manipulability and their economical efficiency. We show that one particular system, called Single Transferable Vote (STV), is largely more resistant to manipulability than the natural system which tries to get the economical optimum. Moreover, STV manages to select paths close to the economical optimum, whether the participants try to cheat or not.

4.3. Fault Tolerance in Distributed Networks

4.3.1. Wait-Freedom with Advice

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Eli Gafni, Petr Kuznetsov.

In [14], we motivate and propose a new way of thinking about failure detectors which allows us to define, quite surprisingly, what it means to solve a distributed task *wait-freeusing a failure detector*. In our model, the system is composed of *computation* processes that obtain inputs and are supposed to output in a finite number of steps and *synchronization* processes that are subject to failures and can query a failure detector. We assume that, under the condition that *correct* synchronization processes take sufficiently many steps, they provide the computation processes with enough *advice* to solve the given task wait-free: every computation process outputs in a finite number of its own steps, regardless of the behavior of other computation processes. Every task can thus be characterized by the *weakest* failure detector that allows for solving it, and we show that every such failure detector captures a form of set agreement. We then obtain a complete classification of tasks, including ones that evaded comprehensible characterization so far, such as renaming or weak symmetry breaking.

4.3.2. Partial synchrony based on set timeliness

Participants: Markos Aguilera, Carole Delporte-Gallet, Hugues Fauconnier, Sam Toueg.

We introduce in [1], a new model of partial synchrony for read-write shared memory systems. This model is based on the simple notion of set timeliness—a natural generalization of the seminal concept of timeliness in the partially synchrony model of Dwork et al. (J. ACM 35(2):288–323, 1988). Despite its simplicity, the concept of set timeliness is powerful enough to define a family of partially synchronous systems that closely match individual instances of the $t-$resilient $k-$set agreement problem among n processes, henceforth denoted $(t, k, n)-$agreement. In particular, we use it to give a partially synchronous system that is synchronous enough for solving $(t, k, n)-$ agreement, but not enough for solving two incrementally stronger problems, namely, $(t + 1, k, n)-$agreement, which has a slightly stronger resiliency requirement, and $(t, k-1, n)-$agreement, which has a slightly stronger agreement requirement. This is the first partially synchronous system that separates these sub-consensus problems. The above results show that set timeliness can be used to study and compare the partial synchrony requirements of problems that are strictly weaker than consensus.

4.3.3. Byzantine Agreement with Homonyms in Synchronous Systems

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Hung Tran-The.

In [15], [6], we consider the Byzantine agreement problem (BA) in synchronous systems with homonyms. In this model different processes may have the same authenticated identifier. In such a system of n processes sharing a set of l identifiers, we define a distribution of the identifiers as an integer partition of n into l parts $n_1, ..., n_l$ giving for each identifier i the number of processes having this identifier.

Assuming that the processes know the distribution of identifiers we give a necessary and sufficient condition on the integer partition of n to solve the Byzantine agreement with at most t Byzantine processes. Moreover we prove that there exists a distribution of l identifiers enabling to solve Byzantine agreement with at most t Byzantine processes if and only if $n > 3t$, $l > t$ and where $r = n \mod l$.

This bound is to be compared with the $l > 3t$ bound proved in Delporte-Gallet et al. (2011) when the processes do not know the distribution of identifiers.

4.3.4. Homonyms with forgeable identifiers

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Hung Tran-The.

In [16], we refine the Byzantine Agreement problem (BA) in synchronous systems with homonyms, in the particular case where some identifiers may be forgeable. More precisely, the n processes share a set of l $(1 \leq l \leq n)$ identifiers. Assuming that at most t processes may be Byzantine and at most k $(t \leq k \leq l)$ of these identifiers are forgeable in the sense that any Byzantine process can falsely use them, we prove that Byzantine Agreement problem is solvable if and only if $l > 2t + k$. Moreover we extend this result to systems with authentication by signatures in which at most k signatures are forgeable and we prove that Byzantine Agreement problem is solvable if and only if $l > t + k$.

4.4. Discrete Optimization Algorithms

4.4.1. Estimating satisfiability

Participants: Yacine Boufkhad, Thomas Hugel.

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 [3], 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.

4.4.2. Attractive force search algorithm for piecewise convex maximization problems

Participants: Dominique Fortin, Ider Tseveendorj.

In [8], 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.

4.4.3. B-spline interpolation: Toeplitz inverse under corner perturbations

Participant: Dominique Fortin.

For Toeplitz matrices associated with degree 3 and 4 uniform B-spline interpolation, the inverse may be analytically known [7], saving the standard inverse calculations. It generalizes to any degree as a row of the Eulerian numbers triangle.

HIPERCOM Project-Team

6. New Results

6.1. Time Slot Assignment in Wireless Sensor Networks

Participants: Pascale Minet, Ridha Soua, Erwan Livolant.

6.1.1. NP-completeness of the Time Slot Assignment problem

In data gathering applications, wireless sensor networks (WSNs) collect data from sensor nodes towards a sink in a multi-hop convergecast structure. Assigning equal channel access to each node may lead to congestion and inefficient use of the bandwidth. That is why we focus on traffic-aware solutions. More precisely, we investigate the Time Slot Assignment problem, where nodes are assigned time slots to transmit their data to the sink, while minimizing the total number of slots. We considered the generalized h-hop Time Slot Assignment problem for any positive integer h, where any two nodes that are less than or equal to h-hop away are not scheduled simultaneously. We proved its NP-completeness.

6.1.2. Multichannel Slot Assignment

The throughput requirement of data gathering applications is difficult to meet with a single wireless channel. Furthermore, the considered channel may be temporarily jammed. That is why, we focus on a multichannel time slot assignment that minimizes the data gathering cycle. We first formalize the problem as a linear program and compute the optimal time needed for a raw data convergecast in various multichannel topologies (linear, multi-line, tree). These optimal times apply to nodes equipped with one or several radio interfaces. This work generalizes the results established by Incel. We then propose our algorithm called MODESA and prove its optimality in various multichannel topologies. We evaluate its performances in terms of number of slots, maximum buffer size and number of active/sleep switches per node. Furthermore, we present variants of MODESA achieving a load balancing between the channels used.

6.1.3. Multisink Multichannel Slot Assignment

We generalize this work, taking into account the existence of several sinks. We focus on the data gathering problem with differentiated traffic, each addressed to a specific sink in multichannel WSNs. In order to find a collision-free optimized multichannel time slot assignment that minimizes the data gathering cycle, we propose a centralized traffic-aware algorithm called MUSIKA. We formulate the problem as a linear program and compute the optimal time needed for a raw data convergecast in various multichannel topologies (linear, multi-line, tree). More generally, we run simulations on various network topologies to evaluate the performance of MUSIKA in terms of cycle length, maximum buffer size and slot reuse ratio for different use cases: redundant functional processing chains, different application functionalities per sink.

6.2. Multi-Sink Wireless Sensor deployment and energy analysis

Participants: Paul Mühlethaler, Nadjib Achir.

We propose a general framework for multi-sink Wireless Sensors networks (WNSs). This framework is devoted to computing the optimal deployment of sinks for a given maximum number of hops between nodes and sinks. This framework allows an estimation of the energy consumption to be computed. We consider the energy consumed due to reporting, forwarding and overhearing. In contrast to reporting and forwarding, the energy used in overhearing is difficult to estimate because it is dependent on the packet scheduling. We determine the upper-bound and lower-bound of overhearing. We also propose another estimation which can simulate non interfering parallel transmissions which is more tractable in large networks. We note that overhearing largely predominates in energy consumption. A large part of the optimizations and computations carried out in this paper are obtained using ILP formalization.

6.3. WSN Redeployment

Participants: Pascale Minet, Saoucene Mahfoudh Ridene, Ines Khoufi.

This is a joint work with Telecom SudParis: Anis Laouiti.

6.3.1. Centralized redeployment algorithm based on Virtual Forces

In many applications (e.g military, environment monitoring), wireless sensors are randomly deployed in a given area. Unfortunately, this deployment is not efficient enough to ensure full area coverage and total network connectivity. Hence, all the considered area must be covered by sensors ensuring that any event is detected in the sensing range of at least one sensor. In addition, the sensor network must be connected in terms of radio communication in order to forward the detected event to the sink(s). Thus, a redeployment algorithm has to be applied in order to achieve these two goals. In this context, we have proposed redeployment algorithms based on virtual forces. First, we have designed and simulated a centralized algorithm called CVFA. This algorithm is executed by a specific node which has global information of node positions.

6.3.2. Distributed redeployment algorithm based on Virtual Forces

Then, we proposed DVFA, Distributed Virtual forces Algorithm. Each node in the network executes DVFA and computes its new position based on information collected from its neighbors.
Performance evaluation shows that both CVFA and DVFA give very good coverage rate (between 98% and 100%) and ensure the connectivity between sensors.

6.3.3. Distributed redeployment algorithm based on Virtual Forces in the presence of obstacles

Moreover, in a real environment, obstacles such as trees, walls and buildings may exist and they may impact the deployment of wireless sensors. Obstacles can prohibit the network connectivity between nodes and create some uncovered holes or some accumulation of sensors in the same region. Consequently, an efficient wireless sensors deployment algorithm is required to ensure both coverage and network connectivity in the presence of obstacles. We have focused on this problem and enhanced our Distributed Virtual Force Algorithm (DVFA) to cope with obstacles. Simulation results show that DVFA gives very good performances even in the presence of obstacles.

6.4. Mesh Network Planning: Deployment and Canal Allocation

Participant: Nadjib Achir.

This is a joint work with University Paris XIII: A. Farsi, K. Boussetta.
We deal with the Wireless LAN planning problem. We study this problem and we propose to couple its two major issues: AP placement and channel assignment to treat them jointly. Here, we propose a novel fast and scalable three-phase heuristic algorithm (TPHA). Our proposal is able to resolve the defined multiobjective problem to provide (1) the efficient number of Access Points (APs) to be deployed, while (2) ensuring the coverage of all Test Points (TPs) and (3) maximizing their nominal data rate. To achieve the first objective, we propose an heuristic called MCL-ILP combining the quick decision making based on the Markovian CLustering algorithm and the exact solution provided by the Integer Linear Programming. Hence, a TPs-based Least Interfering Channel Search algorithm (TLICS) has been proposed for channel assignment to improve the throughput at TP locations. However, the Virtual Forces-based WLAN Planning Algorithm namely VFPA considers the results delivered by the two previous algorithms as an initial solution and tries to enhance it by adjusting the APs' positions and re-assigning their operating frequencies. Computational results exhibit that our proposal is highly beneficial to designing WLANs.

6.5. Routing in MANETs using slotted Aloha. End-to-end delays

Participants: Paul Mühlethaler, Iskander Banaouas.

This is a joint work with TREC: B. Blaszczyzyn.

Planar Poisson models with the Aloha medium access scheme have already proved to be very useful in studies of mobile ad-hoc networks (MANETs). However, it seems difficult to quantitatively study the performances of end-to-end routing in these models. In order to tackle this problem, in this paper we study a *linear stationary route embedded in an independent planar field of interfering nodes*. We consider this route as an idealization of a "typical" route in a MANET obtained by some routing mechanism. Such a decoupling allows us to obtain many numerically tractable expressions for local and mean end-to-end delays and the speed of packet progression, assuming slotted Aloha MAC and the Signal-to-Interference-and-Noise Ratio (SINR) capture condition, with the usual power-law path loss model and Rayleigh fading. These expressions show how the network performance depends on the tuning of Aloha and routing parameters and on the external noise level. In particular we show a need for a well-tuned lattice structure of fixed relaying nodes, which helps to relay packets on long random routes in the presence of a non-negligible noise. We also consider a *Poisson-line MANET model*, in which *all* nodes are located on roads forming a Poisson-line process. In this case our linear route is rigorously (in the sense of Palm theory) the typical route in this Poisson-line MANET.

6.6. Cognitive networks using a darwinian approach

Participant: Paul Mühlethaler.

This is a joint work with Alcatel Bell Labs: Philippe Jacquet.
We present a new approach for cognitive radio. In the usual approach the secondary network is in charge of monitoring the channel to determine whether or not the primary network is active in the area. If it is not, the secondary network is allowed to use the spectrum. In the new access scheme we propose, the primary network encompasses the techniques which allow it to capture the bandwidth even if the secondary network is transmitting in the area. The access scheme of the primary network preempts the secondary network activity. We present an access scheme which preempts the IEEE 802.11 decentralized scheme. This protocol is a generalized Carrier Sense Multiple Access scheme using active signaling. Instead of only sensing the carrier, this algorithm also transmits bursts of signal which may be sensed by the other nodes. If so, they give up the selection process. We show that this scheme preempts the IEEE 802.11 decentralized access scheme if the bursts transmitted by the node in the primary network are made up of special sequences which alternate between bursts of signal and periods of sensing. These sequences called (d, k) sequences encompass a minimum number d and a maximum number of k successive zeros during which the node senses the channel to find other possible concurrent transmissions. In practice we use $d = 0$ and k depends on the duration of the IEEE 802.11 interframe space and the duration of a signaling burst. We compute the number of $(0, k)$ sequences with respect to the length n of the sequence. We also show that (d, k) sequences (with $2d > k$) can be used if, by mistake, during the signaling phase one burst is not detected. We evaluate the number of such sequences.

6.7. Massive mobile dense wireless networks

Participants: Aline Carneiro Viana, Ana Cristina B. Kochem Vendramin, Kanchana Thilakarathna, Eduardo Mucceli.

routing protocols, analytical models, content distribution.

6.7.1. Scientific achievements

6.7.1.1. Social Relationship Classified

Understanding human mobility is of fundamental importance when designing new communication protocols that exploit opportunistic encounters among users. In particular, human behavior is characterized by an elevated rate of regularity, but random events are always possible in the routines of individuals as hardly predictable situations that deviate from the regular pattern and are unlikely to arise repeatedly in the future. These random events veil the ordinary patterns by introducing a significant amount of noise, thus making the process of knowledge discovery in social dataset a complex task. However, the ability to accurately identify random and social events in large datasets is essential to social analysis as well as to applications that rely

on a precise description of human routines, such as recommendation systems, forwarding strategies and ad-hoc message dissemination schemes focusing on coverage efficiency with a limited number of redundant messages. In such a context, we have proposed a strategy to analyze wireless network scenarios where mobile users interact in a rational manner, reflecting their interests and activity dynamics. Our strategy, named Random rElationship ClASsifier sTrategy (RECAST), allows to classify user relationships, separating random interactions from different kinds of social ties. The goal is achieved by observing how the real system differs from an equivalent one where entities decisions are completely random. We have evaluate the effectiveness of RECAST classification on datasets of real-world user contacts in diverse networking contexts. Our analysis unveils significant differences in the relationship dynamics of the datasets, proving that the evaluation of network protocols on a single dataset cannot lead to conclusions of general validity.

6.7.1.2. Social-aware Forwarding Protocol

Pervasiveness of computing devices, ubiquitous wireless communication, emergence of new applications, and cloud services are examples of current new emerging factors that emphasize the increasing need for adaptive networking solutions. The adaptation, most of the time, requires the design of more interdisciplinary approaches as those inspired by techniques coming from biology, social structures, games, and control systems. The approach we consider brings together solutions from different but complementary domains - i.e., networking, biology, and complex networks - aiming to deal with the problem of efficient data delivery in mobile and intermittently connected networks. For this, we have designed the Cultural Greedy Ant (CGrAnt) protocol to solve the problem of data delivery in mobile and intermittently connected networks referred as Delay Tolerant Networks (DTNs). CGrAnt is a hybrid Swarm Intelligence-based forwarding protocol designed to deal with the dynamic and complex environment of DTNs resulting from users mobility or varying conditions of wireless communications. CGrAnt is based on (1) Cultural Algorithms (CA) and Ant Colony Optimization (ACO) and (2) metrics which characterize opportunistic social connectivity between wireless users. CA and ACO are used to direct the network traffic, taking into account a set of social-aware metrics that may infer relevant structures in meeting regularities and mobility patterns of users. The most promising message forwarders are selected through a greedy transition rule based on local and global information captured from the DTN environment. Through simulation, we have analyzed the influence of ACO operators and CA's knowledge on CGrAnt performance. We have then compared the performance of CGrAnt with PROPHET and Epidemic protocols under varying networking parameters. Results have shown that CGrAnt achieves the highest delivery ratio and lowest byte redundancy.

6.7.1.3. Opportunistic Content Dissemination

Here, we focus on dissemination of content for delay tolerant applications/services, (i.e. content sharing, advertisement propagation, etc.) where users are geographically clustered into communities. Due to emerging security and privacy related issues, majority of users are becoming more reluctant to interact with strangers and are only willing to share information/content with the users who are previously identified as friends. In this environment, opportunistic communication will not be effective due to the lack of known friends within the communication range. Thus, we have proposed a novel architecture that addresses the issues of lack of trust, timeliness of delivery, loss of user control, and privacy-aware distributed mobile social networking by combining the advantages of distributed decentralized storage and opportunistic communications. We have formally defined a content replication problem in mobile social networks and show that it is computationally hard to solve optimally. Then, we have proposed a community based greedy heuristic algorithm with novel dynamic centrality metrics to replicate content in well-selected users, to maximize the content dissemination with limited number of replication. Using both real world and synthetic traces, we have shown that content replication can attain a large coverage gain and reduce the content delivery latency.

6.7.1.4. Data Offloading-aware Hotspot Deployment

With the steady growth of sales of smart-phones, the demand for services that generate mobile data traffic has grown tremendously. The growing use of traffic data generated from mobile devices overloads the network infrastructure, which is not always prepared to receive such demand. To tackle this problem, we are studying the mobile behavior and resource consumptions of people on a metropolitan area in a major city and turn it into a set of well located WiFi hotspots. For this, we have proposed a data offloading-aware hotspot deployment. It

is methodologically divided as (i) creation of a time dependent weighted graph to represent people's mobility, traffic and its relation with places/locations able to receive a hotspot, (ii) measurement of location's importance and selection of the best-ranked ones. Better positioned hotspots are likely to provide better coverage, and therefore, be able to offload more data.

6.7.2. Collaborations

- Professors Anelise Munaretto and Myriam Regattieri Delgado from Federal Technological University of Parana (UTFPR), Brazil,
- Professors Aruna Seneviratne and Henrik Petander from NICTA and School of EE&T, UNSW, Sydney, Australia,
- Pedro O.S. Vaz de Melo and Antonio A. F. Loureiro, Federal University of Minas Gerais, Brazil,
- Marco Fiore and Frederic Le Mouel from INSA Lyon, France,
- Katia Jaffrès-Runser, University of Toulouse, IRIT/ENSEEIHT, France.

6.8. New services and protocols

Participants: Aline Carneiro Viana, Guilherme Maia.

6.8.1. Scientific achievements

6.8.1.1. Network Discovery

Network discovery is a fundamental task in different scenarios of IEEE 802.15.4-based wireless personal area networks. Scenario examples are body sensor networks requiring health- and wellness-related patient monitoring or situations requiring opportunistic message propagation. Therefore, we have investigated optimized discovery of IEEE 802.15.4 static and mobile networks operating in multiple frequency bands and with different beacon intervals. We designed a linear programming model that allows finding two optimized strategies, named OPT and SWOPT, to deal with the asynchronous and multi-channel discovery problem. We have also proposed a simplified discovery solution, named SUBOPT, featuring a low-complexity algorithm requiring less memory usage. A cross validation between analytical, simulation, and experimental evaluation methods was performed. Our performance studies confirmed improvements achieved by our solutions in terms of first, average, and last discovery time as well as discovery ratio, when compared to IEEE 802.15.4 standard approach and the SWEEP approach known from the literature.

6.8.1.2. Distributed Data Storage

The deployment of large-scale Wireless Sensor Network (WSN) applications (e.g., environment sensing and military surveillance), which operate unattended for long periods of time and generate a considerable amount of data, poses several challenges. One of them is *how to retrieve the sensed data*. To tackle this issue, we have designed ProFlex, a distributed data storage protocol for large-scale heterogeneous wireless sensor networks (HWSNs) with mobile sinks. ProFlex guarantees robustness in data collection by intelligently managing data replication among selected storage nodes in the network. Contrarily to related protocols in the literature, ProFlex considers the resource constraints of sensor nodes and constructs multiple data replication structures, which are managed by more powerful nodes. Additionally, ProFlex takes advantage of the higher communication range of such powerful nodes and uses the long-range links to improve data distribution by storage nodes. When compared with related protocols, we have shown through simulation that Proflex has an acceptable performance under message loss scenarios, decreases the overhead of transmitted messages, and decreases the occurrence of the energy hole problem. Moreover, we have proposed an improvement that allows the protocol to leverage the inherent data correlation and redundancy of wireless sensor networks in order to decrease even further the protocol's overhead without affecting the quality of the data distribution by storage nodes.

6.8.2. Collaborations

- PhD Niels Karowski, Technische Universitat Berlin, Germany,
- Professor Adam Wolisz, Technische Universitat Berlin, Germany,
- Antonio A. F. Loureiro, Federal University of Minas Gerais, Brazil,

MADYNES Project-Team

6. New Results

6.1. Android Security

Participants: Olivier Festor, Abdelkader Lahmadi [contact].

Android-based devices include smartphones and tablets that are now widely adopted by users because they offer a huge set of services via a wide range of access networks (WiFi, GPRS/EDGE, 3G/4G). Android provides the core platform for developing and running applications. Those applications are available to the users over numerous online marketplaces. These applications are posted by developers, with little or no review process in place, leaving the market self-regulated by users. This policy generates a side-effect where users are becoming targets of different malicious applications which the goal is to steal their private information, collect all kind of sensitive data via sensors or abusing granted permissions to make surtaxed calls or messages. To address this security issue, monitoring the behaviour of running applications is a key technique enabling the identification of malicious activities.

During 2012, we have designed and developed a monitoring framework integrating observed network and system activities of a running application. We have developed an embedded NetFlow probe running on android devices to export observed network flow records observed to a collection point for their processing. Our embedded probe includes a new set of IPFIX information elements that we have designed [36] to encapsulate location information within exported flows using the IPFIX protocol.

We have also developed an embedded logging probe that exports available system logs to a collection point. The logs are then centrally processed and correlated with observed network flow records to extract an accurate behavior of an application including its network and in-device activities.

Our monitoring framework is different from available proposed solutions since we build a dynamic model to infer the running behavior of an Android application. This technique allows us to identify patched applications where a malicious activity has been added, cloned applications where the observed behavior is different from the expected behavior and privacy leaks where an application is contacting unexpected services.

6.2. Sensor networks monitoring

Participants: Alexandre Boeglin, Laurent Ciarletta, Olivier Festor, Abdelkader Lahmadi [contact], Emmanuel Nataf, Bilel Saadallah.

Low Power and Lossy Networks (LLNs) are made of interconnected wireless devices with limited resources in terms of energy, computing and communication. The communication channels are low-bandwidth, high loss rate and volatile wireless links subject to failure over time. They are dynamic and the connectivity is limited and fluctuant over time. Each node may loss frequently its connectivity with its neighborhood nodes. In addition, link layer frames have high constrains on their size and throughput is limited. These networks are used for many different applications including industrial automation, smart metering, environmental monitoring, homeland security, weather and climate analysis and prediction. The main issue in those networks is optimal operation combined with strong energy preservation. Monitoring, i.e the process of measuring sampled properties of nodes and links in a network, is a key technique in operational LLNs where devices need to be constantly or temporally monitored to assure their functioning and detect relevant problems which will result in an alarm being for- warded to the enterprise network for analysis and remediation.

During the year 2012, we developed novel approaches for the monitoring of LLNs. We developed and designed a novel algorithm and a supporting framework [18] that improves a poller-pollee monitoring architecture. We empower the poller-pollee placement decision process and operation by exploiting available routing data to monitor nodes status. In addition, monitoring data is efficiently embedded in any messages flowing through the network, drastically reducing monitoring overhead. Our approach is validated through both simulation, implementation and deployment on a 6LoWPAN-enabled network. Both simulations and large-scale testbed experiments assess the efficiency of our monitoring scheme. Results also demonstrate that our approach is less aggressive and less resource consuming than its competitors.

We developed a first fully operational CCNx stack [40] on a wireless sensor network. We implemented CCNx as a native C experimental extension of Contiki, an operating system dedicated to Internet of Things applications. Our extension [33] is based on the reference implementation of CCNx modified to run as a network driver on top of different available MAC protocols implementations in Contiki. Our goal is to design a monitoring and configuration framework that benefits from the content-centric approach to efficiently collect desired management content and apply in-network processing functions for nodes configuration and monitoring. This includes extending naming schema with monitoring oriented processing functions, optimizing data interests to minimize the communication overhead.

6.3. Management and monitoring of P2P networks

Participants: Isabelle Chrisment [contact], Olivier Festor, Juan Pablo Timpanaro.

In 2012, we have addressed operation, monitoring and security issues on several P2P target networks: KAD, BitTorrent and I2P.

Several large scale P2P networks operating on the Internet are based on a Distributed Hash Table. These networks offer valuable services, but they all suffer from a critical issue allowing malicious nodes to be inserted in specific places on the DHT for undesirable purposes (monitoring, distributed denial of service, pollution, etc.). While several attacks and attack scenarios have been documented, few studies have measured the actual deployment of such attacks and none of the documented countermeasures have been tested for compatibility with an already deployed network. In our work, we focus on the KAD network. Based on large scale monitoring campaigns, we demonstrated that the world-wide deployed KAD network suffers large number of suspicious insertions around shared contents and we quantify them. To cope with these peers, we proposed a new efficient protection algorithm based on analyzing the distribution of the peers ID found around an entry after a DHT lookup [3]. The evaluation of our solution showed that it detects the most efficient configurations of inserted peers with a very small false-negative rate, and that the countermeasures successfully filter almost all the suspicious peers. We demonstrate the direct applicability of our approach by implementing and testing our solution in real P2P networks

BitTorrent is a fast, popular, P2P filesharing application focused on fast propagation of content. Its trackerless approach uses a DHT based on Kademlia to search for sources when the hash of the metadata of the content to transfer is known. On the other hand, the eMule network uses the old ED2K protocol for filesharing including a system of priorized queues, but indexation is done through a solid Kademlia based DHT, named Kad. The Kad DHT stands for a search engine, wich provides an extra level to map keywords to file identifiers. We have designed an hybrid approach, compatible with both P2P file-sharing networks, which has the Kad advantages on indexation and the BitTorrent throughput for transfer while maintaining backward compatibility with both of these networks [42]. To validate our proposal we developed a prototype which supports content indexation provided by the Kad network and is able to transfer files using the BitTorrent protocol. Using this prototype, we measured the propagation of new content in clusters of aMule clients, BitTorrent clients, hybrid clients, and a mix of them.

In parallel, we continued our research about being anonymous when downloading from BitTorrent. Anonymous communications have been gaining more and more interest from Internet users as privacy and anonymity problems have emerged. Among anonymous enabled services, anonymous file-sharing is one of the most active one and is increasingly growing. Large scale monitoring on these systems allows us to grasp how they behave, which type of data is shared among users, the overall behavior in the system.

We presented the first monitoring study aiming to characterize the usage of the I2P network, a low-latency anonymous network based on garlic routing [23]. We characterized the file-sharing environment within I2P, and evaluated if this monitoring affects the anonymity provided by the network. We showed that most activities within the network are file-sharing oriented, along with anonymous web-hosting. We assessed the wide geographical location of nodes and network popularity. We also demonstrated that group-based profiling is feasible on this particular network [22].

Dedicated anonymous networks such as Freenet and I2P allow anonymous file-sharing among users. However, one major problem with anonymous file-sharing networks is that the available content is highly reduced, mostly with outdated files, and non-anonymous networks, such as the BitTorrent network, are still the major source of content. We showed that in a 30-days period, 21648 new torrents were introduced in the BitTorrent community, whilst only 236 were introduced in the anonymous I2P network, for four different categories of content. Therefore, how can a user of these anonymous networks access this varied and non-anonymous content without compromising its anonymity? In [24], we improved content availability in an anonymous environment by proposing the first internetwork model allowing anonymous users to access and share content in large public communities while remaining anonymous. We showed that our approach can efficiently interconnect I2P users and public BitTorrent swarms without affecting their anonymity nor their performance. Our model is fully implemented and freely usable.

6.4. Configuration security automation

Participants: Rémi Badonnel [contact], Martin Barrere, Olivier Festor.

The main research challenge addressed in this work is focused on enabling configuration security automation in autonomic networks and services. In particular our objective is 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 that 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 executed during these self-management activities may generate vulnerable configurations. A first part of our work in the year 2012 has been dedicated to the assessment of distributed vulnerabilities and to the elaboration of a collaborative management strategy for supporting their remediation. A configuration vulnerability is not necessarily local but can also be spread over several devices in the autonomic network. We have showed in [8] how such distributed vulnerabilities can be mathematically formalized and described in a machine readable manner, through the specification of the DOVAL (Distributed OVAL) language on top of OVAL (Open Vulnerability and Assessment Language). We have designed and evaluated a dedicated framework for exploiting these vulnerability descriptions, collecting device configurations and detecting distributed vulnerabilities using specific aggregation techniques. Once a vulnerability is identified in the autonomic network, several remediation actions can potentially be performed by the autonomic network over devices. For that purpose, we have introduced an XCCDF-based specification for expressing alternative treatments related to a distributed vulnerability. We have also proposed a collaborative scheme for selecting one of these treatments depending on the current context (device capabilities and willingness to participate) [6]. A second part of our work has focused on the extension of our solution to other environments. In particular we have worked on the integration of our vulnerability assessment strategy over the Android plateform [9]. We have put forward a mathematical model as well as an optimized method that provides solid foundations for this context. By maintaining low-consumption services monitoring the system, the proposed approach minimizes heavy task executions by only triggering assessment activities when configuration changes are detected or new vulnerability definitions are available. In light of this, we have developed a prototype that efficiently performs self-assessment activities, and also introduces dedicated web services for collecting OVAL descriptions and storing assessment results. We have performed an analytical evaluation of the proposed model as well as an extensive set of technical experiments that shows the feasibility of our solution. We are currently working on the issue of past hidden vulnerable states. A network compromised in the past by an unknown vulnerability at that moment may still constitute a potential security threat in the present. Accordingly, past unknown system

exposures are required to be taken into account. We are therefore investigating a novel strategy for identifying also such past hidden vulnerable configurations and increasing the overall security [9].

6.5. Cache Management in CCN

Participants: Thomas Silverston [contact], César Bernardini, Olivier Festor.

The Internet is currently mostly used for accessing content. Indeed, ranging from P2P file sharing to current video streaming services such as Youtube, it is expected that content will count for approximately 86% of the global consumer traffic by 2016.

While the Internet was designed for -and still focuses on- host-to-host communication (IP), users are only interested in actual content rather than source location. Hence, new Information-Centric Networking architectures (ICN) such as CCN, NetInf, Pursuit have been proposed giving high priority to efficient content distribution at large scale. Among all these new architectures, Content Centric Networking (CCN) has attracted considerable attention from the research community [2].

CCN is a network architecture based on named data where a packet address names content, not location. The notion of host as defined into IP does not exist anymore. In CCN, the content is not retrieved from a dedicated server, as it is the case for the current Internet. The premise is that content delivery can be enhanced by including per-node-caching as content traverses the network. Content is therefore replicated and located at different points of the network, increasing availability for incoming requests.

As content is cached along the path, it is crucial to investigate the caching strategy for CCN Networks and to propose new schemes adapted to CCN. We therefore designed *Most Popular Content* (MPC), a new caching strategy for CCN network [10].

Instead of storing all the content at every nodes on the path, MPC strategy caches only popular content. With MPC, each nodes count all the requests for a content and when it has been requested a large amount of time, the content will be cached at each node along the path. Otherwise, the content is not popular; it is transmitted but it is not cached into the network.

We implemented MPC into the ccnSim simulator and evaluate it through extensive simulations.

Our results demonstrate that using MPC strategy allow to achieve a higher Cache Hit in CCN networks and still reduces drastically the number of replicas. By caching only popular content, MPC helps at reducing the cache load at each node and the network resource consumption.

We expect that our strategy could serve as a base for studying name-based routing protocols. Being a suggestion based mechanism, it is feasible to adapt it to manage content among nodes, to predict popularity and to route content to destination. In addition, we are currently investigating the social relationship between users to improve our caching strategy for CCN networks.

6.6. QoS in Wireless Sensor Networks

Participants: François Despaux, Abdelkader Lahmadi, Bilel Nefzi, Hugo Cruz-Sanchez, Ye-Qiong Song [contact].

WSN research focus has progressively been moved from the energy issue to the QoS issue. Typical example is the MAC protocol design, which cares about not only low duty-cycle, but also high throughput with self-adaptation to dynamic traffic changes [21]. Our research on WSN QoS is thoroughly organized in three topics:

- MAC protocol design for both QoS and energy efficiency

[2] http://www.ccnx.org

The main result that we obtained in 2012 is a new hybrid CSMA/TDMA MAC protocol, called Queue-MAC, that dynamically adapts the duty-cycle according to the current network traffic. The queue length of nodes is used as the network traffic indicator. When the traffic increases, the active CSMA period is accordingly extended by adding dynamic TDMA slots, allowing thus to efficiently handle burst traffic under QoS constraints. This protocol is implemented on the STM32W108 SOC chips and compared with both a fixed duty-cycle reference protocol and an optimized IEEE802.15.4 MAC protocol. Through extensive experimental measurements, we showed that our queue-length aware hybrid CSMA/TDMA MAC protocol largely outperforms the compared protocols. The proposed protocol can be easily implemented through slight adaptation of the IEEE802.15.4 standard [25].

Many industrial WSN are based on IEEE802.15.4 standard. One of the critical issues is the scheduling of neighboring coordinators beacons. In [20], we presented TBoPS, a novel technique for scheduling beacons in the cluster-tree topology. TBoPS uses a dedicated period called beacon only period (BOP) to schedule beacons at the beginning of IEEE 802.15.4 superframe. The advantage of TBoPS is that every beacon-enabled node distributively selects a beacon schedule during association phase.

- QoS routing

  For supporting different QoS requirements, routing in WSN must simultaneously consider several criteria (e.g., minimizing energy consumption, hop counts or delay, packet loss probability, etc.). When multiple routing metrics are considered, the problem becomes a multi-constrained optimal path problem (MCOP), which is known as NP-complete. In practice, the complexity of the existing routing algorithms is too high to be implemented on the low cost and power constrained sensor nodes. Recently, Operator calculus (OC) has been developed by Schott and Staples with whom we collaborate. OC can be applied to solving MCOP problem with much lower complexity and can deal with dynamic topology changes (which is the case in duty-cycled WSN). The OC approach has been successfully applied to a concrete routing problem [13]. Its implementation over Contiki on TelosB motes has also been achieved, confirming thus its great potential for developing new QoS routing protocols for WSN.

- End-to-end performance in multi-hop networks

  Probabilistic end-to-end performance guarantee may be required when dealing with real-time applications. For instance, in our ANR QUASIMODO project, we considered an intrusion detection and tracking scenario and analyzed the application requirements with respect to the network QoS. Assuming the use of the extended Kalman filter based tracking technique, we derived the tradeoff relationship between the tracking precision and the delay (from the target position and speed sampling to mobile nodes moving to cover the estimated next step area). In [5] we proposed a novel coordinative moving algorithm for autonomous mobile sensor networks to guarantee that the target can be detected in each observed step while minimizing the amount of moving sensors (so saving energy). In such kind of application context, we aim to provide methods for both network resource allocation and estimating the end-to-end delay in multi-hop WSN. Assuming IEEE802.15.4 WSN with cluster-tree routing, in [16] we addressed the problem of allocating and reconfiguring the available bandwidth using an Admission Control Manager that guarantees that the nodes respect their probabilistic bandwidth assignment when generating data traffic. It has been shown by simulation that using the proposed method, one can obtain desired probabilistic guarantee in both bandwidth and energy efficiency.

  In a more general context of meshed networks, we present an empirical support of an analytical approach, which employs a frequency domain analysis for estimating end-to-end delay in multi-hop networks. The proposed analytical results of the end-to-end delay distribution are validated through simulation and compared with queuing theory based analysis. Our results demonstrate that an analytical prediction schema is insufficient to provide an adequate estimation of the end-to-end

delay distribution function, but it requires to be combined with simulation methods for detailed link and node latency distribution [15].

6.7. Energy in Wireless Sensor Networks

Participants: Emmanuel Nataf [contact], Patrick-Olivier Kamgueu.

The energy sources of sensors in a wireless network rely mainly on batteries and are very limited in their capacity. Several research efforts are focalized on trying to limit the energy consumption in such networks. This is particularly the case in protocol design. Indeed, the communication consumes a large majority of the available energy. To be realistic and efficient, all proposed approaches need to know the energy available at any time in the systems. Unfortunately, most sensors do not provide such information because it requires additional built-in hardware that would drastically increase their cost. Over the last decade very accurate physical battery models that encompass consumption and recovery have been designed. The complexity of these models is however too high to be implemented inside simple sensors. Recent research results have shown that this integration could be possible if some approximations are integrated in the models.

We have worked on integrating such an approcimated model in the sensor operating system. This work allows the simulation of such sensors and the deployement on real devices that will be aware of their remaining energy level without requiring any additionnal costly equipment. A first implementation on simulation tool has given very promising results; sensors can access their energy level and take decision based on this estimate. Firstly, we have studied energy consumption of a sensors network collecting and routing data toward a single destination. Energy cost of the network deployment has been computed and so the network life as a whole. An other result of our work is the comparison of several common link layer access protocols and several data rate transmits [31].

6.8. 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 objective is to dynamically adapt the service exposure with respect to the threat potentiality, while maintaining a low security overhead. In the year 2012, we have pursued our work on online risk management and applied it to more distributed configurations. In that context we have defined in [14] an exposure control solution for P2PSIP networks where the registration and location servers are implemented by a distributed hash table. After having analyzed different attack scenarios, we have designed the underlying risk management architecture and modelled several dedicated countermeasures. We have evaluated the performance and scalability of our approach through extensive experiments performed with the OMNET++ simulator. We have also proposed a trust-based solution for addressing residual attacks in the RELOAD framework. This latter, complementary to our risk management approach, is a peer-to-peer signalling overlay using a central certificate enrolment server and supporting P2PSIP infrastructures. Self-signed certificates can also be used in closed networks, and connections amongst nodes can be secured using an encryption protocol such as TLS. While the RELOAD framework permits to reduce the exposure to threats, P2PSIP networks are still exposed to residual attacks related to the routing and storage activities. For instance, it is trivial for a malicious node to refuse to give the stored information, or to send false routing messages in the network. We have showed how trust mechanisms can be exploited to counter these attacks in an efficient manner. Our work on online risk management has also focused on VoIP services in the Cloud [30]. The integration of IP telephony in this environment permits the delivery and access of new resources and constitutes an important factor for its scalability. While the Cloud has recently served as a basis for security attacks targeting IP telephony, such as SIP brute force attacks from the Amazon EC2 Cloud

infrastructure, we consider that it also provides new possibilities for supporting the security of this service. We have analyzed the applicability of our online risk management approach in the Cloud, and evaluated to what extent security countermeasures may be outsourced as a service. We have mathematically defined a dedicated modelling and detailed different treatment strategies for applying countermeasures in the Cloud. Finally, we have quantified the benefits and costs of these strategies based on a set of experimental results.

6.9. Pervasive Computing

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Adrien Guenard, Yannick Presse.

Vincent Chevrier(MAIA Team), Thomas Navarrette Gutierrez (MAIA Team) and Priyadrsi Nanda (Universit of Technology, Sydney) did contribute to part of this activity.

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. In a related field, Cyber Physical Systems also are technological systems that have to be considered within a physical world and its contraints. They are complex systems where several inter-related phenomena have to be considered. In order to be studied, modeled and evaluated, we propose the use of co-simulation and multimodeling. to be Madynes is focusing on the networking aspects of such systems. We cooperate with toher the Maia team 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 is going towards both closing the loop between humans and systems, physical and computing systems, and taming the complexity, using multi-modeling (to combine the best of each domain specific model) and co-simulation (to design, develop and evaluate) as part of a global conceptual and practical toolbox. We apply this work on UAvs and energy-constrained / location aware services.

In 2012 we worked on the following research topics :

- Continuing the work on multi-modeling and co-simulation, we've participated with the MAIA team on the development of an architecture for the control of complex systems based on multi-agent simulation [32], [2], and a CPS co-simulation (next item), and continue working on the AA4MM framework (Agents and artefacts for Multiple heterogeneous Models).
- In Cyper Physical Systems, we have lead the design and implementation of the Aetournos (Airborne Embedded auTonomOUs Robust Network of Objects and Sensors) platform at Loria. The idea of AETOURNOS is to build a platform which can be at the same time a demonstrator of scientific realizations and an evaluation environment for research works of various teams of our laboratory. It is also its own research domain : building a completely autonomous and robust flock of collaborating UAVs.

  In Madynes, we focus on the CPS and their networks and applications. Those systems consist of numerous autonomous elements in sharp interaction which functioning require a tight coupling between software implementations and technical devices. The collective movements of a flock of flying communicating robots / UAVs, evolving in potentially perturbed environment constitute a good example of such a system. Indeed, if we look at the level of each of the elements playing a role into this system, a certain number of challenges and scientific questions can be studied: respect of real-time constraints of calculations for every autonomous UAV and for the communication between the robots, conception of individual, embedded, distributed or global management systems, development of self-adaptative mechanisms, conception of algorithms of collective movement etc... Furthermore, the answers to each of these questions have to finally contribute to the global functioning of the system. Applying co-simulation technique we plan to develop a hybrid "network-aware

flocking behavior" / "behavior aware routing protocol". The platform is composed of several high-grade research UAVs (Pelican quadcopters and Firefly hexacopters) and lighter models (AR.Drone quacopters). We have provided a working set of tools : multi-simulation behavior / network / physics and generic software development using ROS (Robot Operating System). The UAVs carry a set of sensor for location awareness, their own computing capabilities and several wireless networks.

This work is discribed in a position paper where a first implementation of a formation flight is detailed ([11]).

- Energy-constraint geolocalization, addressing, routing and management of wireless devices: a research collaboration with Fireflies RTLS was started in March 2009 and has ended in 2012. The initial work has been extended in a joint work with the former TRIO Team and a visiting professor from the University of Technology of Sydney. Its focus has been shiffted towards novel adressing and routing scheme minimizing a global energy-cost function in a wireless sensor network location systems [28]. We are proposing a global configuration tool for this matter in regards with given constraints (number of nodes, topology, QoS).

In 2013, we will continue working on the hybrid protocols and on the UAV platform, and apply our co-simulation work to Smart Grids.

MAESTRO Project-Team

5. New Results

5.1. Network Science

Participants: Eitan Altman, Konstantin Avrachenkov, Mahmoud El Chamie, Philippe Nain, Giovanni Neglia, Marina Sokol.

5.1.1. Epidemic models of propagation of content

E. Altman and P. Nain have studied in [96] in collaboration with A. Shwartz (Technion, Israel) and Y. Xu (Univ. Avignon/LIA) the efficiency of the existing methods for reducing availability of non-authorized copyrighted content for free download on the Internet. To model the propagation of the content, they used both branching processes as well as several epidemic models. One of the important finding is that the greatest impact of measures against unauthorized download is obtained whenever some parameter that describes the virality of the content is close to some critical value (which is computed in this work).

5.1.2. Control and game models for malware attack

In collaboration with M. H. R. Khouzani (Ohio State Univ., USA) and S. Sarkar (Univ. of Pennsylvania, USA), E. Altman has used in [31],[33], [32], optimal control theory to study malware attack in networks. The structure of optimal policies is obtained by using the Pontryagin maximum principle. In the first two references, optimal defense policies are studies in the goal of protecting the network. In the third work, the worst case behavior of the attack is identified using control theory. The authors then study in [34] the combined problem of identifying the defensive control that achieves the best performance under the worst possible malware attack. This is done through a zero-sum game context.

5.1.3. Time random walks on time varying graphs

In collaboration with D. Figueiredo (Federal Univ. of Rio de Janeiro, Brazil), B. Ribeiro and D. Towsley (both from the Univ. of Massachusetts at Amherst, USA), P. Nain has studied the behavior of a continuous time random walk (CTRW) on a stationary and ergodic time varying dynamic graph [57]. Conditions have been established under which the CTRW is a stationary and ergodic process. In general, the stationary distribution of the walker depends on the walker rate and is difficult to characterize. However, the stationary distribution has been characterized in the following cases: i) the walker rate is significantly larger or smaller than the rate in which the graph changes (time-scale separation), ii) the walker rate is proportional to the degree of the node that it resides on (coupled dynamics), and iii) the degrees of nodes belonging to the same connected component are identical (structural constraints). Examples are provided that illustrate these theoretical findings.

5.1.4. Quick detection of central nodes

In [50] K. Avrachenkov and M. Sokol, together with N. Litvak (Twente Univ., The Netherlands) and D. Towsley (Univ. of Massachusetts at Amherst, USA) propose a random walk based method to quickly find top k lists of nodes with the largest degrees in large complex networks. The authors show theoretically and by numerical experiments that for large networks the random walk method finds good quality top lists of nodes with high probability and with computational savings of orders of magnitude. They also propose stopping criteria for the random walk method which requires very little knowledge about the structure of the network.

5.1.5. Graph-based semi-supervised learning

In [48] K. Avrachenkov and M. Sokol, together with P. Gonçalves (INRIA project-team RESO) and A. Mishenin (St. Petersburg State Univ., Russia) 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 semi-supervised learning methods. The authors provide new probabilistic interpretation based on random walks and characterize the limiting behaviour of the methods. The random walk based interpretation allows one 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. The theoretical results are illustrated with two realistic datasets, characterizing different challenges: "Les Misérables" characters social network and Wikipedia hyper-link graph. It appears that the PageRank based method can classify the Wikipedia articles with very good precision and perfect recall employing only the information about the hyper-text links.

In [47] K. Avrachenkov and M. Sokol, together with P. Gonçalves (INRIA project-team RESO) and A. Legout (INRIA project-team PLANETE) apply the theoretical results of [48] to classification of content and users in BitTorrent. The general intuition behind the application of the graph based semi-supervised learning methods is that the users with similar interests download similar contents. PageRank based semi-supervised learning method was chosen as it scales well with very large volumes of data. The authors provide recommendations for the choice of parameters in the PageRank based semi-supervised learning method, and show, in particular, that it is advantageous to choose labelled points with large PageRank score.

5.1.6. Optimal weight selection in average consensus protocols

In average consensus protocols, nodes in a network perform an iterative weighted average of their estimates and those of their neighbors. The protocol converges to the average of initial estimates of all nodes found in the network. The speed of convergence of average consensus protocols depends on the weights selected on links (to neighbors). In [92] K. Avrachenkov, M. El Chamie and G. Neglia address how to select the weights in a given network in order to have a fast speed of convergence for these protocols. They approximate the problem of optimal weight selection by the minimization of the Schatten p-norm of a matrix with some constraints related to the connectivity of the underlying network. They then provide a totally distributed gradient method to solve the Schatten norm optimization problem. By tuning the parameter p in the proposed minimization, it is possible to simply trade-off the quality of the solution (i.e. the speed of convergence) for communication/computation requirements (in terms of number of messages exchanged and volume of data processed). Simulation results on random graphs and on real networks show that this approach provides very good performance already for values of p that only needs limited information exchange. The weight optimization iterative procedure can also run in parallel with the consensus protocol and form a joint consensus–optimization procedure.

5.1.7. Reducing communication overhead of average consensus protocols

The average consensus protocol converges only asymptotically to consensus and implementing a termination algorithm is challenging when nodes are not aware of some global information (e.g. the diameter of the network or the total number of nodes). In [93] K. Avrachenkov, M. El Chamie and G. Neglia propose a totally distributed algorithm for average consensus where nodes send more messages when they have large differences in their estimates, and reduce their message sending rate when the consensus is almost reached. The convergence of the system is guaranteed to be within a predefined margin from the true average and the algorithm gives a trade-off between the precision of consensus and the number of messages send in the network. The proposed algorithm is robust against nodes changing their estimates and can also be applied in dynamic networks with faulty links.

5.2. Wireless Networks

Participants: Eitan Altman, Philippe Nain, Giovanni Neglia.

5.2.1. Estimation of population sizes in sensor networks

We have been working on several problems related to the estimation of population sizes. In collaboration with D. Kumar (IBM Research Center, Hawthorne, USA) and T. Başar (Univ. of Illinois at Urbana-Champaign, USA), E. Altman develops in [73] a Wiener filter that allows to estimate the number of sensors that cover the space at some selected points. The authors take advantage of spatial correlations between the number of sensors covering different points in order to derive the filter. We note that causality is not an issue in space, in contrast to filtering at different points in time.

In collaboration with A. Ali, T. Chahed and M. K. Panda (Telecom SudParis, France), D. Fiems (Gent Univ., Belgium), and L. Sassatelli (I3S, Univ. Nice Sophia Antipolis - CNRS, France), E. Altman has used in [37] Kalman filtering theory in order to estimate the number of mobiles in a delay tolerant ad-hoc network which have a copy of a broadcasted message.

5.2.2. Cellular networks: Small cells

Analysing performance measures of cellular systems combines tools from queueing theory and stochastic processes, on one hand, and geometric considerations on the other hand. In [72], V. Kavitha (Mymo Wireless, Bangalore, India), S. Ramanath (Lekha Wireless Solutions, Bangalore, India), and E. Altman compute the time it takes to transmit a file taking into account the channel conditions which vary due to mobility of terminals. Mobility considerations play a key role in small cells since handover may occur way before the transmission of the file ends.

5.2.3. Multi scale fairness concepts for resource allocation in wireless networks

In many applications that require resources, one needs these resources within some given deadline. These impose constraints when attempting to allocate resources fairly. In [14], E. Altman, K. Avratchenkov and S. Ramanath have extended the α fairness concept by Mo and Walrand so as to include time constraints. They study the question of how to compute such constrained fair allocation, and derive some asymptotic properties of constrained fair assignment.

5.2.4. Self organization in cellular communications

Self organization is an approach to design networks so as to allow them to configure in an automatic way. This allows to reduce the complexity in systems containing thousands of mobiles and a huge number of small cells. In cellular networks, self organization can be used for deciding on time or frequency reuse according to the interference in these time and frequency slots from other cells. The impact of self organization on communications are derived in [55] and [21] by R. Combes, and Z. Altman (Orange Labs, Issy les Moulineaux), in collaboration with E. Altman.

5.2.5. Streaming over wireless

In [75], E. Altman and M. Haddad study in collaboration with T. Jiménez and R. El-Azouzi (Univ. Avignon/LIA) and S.-E. Elayoubi (Orange Labs, Issy les Moulineaux) streaming service over cellular networks. The purpose is to obtain the exact distribution of the number of buffer starvations within a sequence of N consecutive packet arrivals. This is then applied to optimize the quality of experience (QoE) of media streaming service over cellular networks by exploiting the tradeoff between the start-up delay and the starvation.

5.2.6. Wireless network security

The operation of a wireless network relies extensively on exchanging messages over a universally known channel, referred to as the control channel. The network performance can be severely degraded if a jammer launches a denial-of-service (DoS) attack on such a channel.

In [94], P. Nain, M. Krunz, H. Rahbari and M. J. Abdel Rahman (all three from Univ. of Arizona, USA) design frequency hopping (FH) algorithms that mitigate DoS attacks on the control channel of an asynchronous ad hoc network. More specifically, three FH algorithms (called NUDoS, KMDoS, and NCMDoS) are developed for establishing unicast (NUDoS) and multicast (KMDoS and NCMDoS) communications in the presence of multiple jammers. KMDoS and NCMDoS provide different tradeoffs between speed and robustness to node compromise. These algorithms are fully distributed, do not incur any additional message exchange overhead, and can work in the absence of node synchronization. Furthermore, KMDoS and NCMDoS have the attractive feature of maintaining the multicast group consistency. NUDoS exploits the grid quorum system, whereas KMDoS and NCMDoS use the uniform k-arbiter and the Chinese remainder theorem (CRT) quorum systems, respectively. Extensive simulations are used to evaluate these algorithms.

5.3. Network engineering games

Participants: Eitan Altman, Konstantin Avrachenkov, Ilaria Brunetti, Richard Combes, Julien Gaillard, Majed Haddad, Manjesh Kumar Hanawal, Alexandre Reiffers.

5.3.1. Fairness

Anti-trust laws have been introduced by many countries in the last century. This is due to the perception that free competition is better for society. This motivated H. Kameda (Univ. Tsukuba, Japan), C. Touati and A. Legrand (MESCAL, INRIA - CNRS) in cooperation with E. Altman, to define in [28] a fairness concept related to the outcome of competition, which is the Nash equilibrium concept.

5.3.2. Association problem

In [70], E. Altman and M. Haddad study in collaboration with C. Hasan and J.-M. Gorce (SOCRATE, INRIA - INSA) games related to the association problem of mobiles to an access point. It consists of deciding to which access point to connect. Here the choice is between two access points or more, where the access decisions may depend on the number of mobiles connected to each one of the access points. New results were obtained using elementary tools in congestion and crowding games.

5.3.3. Association and placement

The location of a base station has an impact on the throughuput of arriving mobiles that decide to connect to it. Given a cooperative behavior among base stations, E. Altman derives in [54] in collaboration with A. Coluccia (Univ. Salento, Italy) the equilibrium association policy and maximizes its performance by a suitable cooperative positioning of the base stations. The non-cooperative related model was studied in [16] by E. Altman, in collaboration with A. Kumar, C. Singh and R. Sundaresan (all three from IISc, Bangalore, India).

5.3.4. Power control with energy state

In [42] and [64], E. Altman, M. Haddad, J. Gaillard study with D. Fiems (Gent Univ., Belgium) a power control game over a collision channel. Each player has an energy state. When choosing a higher transmission power, the chance of a successful transmission (in the presence of other interference) increases at the cost of a larger decrease in the energy state of the battery. This dynamic game is studied when restricting to simple non-dynamic strategies that consist of choosing a given power level that is maintained during the lifetime of the battery. Surprising paradoxes were identified in the proposed Hawk and Dove game.

5.3.5. Routing games

In [65], M. Haddad, E. Altman and J. Galliard study in collaboration with D. Fiems (Gent Univ., Belgium) a sequential dynamic routing game on a line, where the decision of a user is spatio-temporal control. Each user ships its demand over time on a shared resource. Explicit expressions of the equilibrium of such systems are presented and compared to the global optimum case. The basic idea is taken from a previous paper on this subject by M. K. Hanawal (also with Univ. Avignon/LIA) and E. Altman, in collaboration with R. El-Azouzi (Univ. Avignon/LIA) and B. Prabhu (CNRS - LAAS), who show in [67] that one may transform the time dimension into a spatial component and thus obtain an equivalent standard routing game (where time plays no role) with infinitely many nodes.

5.3.6. Bayesian games in networking

We have considered several problems in networks in which decision makers have asymmetrical information. One of these is how one agent may benefit from revealing part of his information? We considered two types of hierarchical scenarios. In the first, we assume that an agent signals some information to another agent who then chooses an action based on that signal. This action determines the utility of both agents. In the second scenario, a player takes an action (such as pricing) and then the second player reacts to it. Both players' utilities depend on the actions of the two players. The action of the first player may reveal to the second player some of his private information. We use the framework of signalling game to solve the first type of problem and that of Bayesian game to solve the second. Other problems include pricing access to the Internet with partial information [52] (by I. Brunetti (Univ. Bologna, Italy), M. Haddad (Univ. Avignon/LIA) and E. Altman). In [45], M. Haddad and E. Altman, in collaboration with P. Wiecek (Wroclaw Univ. of Technology, Poland), apply Bayesian games for the association problem in which users have to decide to which access point to connect.

5.3.7. Jamming

We have been working on various models that capture different aspects of jamming (on purpose noise generation). Jamming with partial information is studied in [51] using Bayesian games, by M. Haddad (Univ. Avignon/LIA), E. Altman and S. Azad, as well as [62] and [63] by E. Altman in collaboration with A. Garnaev (St. Petersburg State Univ., Russia) and Y. Hayel (Univ. Avignon/LIA). With K. Avrachenkov, they further consider a dynamic jamming problem in [61]. In all these models the jammer creates interference to the data packets. In [29] V. Kavitha and R. El-Azouzi (Univ. Avignon/LIA), R. Sundaresan (IISc, Bangalore, India), and E. Altman study a different type of jamming game. The jammer attacks the signalling channel and not the data itself. A Bayesian game is formulated and solved there.

5.3.8. Network neutrality and collusions

Network neutrality is a key issue in the future Internet. It is related to the question of whether the access to Internet will remain a universal service or whether it would be regulated by market forces according to economic interests of those that control the Internet access. One form of network non-neutrality is when an ISP gives preferential treatment to one content provider over others. We call this "collusion" or "vertical monopoly". In collaboration with T. Jimenez and Y. Hayel (Univ. Avignon/LIA), E. Altman studies this in [71] along with "horizontal monopolies" that may occur when several ISPs merge. They introduce a new concept of "price of collusion" and identify in [44] cases in which not only consumers loose from collusions but also the colliding agents, as also seen in a different model for network non-neutrality given in [69] by M. K. Hanawal (also with Univ. Avignon/LIA) and E. Altman in collaboration with R. Sundaresan (IISc, Bangalore, India). This is related to a special kind of Braess type paradox.

5.3.9. Competition over popularity in social networks

We focus on competition of video contents for popularity. We analyze the impact of sharing, embedding, advertisement and other actions by the users for increasing the popularity and visibility. This then allowed E. Altman in [80], [38] and [95] to propose stochastic game models and to fully determine the equilibrium policy. He further proposes a dynamic game for the study of partial information and obtain the equilibrium policies and equilibrium performance. In [39], [79] the results are further extended for the wireless context.

5.3.10. Stochastic geometry methods for wireless design issues

Stochastic geometry seems to be the adequate tool in order to model correctly randomness in the location of networks elements such as the mobile terminals and the fixed base stations. Modeling the locations of both as independent spatial processes, In [66] and [25], M. K. Hanawal and E. Altman study in collaboration with F. Baccelli (TREC, INRIA - ENS) properties of Nash equilibria obtained in a multiple access game. They also derive the saddle point obtained in jamming games [68].

5.3.11. In which content to specialize

E. Altman considers in [40] the question of how should a content provider decide in which content to specialize. He shows that the problem is equivalent to the so called "Crowding" games, which allows him to prove the existence of a pure equilibrium. The conclusion is then that there is no gain by diversifying in several contents.

5.3.12. Cognitive radio

In collaboration with J. Elias (Univ. Paris Descartes-Sorbonne) and F. Martignon (LRI-Univ. Paris-Sud), E. Altman study in [56] the question of which priority level to use in a cognitive radio network: higher priority (primary user) or lower one (secondary user). The utilities are function of both the price and the quality of service. After deriving an equilibrium in this game problem, the authors study the question of how to choose prices so as to induce efficient equilibria.

5.3.13. Constrained games

In collaboration with A. Galindo-Serrano and L. Guipponi (CTTC, Spain), E. Altman studies in [60] a game theoretical problem of power control in several base stations with a coupled constraint: the interference at a given point in space should be upper bounded by some constant. The authors establish the existence of a continuum of constrained equilibria to this type of games and show that there is a unique one with some desirable scaling properties (i.e. that consitutes a normalized Nash equilibrium).

5.3.14. Dynamic coalition games

In collaboration with M. K. Panda and T. Chahed (Telecom SudParis, France), E. Altman considers the question of whether to join a multicast session or not. In contrast to many queueing problems, the congestion here is a desirable property, since the cost per user decreases as the number of users connected to the multicast session increases. In [74] the equilibrium policies are derived; these exhibit a surprising structure.

5.3.15. Evolutionary games

The relatively young theory of Evolutionary games considers a large number of interactions between pairs of randomly selected players. It is thus based on a relatively narrow scope in which the one that interacts is the player. In collaboration with Y. Hayel (Univ. Avignon/LIA) and E. V. Belmega (ETIS/ENSEA - Univ. Cergy-Pontoise - CNRS), E. Altman has been developing in [26] an alternative theory of evolutionary game in which a player consists of a group of interacting agents. This is in line with today's understanding of evolution of species (e.g. Dawkins' book "The Selfish Gene" in which the player is the gene of the species). We plan to apply this to energy dependent power control in wireless systems. We also plan to apply these in other areas such as the evolution of languages over social networks, in which some preliminary results (over Twitter) were already obtained in [81] by E. Altman and Y. Portilla (Univ. Avignon/LIA).

5.4. Green networking

Participants: Sara Alouf, Nicaise Choungmo Fofack, Delia Ciullo, Alain Jean-Marie.

5.4.1. Analysis of power saving in cellular networks with continuous connectivity

We have pursued our effort in the analysis of the continuous connectivity mode used in 4G cellular networks. Assuming Poisson traffic at each user, S. Alouf and V. Mancuso (Institute IMDEA Networks, Madrid, Spain) analyze the impact of 3GPP-defined power saving mechanisms on the performance of users with continuous connectivity. Each downlink mobile user's traffic is seen as $M/G/1$ queue, and the base station's downlink traffic as an $M/G/1\,PS$ queue with multiple classes and inhomogeneous vacations. The model is validated through packet-level simulations in [35]; its results show that consistent power saving can be achieved in the wireless access network.

The case of web traffic is investigated in [13] where the same authors, with the participation of N. Choungmo Fofack, perform in addition a sensitivity analysis to assess the impact of model parameters on the performance and cost metrics. It is found that significant power save can be achieved while users are guaranteed to experience high performance. Important outcomes of this work include the need to limit the number of active users in a cell (to less than 350 users – reasonable for 3GPP LTE, 802.16 and HSPA networks) in order to limit the web page download time, and the need to limit the web page size as large pages can dramatically decrease the energy saving. A *green attitude* would be to design web sites with short pages having few embedded objects, enabling cellular operators to use reasonable power save parameters and yet achieve a dramatic cost economy at both base station and mobile user sides, without any quality degradation.

5.4.2. Analysis of base station sleep modes in cellular networks

D. Ciullo, L. Chiaraviglio (INRIA project-team MASCOTTE), M. Ajmone Marsan (Politecnico di Torino, Italy and Institute IMDEA Networks, Spain), M. Mellia and M. Meo (Politecnico di Torino, Italy) study in [78] base station sleep modes. Putting into sleep mode some base stations in periods of low traffic improves the energy efficiency of cellular access networks. Two schemes are considered whether the sleep mode is activated once per day or multiple times per day having progressively fewer active base stations. For both schemes, the optimal base station sleep times are identified according to the traffic. Considering real traffic traces, the study reveals that significant energy saving can be achieved, the actual value strongly depending on the traffic pattern. An important result is that most of the potential savings can be attained with a single daily sleep mode, avoiding the increased complexity coming from the use of multiple sleep modes per day.

5.4.3. Analysis of sleep modes in backbone networks

The case of backbone networks is considered in [86] where L. Chiaraviglio (INRIA project-team MASCOTTE)), D. Ciullo, M. Mellia and M. Meo (Politecnico di Torino, Italy) formulate a theoretical model based on random graph theory. This model allows to estimate the potential gains achievable by adopting sleep modes in fixed networks where some devices consume energy proportionally to the handled traffic. Putting a given fraction of devices in sleep mode reduces the energy these consume but increases the energy consumed by the devices still active due to the additional load these have to handle. The model of [86] allows to predict how much energy can be saved in different scenarios. The results show that sleep modes can be successfully combined with load proportional solutions. However, if the static power consumption component is one order of magnitude less than the load proportional component, then sleep modes are no longer convenient. Thanks to random graph theory, this model gauges the impact of different properties of the network topology.

5.5. Content-oriented systems

Participants: Konstantin Avrachenkov, Nicaise Choungmo Fofack, Delia Ciullo, Philippe Nain, Giovanni Neglia, Marina Sokol.

5.5.1. Performance analysis of peer-assisted Video-on-Demand (VoD) systems

In [88] and [97], D. Ciullo, V. Martina and E. Leonardi (Politecnico di Torino, Italy), M. Garetto (Università di Torino, Italy), and G. L. Torrisi (CNR, Italy) consider peer-assisted Video-on-Demand systems. Some of the essential aspects of such systems are peer churn, bandwidth heterogeneity, and Zipf-like video popularity. The authors propose an analytical framework to tightly characterize the scaling laws for the additional bandwidth that servers must supply to guarantee perfect service, taking into account these essential aspects.

The results in [88] and [97] reveal that the catalog size and the content popularity distribution have a huge effect on the system performance. Also, users' cooperation can effectively reduce the servers' burden for a wide range of system parameters, confirming it as an attractive solution to limit the costs incurred by content providers as the system scales to large populations of users. Moreover, in [89] the same authors provide important hints for the design of efficient peer-assisted VoD systems under server capacity constraints.

5.5.2. Analysis of TTL-based cache networks

N. Choungmo Fofack, P. Nain and G. Neglia, together with D. Towsley (Univ. of Massachusetts at Amherst, USA) introduced in [87] a novel Time-To-Live (TTL) replacement policy to manage a set of documents buffering routers in information-centric networks. The TTL policy assigns a timer to each content stored in the cache and redraws the timer at each content request. In [53] they have showed that this TTL policy is more general than other policies like least frequently used (LRU), first-in-first-out (FIFO) or random (RND) as it mimics their behavior under an appropriate choice of its parameters. While exact formulas for the performance metrics of interest (hit/miss processes) are derived for a linear network and a tree network with one root cache and N leaf caches, for more general networks, an approximate solution is found with relative errors smaller than 10^{-3} and 10^{-2} for exponentially distributed and constant TTLs respectively. It is demonstrated in [53] that the TTL model can be implemented and used to optimize a multi-content cache network under realistic constraints such as the cache size limitation.

5.5.3. CCN interest routing as multi-armed bandit problem

In [49] K. Avrachenkov and P. Jacko (BCAM, Spain) consider Content Centric Network (CCN) interest forwarding problem as a Multi-Armed Bandit (MAB) problem with delays. The authors investigate the transient behaviour of the ϵ-greedy, tuned ϵ-greedy and Upper Confidence Bound (UCB) interest forwarding policies. Surprisingly, for all the three policies very short initial exploratory phase is needed. It is demonstrated that the tuned ϵ-greedy algorithm is nearly as good as the UCB algorithm, commonly reported as the best currently available algorithm. The uniform logarithmic bound for the tuned ϵ-greedy algorithm in the presence of delays is proved. In addition to its immediate application to CCN interest forwarding, the new theoretical results for MAB problem with delays represent significant theoretical advances in machine learning discipline.

In [46] K. Avrachenkov together with L. Cottatellucci and L. Maggi (both from Eurecom, France) consider the choice of CCN Access Points (APs) when CCN APs are wireless base stations. It is assumed that the slow fading channel attenuations follow an autoregressive model. In the single user case, the authors formulate this selection problem as a restless multi-armed bandit problem and propose two strategies to dynamically select a band at each time slot. The objective is to maximize the SNR in the long run. Each of these strategies is close to the optimal strategy in different regimes. In the general case with several users, the authors formulate the problem as a stochastic game with uncountable state space, where the objective is the SINR. Then the authors propose two strategies to approximate the best response policy for one user when the other users' strategy is fixed.

5.6. Advances in methodological tools

Participants: Eitan Altman, Konstantin Avrachenkov, Alain Jean-Marie, Philippe Nain.

5.6.1. Perturbation analysis

In [17] K. Avrachenkov, together with R. Burachik, J. Filar V. Gaitsgory (Univ. of South Australia, Australia), study a linear programming problem with a linear perturbation introduced through a parameter $\epsilon > 0$. The authors identify and analyze an unusual asymptotic phenomenon in such a linear program. Namely, discontinuous limiting behavior of the optimal objective function value of such a linear program may occur even when the rank of the coefficient matrix of the constraints is unchanged by the perturbation. The authors show that, under mild conditions, this phenomenon is a result of the classical Slater constraint qualification being violated at the limit and propose an iterative, constraint augmentation approach for resolving this problem.

5.6.2. Zero-sum games

In [18] K. Avrachenkov, together with L. Cottatellucci and L. Maggi (both from Eurecom, France), study zero-sum two-player stochastic games with perfect information. The authors propose two algorithms to find the uniform optimal strategies and one method to compute the optimality range of discount factors. The convergence in finite time for one algorithm is proved. In particular, the uniform optimal strategies are also optimal for the long run average criterion and, in transient games, for the undiscounted criterion as well.

5.6.3. Approximations in semi-Markov zero-sum games

In conjunction with E. Della Vecchia and S. Di Marco (both from National Univ. Rosario, Argentina), A. Jean-Marie has pursued the studies on the Rolling Horizon procedure and other approximations in stochastic control problems. Their first study on convergence conditions for average-cost MDPs has been published in [23].

They have then turned to the case of discounted semi-Markov zero-sum games. Generalizing previous contributions of the literature, they have established existence conditions and geometric convergence results when action spaces are compact and rewards possibly unbounded. The bounds they obtain hold for the Rolling Horizon procedure as well as for variants called Approximate Rolling Horizon [91]. In the same semi-Markovian context, they have also performed a sensitivity analysis of the model with respect to its parameters: cost function, discount factor, transition probabilities and state space [90].

5.6.4. Retrial queues

In [84] K. Avrachenkov and P. Nain, in collaboration with U. Yechiali (Tel Aviv Univ.), consider a retrial system with two input streams and two orbit queues. More specifically, there are two independent Poisson streams of jobs feeding a single-server service system having a limited common buffer that can hold at most one job. If a type-i job (i=1,2) finds the server busy, it is blocked and routed to a separate type-i retrial (orbit) queue that attempts to re-dispatch its jobs at its specific Poisson rate. This creates a system with three dependent queues. Such a queueing system serves as a model for two competing job streams in a carrier sensing multiple access system. The authors study the queueing system using multi-dimensional probability generating functions, and derive its necessary and sufficient stability conditions while solving a boundary value problem. Various performance measures are calculated and numerical results are presented.

5.6.5. Branching processes

In collaboration with D. Fiems (Gent Univ., Belgium), E. Altman introduces in [41] non-standard new branching processes and applies them to evaluate queueing processes. The processes are characterized by replacing the standard Algebra involved in the definition of branching processes by the max-plus algebra. Among the applications introduced are (i) polling systems with infinite server, and (2) new Cruz type bounds for systems with feedback.

Standard branching have been used in the past to study polling systems. In [30] V. Kavitha (LIA/Univ. Avignon) and E. Altman have revisited this method and applied it to spatial sensors, that receive or send data via a mobile relay or base stations. They derive conservation laws for this continuous state space polling system which allows them to compute optimal polling strategies.

D. Fiems (Gent Univ., Belgium) and E. Altman have further used in [24] semi-linear processes, which extend branching processes, to compute expected waiting times in polling systems with generally distributed walking times (the standard i.i.d. assumption is replaced with the assumption that the walking times are stationary ergodic).

In [22], the problem of parallel TCP connections is studied by O. Czerniak and U. Yechiali (Tel Aviv Univ., Israel), in collaboration with E. Altman, for a model in which, when the sum of throughputs reaches some value, there is a loss. It is assumed that the connection to suffer the loss is chosen according to a round robin policy. The expected throughputs of the connections are computed using an approach based on multitype branching processes.

MASCOTTE Project-Team

6. New Results

6.1. Network Design and Management

Participants: Gianlorenzo D'Angelo, Jean-Claude Bermond, Khoa Phan, David Coudert, Frédéric Giroire, Joanna Moulierac, Nicolas Nisse, Ronan Pardo Soares, Stéphane Pérennes, Issam Tahiri.

6.1.1. Network Design

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. 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. All-Optical Label Switching, AOLS

All-Optical Label Switching (AOLS) is a promising technology that performs packet forwarding without any optical-electrical-optical conversions, thus speeding up the forwarding. However, the cost of this technology requires limiting the number of labels needed to ensure the forwarding when routing a set of requests using GMPLS technology. In particular, this prevents the usage of label swapping techniques.

We have studied the routing problem in this context using label stacking techniques. We have formalized the problem by associating to each routing strategy a logical hypergraph, called a hypergraph layout, whose hyperarcs are dipaths of the physical graph, called tunnels in GMPLS terminology. We defined a cost function for the hypergraph layout, depending on its total length plus its total hop count. Minimizing the cost of the design of an AOLS network can then be expressed as finding a minimum cost hypergraph layout. In [24], we prove hardness results for the problem. On the other hand, we provide approximation algorithms, in particular an $O(\log n)$-approximation for symmetric directed networks. We focused on the case where the physical network is a directed path, providing a polynomial-time dynamic programming algorithm first for one source, and then for a fixed number k of sources running in time $O(n^{k+2})$.

6.1.1.2. Protocols

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, 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 [59], [60], a hybrid model of Application Layer Multicast (ALM) and Xcast6 that can work for large multicast groups. We have shown the feasibility of our model by simulation and comparison with IP multicast and NICE protocols.

Congestion control is a distributed algorithm to share network bandwidth among competing users on the Internet. In the common case, quick response time for mice traffic (http traffic) is desired when mixed with elephant traffic (ftp traffic). The current approach using loss-based with Additive Increase, Multiplicative Decrease (AIMD) is too greedy and eventually, most of the network bandwidth would be consumed by elephant traffic. As a result, it causes longer response time for mice traffic because there is no room left at the routers. MaxNet is a new TCP congestion control architecture using an explicit signal to control transmission rate at the source node. In [60] , we show that MaxNet can control well the queue length at routers and therefore the response time to http traffic is several times faster than with TCP Reno/RED.

6.1.1.3. Shared Risk Link Group

The notion of *Shared Risk Link Group*, SRLG has been introduced to capture multiple correlated failures in a network. A SRLG is a set of links that fail simultaneously if a given event (risk) occurs. In such multiple failures scenario, the problem of Diverse Routing consists in finding two SRLG-disjoint paths between a pair of nodes. We consider in [42], [66] such problem for localized failures, when all the links of a SRLG verify the star property i.e. when they are incident to the same node. We prove that in this case the problem is in general NP-complete and determine some polynomial cases.

6.1.1.4. Data Gathering in Radio Networks

We study the problem of gathering information from the nodes of a radio network into a central node. We model the network of possible transmissions by a graph and consider a binary model of interference in which two transmissions interfere if the distance in the graph from the sender of one transmission to the receiver of the other is d_I or less. A *round* is a set of non-interfering transmissions. In [25], we determine the exact number of rounds required to gather one piece of information from each node of a square two-dimensional grid into the central node. The even case uses a method based on linear programming duality to prove the lower bound, and sophisticated algorithms using the symmetry of the grid and non-shortest paths to establish the matching upper bound. We then generalize our results to hexagonal grids.

Other results on multi-interface networks were obtained outside of MASCOTTE [30], [31], [55].

6.1.2. Routing

The problem of finding and updating shortest paths in distributed networks is considered crucial in today's practical applications. In the recent past, there has been a renewed interest in designing new efficient distance-vector algorithms (e.g., the distributed Bellman-Ford method implemented in the routing information protocol, RIP) as an alternative to link-state solutions (e.g., open shortest path first, OSPF) for large-scale distributed networks such as the autonomous systems topology of the Internet.

This year, we have proposed a new loop-free distance-vector routing algorithm, called LFR (Loop Free Routing), which is able to update the shortest paths of a distributed network with n nodes in fully dynamic scenarios [47]. We compared experimentally this new algorithm with DUAL, one of the most popular loop-free distance vector algorithms which is part of CISCO's EIGRP protocol. Our experiments on CAIDA IPv4 routed /24 topology dataset show that LFR out-performs DUAL in terms of memory requirements and number of messages.

We then proposed a new technique, called Distributed Computation Pruning (DCP) [48], for reducing the total number of messages sent and the space occupancy per node of every distance-vector routing algorithm based on shortest paths. We have evaluated experimentally the combination of DCP with DUAL and with LFR. We have observed that these combinations lead to a significant gain both in terms of number of messages sent and memory requirements per node.

We have also considered routing problems arising in road networs. In particular, we have conducted a theoretical study of the graph-augmentation problem of adding shortcuts in order to speedup route planning techniques [23]. We studied the algorithmic complexity of the problem and proposed approximation algorithms for a special graph class. We have also investigated ILP-based exact approaches and show how to stochastically evaluate a given shortcut assignment on graphs that are too large to do so exactly.

6.1.2.1. Compact routing

With the constant increase of the number of routing entries in the Internet, the size of the routing tables stored at router nodes increases drastically. Routing schemes such as BGP are showing their limits in terms of update time, search time, cost of signaling, etc. and alternatives have to be proposed. In particular, compact routing schemes propose interesting trade-offs between the size of the routing tables and the quality of the routes. They also take advantage of the particular properties arising in large scale networks such as low (logarithmic) diameter and high clustering coefficient.

High clustering coefficient implies the existence of few large induced cycles. Considering this fact, we proposed in [37] a routing scheme that computes short routes in the class of k-chordal graphs, i.e., graphs with no induced cycles of length more than k. Our routing scheme achieves an additive stretch of at most $k-1$, and the routing tables are computed with a distributed algorithm which uses messages of size $O(\log n)$ and takes $O(D)$ time, where D is the diameter of the network.

We also used *cops-and-robber* games (See Section 6.2.1.2) to propose the first compact routing scheme for k-chordal graphs using routing tables, addresses and headers of size $O(\log n)$ bits and achieving an additive stretch of $O(k \log \Delta)$ [58], [57], [77]. This scheme is based on a new structural decomposition for a graph class including k-chordal graphs: we proposed a quadratic algorithm that, given a graph G and $k \geq 3$, either returns an induced cycle larger than k in G, or computes a *tree-decomposition* of G, each *bag* of which contains a dominating path with at most $k-1$ vertices. We thus proved that any k-chordal graph with maximum degree Δ has treewidth at most $(k-1)(\Delta-1)+2$, improving the $O(\Delta(\Delta-1)^{k-3})$ bound of Bodlaender and Thilikos (1997). Moreover, any graph admitting such a tree-decomposition has small hyperbolicity.

In addition, we have pursued our investigation of the kind of structural graph properties that can or cannot be deduced from local (partial) views of the network. Such knowledge is crucial for the design of routing schemes. To this end, we have exhibited a hierarchy of problems and distributed models of computation [40].

6.1.2.2. Routing models evaluation

The evaluation of new routing models asks for large-scale and intensive simulations. However, existing routing models simulators such as DRMSim are limited in terms of the number of routing table entries it can dynamically process and control on a single computer. Therefore, we have conducted a feasibility study of the extension of DRMSim so as to support the Distributed Parallel Discrete Event paradigm [46]. We have studied several distribution models and their associated communication overhead. We have in particular evaluated the expected additional time (in hours) required by a distributed simulation of BGP (border gate protocol), the current interdomain routing protocol of the Internet, compared to its sequential simulation. We show that such a distributed simulation of BGP is possible with a reasonable time overhead.

6.1.2.3. Reconfiguration

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 consists of scheduling 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 and established NP-completeness and inapproximability results.

This year, we proved the monotonicity of the *process strategy* game [78], the graph searching game modeling the routing reconfiguration problem. Then, we have investigated on the influence of physical layer impairment constraints on the reconfiguration problem [41]. Setting up a new wavelength in a fiber of a WDM network requires recalibrating the other wavelengths passing through this fiber. This induces a cost (e.g., time, energy, degradation of QoS) that depends nonlinearly on the number of wavelengths using the fiber. Therefore, the order in which requests are switched affects the total cost of the operation. We have studied the corresponding optimization problem by modeling the cost of switching a request as a non-linear function depending on the load of the links used by the new lightpath. We have proved that determining the optimal rerouting order is NP-complete for a 2-nodes network, established general lower and upper bounds, identified classes of instances where the problem can be solved in polynomial time, and proposed a heuristic algorithm.

6.1.3. Energy efficiency

Recently, energy-aware routing has gained increasing popularity in the networking research community. The idea is that traffic demands are aggregated over a subset of the network links, allowing other links to be turned off to save energy. We develop several methods to improve routing protocols for backbone, wireless and content delivery networks. 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,... The goal thus is to find a routing that minimizes the (weighted) number of active network elements used when routing. In [62], we exhibit that the power consumption can be reduced of approximately 33 MWh for a medium-sized backbone network.

In [54], we propose GreenRE - a new energy-aware routing model with the support of the new technique of data redundancy elimination (RE). Based on real experiments on Orange Labs platform and on simulations on several network topologies, we show that GreenRE can gain further 30% energy savings in comparison with the traditional energy-aware routing model.

One of the new challenges facing research in wireless networks is the design of algorithms and protocols that are energy aware. In [33], we use for the first time the evolving graph combinatorial model as a tool to prove an NP-Completeness result, namely that computing a Minimum Spanning Tree of a planar network in the presence of mobility is actually NP-Complete.

Recently, there is a trend to introduce content caches as an inherent capacity of network equipment, with the objective of improving the efficiency of content distribution and reducing network congestion. In [63], we study the impact of using in-network caches and CDN cooperation on an energy-efficient routing: up to 23% of power can be saved in the backbone this way.

In [32], we study the energy efficiency of the networking part of data centers, accounting for between 10-20% of the total power consumption. We proposed a novel approach, called VMPlanner, for power reduction in the virtualization-based data centers. The idea of VMPlanner is to optimize both virtual machine placement and traffic flow routing so as to turn off as many unneeded network elements as possible for power saving.

Finally, in [56], [38], we summarize the main research results of the last years for energy efficiency for backbone, wireless, cellular and content distribution networks and highlight the main challenges of the field. Results are given for two operator networks, considering power and traffic forecasts for 2020.

6.2. Graph Theory

Participants: Julio Araújo, Jean-Claude Bermond, Frédéric Giroire, Frédéric Havet, František Kardoš, Ana Karolinna Maia, Remigiusz Modrzejewski, Leonardo Sampaio, Michel Syska.

6.2.1. Algorithms in graphs

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.

6.2.1.1. 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.

Degree Constraint Subgraphs. A natural question in current social networks is *How do one find a small community (subgraph) in which anyone as at least d friends (neighbors)?* This problem can be modelled as degree-constrained subgraph problems where the objective is to find an optimal weighted subgraph, subject to certain degree constraints (in which each node has degree at most d), in a weighted graph. When $d = 2$, the problem is easy to solve since one simply needs to compute the girth of the graph. In [16], we proved that the problem is not in Apx for any $d \geq 3$. The proof is obtained by a reduction from Vertex Cover in regular graphs, followed by the use of an error amplification technique. On the positive side, we give an $\frac{n}{\log n}$-approximation

algorithm for the class of graphs excluding a fixed graph H as a minor (including planar or bounded genus graphs), using dynamic programming.

Hyperbolicity in Large graphs. Hyperbolicity is a geometric notion that measure how the various shortest paths connecting two vertices can diverge in a graph. Knowing its value provides information on the geometry of the network, moreover it has practical implications for shortest path routing. Hyperbolicity can be computed in polynomial time algorithm ($\Theta(n^4)$). This is far from being practical for large graphs. So, in [69] we proposed a scalable algorithm for this problem. We also led some computational experiments of our algorithms on large-scale graphs.

Hull Number of graphs. In [64], we study the (geodesic) hull number of graphs. For any two vertices $u, v \in V$ of a connected undirected graph $G = (V, E)$, the closed interval $I[u, v]$ of u and v is the set of vertices that belong to some shortest (u, v)-path. For any $S \subseteq V$, let $I[S] = \bigcup_{u,v \in S} I[u, v]$. A subset $S \subseteq V$ is (geodesically) convex if $I[S] = S$. Given a subset $S \subseteq V$, the convex hull $I_h[S]$ of S is the smallest convex set that contains S. We say that S is a hull set of G if $I_h[S] = V$. The size of a minimum hull set of G is the hull number of G, denoted by $hn(G)$. First, we show a polynomial-time algorithm to compute the hull number of any P_5-free triangle-free graph. Then, we present four reduction rules based on vertices with the same neighborhood. We use these reduction rules to propose a fixed parameter tractable algorithm to compute the hull number of any graph G, where the parameter can be the size of a vertex cover of G or, more generally, its neighborhood diversity, and we also use these reductions to characterize the hull number of the lexicographic product of any two graphs. More on the hull number of graphs may be found in Araujo's thesis [13].

6.2.1.2. 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.

Web Caching & the surfer Game. A surprising application of some variant of pursuit-evasion games (namely Cops and Robber games) is the problem for a web-browser to download documents in advance while an internaut is surfing on the Web. In [53], [52], 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.

Connected Graph Searching. Another variant of pursuit-evasion games is graph searching which is mainly related to graph decompositions. For instance, the minimum number of searchers needed to capture an invisble fugitive in a graph is equal to its pathwidth plus one. In [21], we investigated the connected variant of this game. A strategy is called connected if the clear part (the part where the fugitive cannot stand) always induces a connected subgraph. The main motivation for studying connected graph searching is the design of distributed protocols allowing searchers to compute a capture strategy (see also Section 6.2.1.3). [21] gathers most of the results of the last decade concerning connected graph searching, mainly focussing on the cost of connectivity in terms of number of searchers.

6.2.1.3. 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 promissing approach (somehow related to Cops and Robber in Section 6.2.1.2) to adress some models of distributed computing. We considered distributed or even self-stabilizing algoritms for gathering and graph searching problems.

Graph Searching and Routing Reconfiguration. In [29], we developed a generic distributed algorithm for computing and updating various parameters on trees including the process number (see Section 6.1.2.3), and other related graph searching parameters (see Section 6.2.1.2). We also proposed an incremental version of the algorithm allowing to update these parameters after addition or deletion of any tree edge.

Robots in anonymous networks. Motivated by the understanding of the limits of distributed computing, 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 various problems such as graph exploration, graph searching and gathering in various graph classes. 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. More precisely, we design algorithms for the gathering in rings [51], [70], grid [50] and trees [61]. We also proposed a general approach [71] to solve the three problems in rings. Finally, in [67], [44], [43], algorithms are designed to solve the graph searching problem in trees.

6.2.2. Structural graph theory

6.2.2.1. Directed graphs

Graph theory can be roughly partitioned into two branches: the areas of undirected graphs and directed graphs (digraphs). Even though both areas have numerous important applications, for various reasons, undirected graphs have been studied much more extensively than directed graphs. One of the reasons is that many problems for digraphs are much more difficult than their analogues for directed graphs. For example, one of the cornerstones of modern (undirected) graph theory is Minor Theory of Robertson and Seymour. Unfortunately, we cannot expect an equivalent for directed graphs. Minor Theory implies in particular that, for any fixed H, detecting a subdivision of H in an input graph G can be performed in polynomial time by the Robertson and Seymour linkage algorithm. In contrast, the analogous subdivision problem for digraph can be either polynomial-time solvable or NP-complete, depending on the fixed digraph H. In [65], we give a number of examples of polynomial instances, several NP-completeness proofs as well as a number of conjectures and open problems. We also investigated the related problem in which we want to detect an *induced* subdivision of H. Already, for undirected graphs the complexity of this problem depends on H. In [20], we show that for digraphs the complexity of this problem depends on H and on whether the input digraph G 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.

In a directed graph, a *star* is an arborescence with at least one arc, in which the root dominates all the other vertices. A *galaxy* is a vertex-disjoint union of stars. In [34], we consider the Spanning Galaxy problem of deciding whether a digraph D has a spanning galaxy or not. We show that although this problem is NP-complete (even when restricted to acyclic digraphs), it becomes polynomial-time solvable when restricted to strong digraphs. In fact, we prove that restricted to this class, the Spanning Galaxy problem is equivalent to the problem of deciding if a strong digraph has a strong digraph with an even number of vertices. We then show a polynomial-time algorithm to solve this problem. We also consider some parameterized versions of the Spanning Galaxy problem. Finally, we improve some results concerning the notion of *directed star arboricity* of a digraph D, denoted $dst(D)$, which is the minimum number of galaxies needed to cover all the arcs of D. We show in particular that $dst(D) \leq \Delta(D) + 1$ for every digraph D and that $dst(D) \leq \Delta(D)$ for every acyclic digraph D.

Hypergraphs are a generalization of graphs, in which every edge is incident to a set of vertices of any size (not necessarily 2). Like for digraphs, a lot fewer is known about them than about graphs. The two notions of eulerian and hamoltinians cycles have been extensively studied for graphs and digraphs. The analogue notion of eulerian cycle in a hypergraph was only introduced in 2010 by Lonc and Naroski. In [72], we introduce the notions of eulerian and hamiltonian circuits in directed hypergraphs. We show that both associated decision problems are NP-complete. Some necessary conditions for a dihypergraph to be have an eulerian circuit are presented. We exhibit some families of hypergraphs for which those are sufficient conditions. We also generalize a part of the properties of eulerian digraphs to the uniform and regular directed hypergraphs. Finally, we show that the de Bruijn and Kautz dihypergraphs are eulerian and hamiltonian in most cases.

6.2.2.2. Graph colouring

We mainly study graph colouring problems that model channel assignment problems.

A well-known such general problem is the following: we are given a graph G, whose vertices correspond to transmitters, together with an edge-weghting w. The weight of an edge corresponds to the minimum separation between the channels on its endvertices to avoid interferences. (If there is no edge, no separation is required, the transmitters do not interfere.) We need to assign positive integers (corresponding to channels) to the vertices so that for every edge e the channels assigned to its endvertices differ by at least $w(e)$. The goal is to minimize the largest integer used, which corresponds to minimizing the *span* of the used bandwidth.

We mainly studied a particular, yet quite general, case, called *backbone colouring*, in which there are only two levels of interference. So we are given a graph G and a subgraph H, called *the backone*. Two adjacent vertices in H must get integers at least q apart, while adjacent vertices in G must get integers at distance at least 1. The minimum span is this case is called the q-backbone chromatic number and is denoted $BBC_q(G,H)$. Backbone forests in planar graphs are of particular interests. In [74], we give a series of NP-hardness results as well as upper bounds for $BBC_q(G,H)$, depending on the type of the forest (matching, galaxy, spanning tree). Eventually, we discuss a circular version of the problem. In [73], we also consider a list version of the problem in which every vertex must be assigned an integer in its own list of available ones. We provide bounds using the Combinatorial Nullstellensatz for the list version on the channel assignment problem. Through this result and through structural approaches, we obtain good upper bounds for forests and matching backbone in planar graphs. In [68], we give an evidence to a conjecture of Broersma et al. stating that $BBC_2(G,T) \leq 6$, for every planar graph G and spanning tree T. We prove this conjecture in the particular case when T has diameter at most 4.

Another meaningful and very well-studied particular case of backbone colouring is $L(p,1)$-labelling, which is p-backbone colouring of (G^2,G), where G^2 is the square of G (the graph with same vertex set as G, in which two vertices are adjacents if they are at distance at most 2 in G). Griggs and Yeh conjecture in 1992, that for every graph with maximum degree $\Delta \geq 2$, $BBC_2(G^2,G) \leq \Delta^2+1$. In [36], we prove this conjecture when Δ is large. In fact, we prove a more general statement. We prove for any q and sufficiently large Δ, if $\Delta(H) \leq \Delta^2$ and $\Delta(G) \leq \Delta$, then $BBC_q(H,G) \leq \Delta^2+1$. Our result also holds for the list version.

In [17], we studied another colouring problem motivated by a practical frequency assignment problem and, up to our best knowledge, new. In wireless networks, a node interferes with other nodes, the level of interference depending on numerous parameters: distance between the nodes, geographical topography, obstacles,... We model this with a weighted graph (G,w) where the weight function w on the edges of G represents the noise (interference) between the two end-vertices. The total interference in a node is the sum of all the noises of the nodes emitting on the same frequency. A *weighted t-improper k-colouring* of (G,w) is a k-colouring of the nodes of G (assignment of k frequencies) such that the interference at each node does not exceed the threshold t. We consider the Weighted Improper Colouring problem which consists in determining the *weighted t-improper chromatic number* defined as the minimum integer k such that (G,w) admits a weighted t-improper k-colouring. We also consider the dual problem, denoted the Threshold Improper Colouring problem, where, given a number k of colours, we want to determine the minimum real t such that (G,w) admits a weighted t-improper k-colouring. We show that both problems are NP-hard and present general upper bounds for both problems; in particular we show a generalisation of Lovász's Theorem for the weighted t-improper chromatic number. 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 neighbours and a noise of intensity 1/2 for the nodes at distance two. We derive the weighted t-improper chromatic number for all values of t.

Since some of the channel assignment problems must be done on-line, we are interested in some on-line graph colouring heuristics. We only studied such heuristics for the classical proper colouring. The easiest one, and the most widespread one, is the greedy algorithm, which colours the vertices one after another, giving to each vertex the smallest possible positive integer that is not already used by one of its neighbours. The *Grundy number* of a graph G is the largest number of colours used by any execution of the greedy algorithm to colour G. In [27], we give new bounds on the Grundy number of the different product of two graphs. The problem of determining the Grundy number of G is polynomial-time solvable if G is a P_4-free graph and NP-hard if G is a P_5-free graph. In [19], we define a new class of graphs, the *fat-extended P_4-laden graphs*, and we show a polynomial-time algorithm to determine the Grundy number of any graph in this class. Our class intersects the

class of P_5-free graphs and strictly contains the class of P_4-free graphs. More precisely, our result implies that the Grundy number can be computed in polynomial time for any graph of the following classes: P_4-reducible, extended P_4-reducible, P_4-sparse, extended P_4-sparse, P_4-extendible, P_4-lite, P_4-tidy, P_4-laden and extended P_4-laden, which are all strictly contained in the fat-extended P_4-laden class.

A colouring c of a graph $G = (V, E)$ is a *b-colouring* if in every colour class there is a vertex whose neighborhood intersects every other colour classes. Such a colouring appears, when we try to optimize on-line the colouring of a graph, by changing the colour of all vertices of a colour class if it is possible. The *b-chromatic number* of G, denoted $\chi_b(G)$, is the greatest integer k such that G admits a b-coloring with k colours. A graph G is *tight* if it has exactly $m(G)$ vertices of degree $m(G) - 1$, where $m(G)$ is the largest integer m such that G has at least m vertices of degree at least $m - 1$. Determining the b-chromatic number of a tight graph had been shown to be NP-hard even for a connected bipartite graph. In [35], we show that it is also NP-hard for a tight chordal graph, and that the b-chromatic number of a split graph can be computed in polynomial time. Then we define the b-closure and the partial b-closure of a tight graph, and use these concepts to give a characterization of tight graphs whose b-chromatic number is equal to $m(G)$. This characterization is used to develop polynomial-time algorithms for deciding whether $\chi_b(G) = m(G)$, for tight graphs that are complement of bipartite graphs, P_4-sparse and block graphs. We generalize the concept of pivoted tree introduced by Irving and Manlove and show its relation with the b-chromatic number of tight graphs.

Many more results on greedy colourings and b-colourings have been proved in Sampaio's thesis [14].

We studied other variations of graph colouring. 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 (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.

For a connected graph G of order at least 3 and a k-labelling $c : E(G) \to \{1, 2, \cdots, k\}$ of the edges of G, the *code* of a vertex v of G is the ordered k-tuple $(n_1, \cdots, n_k)$, where n_i is the number of edges incident with v that are labelled i. The k-labelling c is detectable if every two adjacent vertices of G have distinct codes. The minimum positive integer k for which G has a detectable k-labelling is the *detection number* of G. In [76], we show that it is NP-complete to decide if the detection number of a cubic graph is 2. We also show that the detection number of every bipartite graph of minimum degree at least 3 is at most 2. Finally, we give some sufficient condition for a cubic graph to have detection number 3.

PLANETE Project-Team

6. New Results

6.1. Towards Data-Centric Networking

Participants: Chadi Barakat, Damien Saucez, Jonathan Detchart, Mohamed Ali Kaafar, Ferdaouss Mattoussi, Marc Mendonca, Xuan-Nam Nguyen, Vincent Roca, Thierry Turletti.

- **DTN**

  Delay Tolerant Networks (DTNs) stand for 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 heterogenous 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. Along the main ideas of HBSD, we worked on a content optimal-delivery algorithm, CODA, that distributes content to multiple receivers over a DTN. CODA assigns a utility to each content item published in the network; this value gauges the contribution of a single content replica to the network's overall delivery-rate. CODA performs buffer management by first calculating the delivery-rate utility of each cached content-replica and then discarding the least-useful item. When an application requests content, the node supporting the application will look for the content in its cache. It will immediately deliver it to the application if the content is stored in memory. In case the request cannot be satisfied immediately, the node will store the pending request in a table. When the node meets another device, it will send the list of all pending requests to its peer; the peer device will try to satisfy this list by sending the requester all the matching content stored in its own buffer. A meeting between a pair of devices might not last long enough for all requested content to be sent. We address this problem by sequencing transmissions of data in order of decreasing delivery-rate utility. A content item with few replicas in the network has a high delivery rate utility; these items must be transmitted first to avoid degrading the content delivery-rate metric. The node delivers the requested content to the application

as soon as it receives it in its buffer. We implement CODA over the CCNx protocol, which provides the basic tools for requesting, storing, and forwarding content. Detailed information on CODA and the implementation work carried out herein can be found in [76].

- **Naming and Routing in Content Centric Networks**

  Content distribution prevails in todays 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 proposed 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 encaspulated 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. See [100] for more details.

- **On the fairness of CCN**

  Content-centric networking (CCN) is a new paradigm to better handle contents in the future Internet. Under the assumption that CCN networks will deploy a similar congestion control mechanism than in today's TCP/IP (i.e., AIMD), we built an analytical model of the bandwidth sharing in CCN based on the "square-root formula of TCP". With this model we can compare CCN download performance to what users get today. We consider different factors such as the way CCN routers are deployed, the popularity of contents, or the capacity of links and observe that when AIMD is used in a CCN network less popular content throughput is massively penalised whilst the individual gain for popular content is negligible. Finally, the main advantage of using CCN is the decrease of load at the server side. Our observations advocate the necessity to clearly define the notion of fairness in CCN and to design a proper congestion control to avoid less popular contents to become hardly accessible in tomorrow's Internet.

  Our results [75] clearly point to a fairness issue if AIMD is used with CCN. Indeed, combining blindly AIMD and CCN can severely worsen the download throughput of less popular contents with respect to the today's Internet due to subtle interactions with in-network caching strategies. The way cache memories are distributed within chain topologies has been investigated too, showing that for small and heterogeneous cache spaces, placing the biggest caches close to clients improves performance due to a smaller RTT on average. On the other hand, CCN can significantly reduce the load at the server side independently of the cache allocation strategy. Our findings advocate the urge of clearly defining the notion of fairness in CCN and designing congestion control algorithms able to limit the unfairness observed between contents of different popularities. The work is currently used within the IRTF ICNRG research group in order to motivate and define an appropriate congestion control mechanism for information centric networks like CCN. Moreover, we are currently validating the analytical results with an implementation of CCN where we can evaluate how much our model

deviates from the reality when contents are of various size or small. The implementation will also be a support to test different congestion control mechanism.

- **CCN to enable profitable collaborative OTT services**

  The ubiquity of broadband Internet and the proliferation of connected devices like laptops, tablets, or TV result in a high demand of multimedia content such as high definition video on demand (VOD) for which the Internet has been poorly designed with the Internet Protocol (IP). Information-Centric Networking and more precisely Content Centric Networking (CCN) overtake the limitation of IP by considering content as the essential element of the network instead of the topology. CCN and its content caching capabilities is particularly adapted to Over-The-Top (OTT) services like Netflix, Hulu, Xbox Live, or YouTube that distribute high-definition multimedia content to millions of consumers, independently of their location. However, bringing content as the most important component of the network implies fundamental changes in the Internet and the transition to a fully CCN Internet might take a long time. Despite this transition period where CCN and IP will co-exist, we have shown that OTT service providers and consumers have strong incentives for migrating to CCN. We also propose a transition mechanism based on the Locator/Identifier Separation Protocol (LISP) [28] that allows the provider to track the demands from its consumers even though they do not download the contents from another consumers instead of the producer itself.

  CCN, compared to IP, provides better security and performance. This last point is very interesting for OTT service providers that deliver multimedia content where performance is a key factor for the adoption of the service by consumers. With CCN, the content can be retrieved from the caches in the different CCN islands, instead of always being delivered by the content publisher. As a result, content retrieval is faster for the consumer and the operational cost of the publisher is reduced. Moreover, as the content is cached by the consumers and because the consumer can provide the content to other consumers, the overall performance increases with the number of consumers instead of decreasing as it is the case in IP today where the content is delivered by the hosting server. This property is particularly interesting because it dampens the effect of flash crowds which are normally very costly for OTT service providers as they have to provision their servers and networks to support them. Using CCN with caching at the consumers has then a direct impact on the profit earned by the OTT service provider as its costs are reduced. However, to benefit from the caching capabilities of consumers, the producer must propose real incentives to its consumers to *collaborate* and cache the content. To understand how incentives can be provided, it is necessary to remember that content in OTT is provided either freely to the consumer or in exchange of a fee. When the content is provided freely, the incomes for the publisher are ensured by advertisements dispersed in the content (e.g., banner, commercial interruptions...). A consumer has incentives to collaborate with the system if it receives some sort of discount, expressed in advertisement reduction or fee reduction. On the one hand, the discount has a cost for the publisher as its revenues will be reduced. On the other hand, the collaboration from its consumers reduces its operational costs. Hence, the publisher must determine the optimal discount, such that it maximises its profit. The situation for the consumer is the exact opposite: its costs are increasing because it is providing content to other consumers but its revenues also increase as it receives a discount on its expenses. We have determined the conditions to respect when deploying OTT with loosely collaborative consumers [99]. We currently refine the results using game theory.

- **Software-Defined Networking in Heterogeneous Networked Environments**

Software-Defined Networking (SDN) has been proposed as a way to facilitate network evolution by allowing networks and their infrastructure to be programmable. In the context of the COMMUNITY associated team with University of California Santa Cruz (see URL http://inrg.cse.ucsc.edu/community/) , we are studying the potential of SDN to facilitate the deployment and management of new architectures and services in heterogeneous environments. In particular, we focus on the fundamental issues related to enabling SDN in infrastructure-less/decentralized networked environments and we use OpenFlow as our target SDN platform. Our plan is to develop a hybrid SDN framework that strikes a balance between a completely decentralized approach like Active Networking and a centralized one such as OpenFlow~[58].

We are also currently evaluating the efficiency of SDN for optimizing caching in content-centric networks. CCN advocates in-network caching, i.e., to cache contents on the path from content providers to requesters. Akthough this on-path caching achieves good overall performance, we have shown that this strategy is far from being the optimal inside a domain. On this purpose, we proposed the notion of off-path caching by allowing deflection of the most popular traffic off the optimal path towards off-path caches available across the domain[100]. Off-path caching improves the global hit ratio and permits to reduce the peering links' bandwidth usage. We are now investigating whether SDN functionalities can be used to implement this optimal caching technique, in particular to identify of the most popular contents, and to configure deflection mechanisms within routers~[94].

- **Application-Level Forward Error Correction Codes (AL-FEC) and their Applications to Broadcast/Multicast Systems**

  With the advent of broadcast/multicast systems (e.g., 3GPP MBMS services), 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 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, with a commercial service that started in April 2012. This is the first adoption of these codes in an international standard. These codes, along with our FLUTE/ALC software, are now part of the server and terminal protocol stack: http://www.rapidtvnews.com/index.php/2012041721327/ntt-data-mse-and-expways-joint-solution-powers-japanese-mobile-tv-service.html.

  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 recent Internet-Drafts [91][92] define 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 for LDPC-Staircase codes. One of them is distributed in open-source, as part of our OpenFEC project (http://openfec.org), 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, is commercialized through an industrial partner, Expway.

  Since May 2012, along with the Expway French company, we are proposing the Reed-Solomon + LDPC-Staircase codes for the 3GPP-eMBMS call for technology, as a candidate for next generation AL-FEC codes for multimedia services. We have shown that these codes offer very good erasure

recovery capabilities, in line with 3GPP requirements, and extremely high decoding speeds, usually significantly faster than that of the other proposals. The final decision is expected for end of January 2013. In any case we have once again showed that these codes provide very good performance, often ahead of the competitors, and an excellent balance between several technical and non technical criteria.

Finally our activities in the context of the PhD of F. Mattoussi include 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 (IETF RFC 5170), while keeping backward compatibility (i.e. LDPC-Staircase "codewords" can be decoded with a GPLDPC-Staircase codec). More information is available in [56][57][55].

- **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 remaing) 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 independant 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 [91] 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 [92], [91].

  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 [64]. 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 RFC6363.

  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: [85] for Reed-Solomon and [88] for LDPC-Staircase. Both documents are close to being published as RFCs.

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. 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 [95] 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 [82]. 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.

- **Security of the Broadcast/Multicast Systems**

  Sooner or later, broadcasting systems will require security services. This is all the more true as heterogeneous broadcasting technologies are used, some of them being by nature open, such as WiFi networks. Therefore, one of the key security services is the authentication of the packet origin and the packet integrity check. To that purpose, 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 and the standard is now published as IETF RFC 6584 [98].

- **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 2011-2012 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 [51].

6.2. Network Security and Privacy

Participants: Claude Castelluccia, Gergely Acs, Mathieu Cunche, Daniele Perito, Lukasz Olejnik, Mohamed Ali Kaafar, Abdelberi Chaabane, Cédric Lauradoux, Minh-Dung Tran.

- *Private Big Data Publication* Public datasets are used in a variety of applications spanning from genome and web usage analysis to location-based and recommendation systems. Publishing such datasets is important since they can help us analyzing and understanding interesting patterns. For example, mobility trajectories have become widely collected in recent years and have opened the possibility to improve our understanding of large-scale social networks by investigating how people exchange information, interact, and develop social interactions. With billion of handsets in use worldwide, the quantity of mobility data is gigantic. When aggregated, they can help understand complex processes, such as the spread of viruses, and build better transportation systems, prevent traffic congestion. While the benefits provided by these datasets are indisputable, they unfortunately pose a considerable threat to individual privacy. In fact, mobility trajectories might be used by a malicious attacker to discover potential sensitive information about a user, such as his habits, religion or relationships. Because privacy is so important to people, companies and researchers are reluctant to publish datasets by fear of being held responsible for potential privacy breaches. As a result, only very few of them are actually released and available. This limits our ability to analyze such data to derive information that could benefit the general public. Here follows some recent results of our activities in this domain.

  Privacy-Preserving Sequential Data Publication [41]: Sequential data is being increasingly used in a variety of applications, spanning from genome and web usage analysis to location-based recommendation systems. Publishing sequential data is of vital importance to the advancement of these applications since they can enable researchers to analyze and understand interesting sequential patterns. However, as shown by the re-identification attacks on the AOL and Netflix datasets, releasing sequential data may pose considerable threats to individual privacy. Recent research has indicated the failure of existing sanitization techniques to provide claimed privacy guarantees. It is therefore urgent to respond to this failure by developing new schemes with provable privacy guarantees. Differential privacy is one of the only models that can be used to provide such guarantees. Due to the inherent sequentiality and high-dimensionality, it is challenging to apply differential privacy to sequential data. In this work, we address this challenge by employing a variable-length n-gram model, which extracts the essential information of a sequential database in terms of a set of variable-length n-grams. Our approach makes use of a carefully designed exploration tree structure and a set of novel techniques based on the Markov assumption in order to lower the magnitude of added noise. The published n-grams are useful for many purposes. Furthermore, we develop a solution for generating a synthetic database, which enables a wider spectrum of data analysis tasks. Extensive experiments on real-life datasets demonstrate that our approach substantially outperforms the state-of-the-art techniques.

  Private Histogram Publishing [33]:

  Differential privacy can be used to release different types of data, and, in particular, histograms, which provide useful summaries of a dataset. Several differentially private histogram releasing schemes have been proposed recently. However, most of them directly add noise to the histogram counts, resulting in undesirable accuracy. In this work, we propose two sanitization techniques that exploit the inherent redundancy of real-life datasets in order to boost the accuracy of histograms. They lossily compress the data and sanitize the compressed data. Our first scheme is an optimization of the Fourier Perturbation Algorithm (FPA) presented in [13]. It improves the accuracy of the initial FPA by a factor of 10. The other scheme relies on clustering and exploits the redundancy between bins. Our extensive experimental evaluation over various real-life and synthetic datasets demonstrates that our techniques preserve very accurate distributions and considerably improve the accuracy of range queries over attributed histograms.

- *Privacy Issues on the Internet* Internet users are being increasingly tracked and profiled. Companies utilize profiling to provide customized, i.e. personalized services to their customers, and hence increase revenues.

  Privacy issues of Targeted Advertising [37]: Behavioral advertising takes advantage from profiles of users' interests, characteristics (such as gender, age and ethnicity) and purchasing activities. For example, advertising or publishing companies use behavioral targeting to display advertisements that closely reflect users' interests (e.g. 'sports enthusiasts'). Typically, these interests are inferred from users' web browsing activities, which in turn allows building of users' profiles. It can be argued that customization resulting from profiling is also beneficial to users who receive useful information and relevant online ads in line with their interests. However, behavioral targeting is often perceived as a threat to privacy mainly because it heavily relies on users' personal information, collected by only a few companies. In this work, we show that behavioral advertising poses an additional privacy threat because targeted ads expose users' private data to any entity that has access to a small portion of these ads. More specifically, we show that an adversary who has access to a user's targeted ads can retrieve a large part of his interest profile. This constitutes a privacy breach because interest profiles often contain private and sensitive information.

  On the Uniqueness of Web Browsing History Patterns [60]: We present the results of the first large-scale study of the uniqueness of Web browsing histories, gathered from a total of $368,284$ Internet users who visited a history detection demonstration website. Our results show that for a majority of users (69%), the browsing history is unique and that users for whom we could detect at least 4 visited websites were uniquely identified by their histories in 97% of cases. We observe a high rate of stability in browser history fingerprints: for repeat visitors, 80% of fingerprints are identical over time, and differing ones were strongly correlated with original history contents, indicating static browsing preferences. We report a striking result that it is enough to test for a small number of pages in order to both enumerate users' interests and perform an efficient and unique behavioral fingerprint; we show that testing 50 web pages is enough to fingerprint 42% of users in our database, increasing to 70% with 500 web pages. Finally, we show that indirect history data, such as information about *categories* of visited websites can also be effective in fingerprinting users, and that similar fingerprinting can be performed by common script providers such as Google or Facebook.
- **Adaptive Password-Strength Meters from Markov Models [38]**

  Passwords are a traditional and widespread method of authentication, both on the Internet and off-line. Passwords are portable, easy to understand for laypersons, and easy to implement for the operator. Thus, password-based authentication is likely to stay for the foreseeable future.

  To ensure an acceptable level of security of user-chosen passwords, sites often use mechanisms to test the strength of a password (often called *pro-active password checkers*) and then reject weak passwords. Hopefully this ensures that passwords are reasonably strong on average and makes guessing passwords infeasible or at least too expensive for the adversary. Commonly used password checkers rely on rules such as requiring a number and a special character to be used. However, as we will show and also has been observed in previous work, the accuracy of such password checkers is low, which means that often insecure passwords are accepted and secure passwords are rejected. This adversely affects both security and usability.

  In this work, we propose to use password strength meters based on Markov-models, which estimate the true strength of a password more accurately than rule-based strength meters. Roughly speaking, the Markov-model estimates the strength of a password by estimating the probability of the n-grams that compose said password. Best results can be obtained when the Markov-models are trained on the actual password database. We show, in this work, how to do so without sacrificing the security of the password database, even when the n-gram database is leaked.

  We show how to build secure adaptive password strength meters, where security should hold even when the n-gram database leaks. This is similar to traditional password databases, where one tries

to minimize the effects of a database breach by hashing and salting the stored passwords. This is not a trivial task. One potential problem is that, particularly strong passwords, can be leaked entirely by an n-gram database (without noise added).

- **Fast Zero-Knowledge Authentication [47]** We explore new area/throughput trade-offs for the Girault, Poupard and Stern authentication protocol (GPS). This authentication protocol was selected in the NESSIE competition and is even part of the standard ISO/IEC 9798. The originality of our work comes from the fact that we exploit a fixed key to increase the throughput. It leads us to implement GPS using the Chapman constant multiplier. This parallel implementation is 40 times faster but 10 times bigger than the reference serial one. We propose to serialize this multiplier to reduce its area at the cost of lower throughput. Our hybrid Chapman's multiplier is 8 times faster but only twice bigger than the reference. Results presented here allow designers to adapt the performance of GPS authentication to their hardware resources. The complete GPS prover side is also integrated in the network stack of the PowWow sensor which contains an Actel IGLOO AGL250 FPGA as a proof of concept.
- **Energy Efficient Authentication Strategies for Network Coding [26]**

  Recent advances in information theory and networking, e.g. aggregation, network coding or rateless codes, have significantly modified data dissemination in wireless networks. These new paradigms create new threats for security such as pollution attacks and denial of services (DoS). These attacks exploit the difficulty to authenticate data in such contexts. The particular case of xor network coding is considered herein. We investigate different strategies based on message authentication codes algorithms (MACs) to thwart these attacks. Yet, classical MAC designs are not compatible with the linear combination of network coding. Fortunately, MACs based on universal hash functions (UHFs) match nicely the needs of network coding: some of these functions are linear $h(x_1 \oplus x_2) = h(x_1) \oplus h(x_2)$. To demonstrate their efficiency, we consider the case of wireless sensor networks (WSNs). Although these functions can drastically reduce the energy consumption of authentication (up to 68% gain over the classical designs is observed), they increase the threat of DoS. Indeed, an adversary can disrupt all communications by polluting few messages. To overcome this problem, a group testing algorithm is introduced for authentication resulting in a complexity linear in the number of attacks. The energy consumption is analyzed for cross-point and butterfly network topologies with respect to the possible attack scenarios. The results highlight the trade-offs between energy efficiency, authentication and the effective throughput for the different MAC modes.
- **Towards Stronger Jamming Model: Application to TH-UWB Radio [35]**

  With the great expansion of wireless communications, jamming becomes a real threat.We propose a new model to evaluate the robustness of a communication system to jamming. The model results in more scenarios to be considered ranging from the favorable case to the worst case. The model is applied to a TH-UWB radio. The performance of such a radio in presence of the different jamming scenarios is analyzed. We introduce a mitigation solution based on stream cipher that restricts the jamming problem of the TH-UWB communication to the more favorable case while preserving confidentiality.
- **Privacy risks quantification in Online social networks**

  In this project, we analyze the different capabilities of online social networks and aim to quantify the privacy risks users are undertaking in this context. Online Social Networks (OSNs) are a rich source of information about individuals. It may be difficult to justify the claim that the existence of public profiles breaches the privacy of their owners, as they are the ones who entered the data and made them publicly available in the first place. However, aggregation of multiple OSN public profiles is debatably a source of privacy loss, as profile owners may have expected each profile's information to stay within the boundaries of the OSN service in which it was created. First we present an empirical study of personal information revealed in public profiles of people who use multiple Online Social Networks (OSNs). This study aims to examine how users reveal their personal information across multiple OSNs. We consider the number of publicly available attributes in public

profiles, based on various demographics and show a correlation between the amount of information revealed in OSN profiles and specific occupations and the use of pseudonyms. Then, we measure the complementarity of information across OSNs and contrast it with our observations about users who share a larger amount of information. We also measure the consistency of information revelation patterns across OSNs, finding that users have preferred patterns when revealing information across OSNs. To evaluate the quality of aggregated profiles we introduce a consistency measure for attribute values, and show that aggregation also improves information granularity. Finally, we demonstrate how the availability of multiple OSN profiles can be exploited to improve the success of obtaining users' detailed contact information, by cross-linking with publicly available data sources such as online phone directories. This work has been published in ACM SIGCOMM WOSN [42].

In a second study, we examine the user tracking capabilities of the three major global Online Social Networks (OSNs). We study the mechanisms which enable these services to persistently and accurately follow users web activity, and evaluate to which extent this phenomena is spread across the web. Through a study of the top 10K websites, our findings indicate that OSN tracking is diffused among almost all website categories, independently from the content and from the audience. We also evaluate the tracking capabilities in practice and demonstrate by analyzing a real traffic traces that OSNs can reconstruct a significant portion of users web profile and browsing history. We finally provide insights into the relation between the browsing history characteristics and the OSN tracking potential, highlighting the high risk properties. This work has also been published in ACM SIGCOMM WOSN [40].

In a third study, we also analyzed the inference capabilities of third parties from seemingly harmless and unconsciously publicly shared data. Interests (or "likes") of users is one of the highly-available on-line information on the web. In this study, 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 friend relationships 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. This is the first time that user interests are used for profiling, and more generally, semantics-driven inference of private data is addressed. Our work received many media attention and was published in the prestigious NDSS symposium [39].

- **On the Privacy threats of hidden information in Wireless communication**

  Wi-Fi protocol has the potential to leak personal information. Wi- Fi capable devices commonly use active discovery mode to find the available Wi-Fi access points (APs). This mechanism includes broadcast of the AP names to which the mobile device has previously been connected to, in plain text, which may be easily observed and captured by any Wi-Fi device monitoring the control traffic. The combination of the AP names belonging to any mobile device can be considered as a Wi-Fi fingerprint, which can be used to identify the mobile device user. Our research investigates how it is possible to exploit these fingerprints to identify links between users i.e. owners of the mobile devices broadcasting such links. In this project, we have used an approach based on the similarity between the Wi-Fi fingerprints, which is equated to the likelihood of the corresponding users being linked. When computing the similarity between two Wi-Fi fingerprints, two dimensions need to be considered : (i) The number of network names in common. Indeed, sharing a network is an indication of the existence of a link, e.g. friends and family that share multiple Wi-Fi networks. (ii) The rarity of the network names in common. Some network names are very common and sharing them does not imply a link between the users. This is the case for public network names such as McDonalds Free Wi-Fi, or default network names such as NETGEAR and Linksys. On the other hand, uncommon network names such as Griffin Family Network or Orange-3EF50 are likely to

indicate a strong link between the users of these networks. Utilising a carefully designed similarity metric, we have been able to infer the existence of social links with a high confidence: 80% of the links were detected with an error rate of 7%. We show that through real-life experiments that owners of smartphones are particularly exposed to this threat, as indeed these devices are carried on persons throughout the day, connecting to multiple Wi-Fi networks and also broadcasting their connection history. There are a number of industry and research initiatives aiming to address Wi-Fi related privacy issues. The deployment of new technology i.e. privacy preserving discovery services, would necessitate software modifications in currently deployed APs and devices. The obvious solution to disable active discovery mode, comes at the expense of performance and usability, i.e. with an extended time duration for the Wi-Fi capable device to find and connect to an available AP. As a possible first step, users should be encouraged to remove the obsolete connection history entries, which may lower the similarity metric and thus reduce the ease of linkage. Our papers illustrating this study have been presented in the WoWMoM'12 conference [45] and in the IEEE MILCOM conference [43].

- **Information leakage in Ads networks**

  In targeted (or behavioral) advertising, users' behaviors are tracked over time in order to customize served ads to their interests. This creates serious privacy concerns since for the purpose of profiling, private information is collected and centralized by a limited number of companies. Despite claims that this information is secure, there is a potential for this information to be leaked through the customized services these companies are offering. In this study, we show that targeted ads expose users' private data not only to ad providers but also to any entity that has access to users' ads. We propose a methodology to filter targeted ads and infer users' interests from them. We show that an adversary that has access to only a small number of websites containing Google ads can infer users' interests with an accuracy of more than 79% (Precision) and reconstruct as much as 58% of a Google Ads profile in general (Recall). This study is the first work that identifies and quantifies information leakage through ads served in targeted advertising. We published a paper illustrating these results in the prestigious Privacy Enhancing Technologies Symposium PETS 2012 [37].

- **Privacy in P2P file sharing systems**

  In this study, we aim at characterizing anonymous file sharing systems from a privacy perspective. We concentrate on a recently deployed privacy-preserving file sharing system: OneSwarm. Our characterisation is based on measurement of several aspects of the OneSwarm system such as the nature of the shared and searched content and the geolocation and number of users. Our findings indicate that, as opposed to common belief, there is no significant difference in downloaded content between this system and the classical BitTorrent ecosystem. We also found that a majority of users appear to be located in countries where anti-piracy laws have been recently adopted and enforced (France, Sweden and U.S). Finally, we evaluate the level of privacy provided by OneSwarm, and show that, although the system has strong overall privacy, a collusion attack could potentially identify content providers. This work has been published in [46].

- **Privacy leakage on mobile devices: the Mobilitics Inria-CNIL project**

  This joint Inria-CNIL (the French data protection agency) project aims at assessing the privacy risks associated to the use of smartphones and tablets, in particular because of personal information leakage to remote third parties. Both applications and the base OS services are considered as potential source of information leakage. More precisely, the goals are to define a platform and a methodology to identify, measure, and see the evolution over the time of privacy risks.

  If similar risks exist with a PC, the situation is more worrying with mobile terminals. The reasons are:

  – the intrusive feature of these terminals that their owner continuously keep with them;
  – the amount of personnal information available on these terminals (mobile terminals aggregate personnal information but also create them, for instance with geolocalisation information);

- the facility with which the owner can personnalize its terminal with new applications;
- the financial incentives that lead companies to collect and use personnal information;
- the fact that the terminal user has no tool (e.g. a "privacy" firewall) to control precisely what information is exchanged with whom. The permissions provided by Android is too coarse grained to be useful, and the new privacy dashboard of IOS 6 does not enable the user to have an idea of how personnal information is used by an authorized application (a one time access to a personnal information and local processing within the application can be acceptable, whereas the periodic transmission of this information to remote servers is not);

The final goals of the Mobilitics project are both to study the situation and trend, but also to make mobile terminal users aware of the situation, and to provide tools that may help them to better control the personnal information flow of their terminal.

6.3. Formal and legal issues of privacy

Participants: Thibaud Antignac, Denis Butin, Daniel Le Métayer.

- **Verification of privacy properties** The increasing official use of security protocols for electronic voting deepens the need for their trustworthiness, hence for their formal verification. The impossibility of linking a voter to her vote, often called voter privacy or ballot secrecy, is the core property of many such protocols. Most existing work relies on equivalence statements in cryptographic extensions of process calculi. We have proposed the first theorem-proving based verification of voter privacy which overcomes some of the limitations inherent to process calculi-based analysis [36]. Unlinkability between two pieces of information is specified as an extension to the Inductive Method for security protocol verification in Isabelle/HOL. New message operators for association extraction and synthesis are defined. Proving voter privacy demanded substantial effort and provided novel insights into both electronic voting protocols themselves and the analysed security goals. The central proof elements have been shown to be reusable for different protocols with minimal interaction.
- **Privacy by design** The privacy by design approach is often praised by lawyers as well as computer scientists as an essential step towards a better privacy protection. The general philosophy of privacy by design is that privacy should not be treated as an afterthought but rather as a first-class requirement during the design of a system. The approach has been applied in different areas such as smart metering, electronic traffic pricing, ubiquitous computing or location based services. More generally, it is possible to identify a number of core principles that are widely accepted and can form a basis for privacy by design. For example, the Organization for Economic Co-operation and Development (OECD) has put forward principles such as the consent, limitation of use, data quality, security and accountability. One must admit however that the take-up of privacy by design in the industry is still rather limited. This situation is partly due to legal and economic reasons: as long as the law does not impose binding commitments, ICT providers and data collectors do not have sufficient incentives to invest into privacy by design. The situation on the legal side might change in Europe though because the regulation proposed by the European Commission in January 2012 (to replace the European Directive 95/46/EC) includes binding commitments on privacy by design.

  But the reasons for the lack of adoption of privacy by design are not only legal and economic: even though computer scientists have devised a wide range of privacy enhancing tools, no general methodology is available to integrate them in a consistent way to meet a set of privacy requirements. The next challenge in this area is thus to go beyond individual cases and to establish sound foundations and methodologies for privacy by design. As a first step in this direction, we have focused on the data minimization principle which stipulates that the collection should be limited to the pieces of data strictly necessary for the purpose, and we have proposed a framework to reason about the choices of architecture and their impact in terms of privacy [53]. The first strategic choices are the allocation of the computation tasks to the nodes of the architecture and the types of communications between the nodes. For example, data can be encrypted or hashed, either to protect

their confidentiality or to provide guarantees with respect to their correctness or origin. The main benefit of a centralized architecture for the "central" actor is that he can trust the result because he keeps full control over its computation. However, the loss of control by a single actor in decentralized architectures can be offset by extra requirements ensuring that errors (or frauds) can be detected *a posteriori*. In order to help the designer grasp the combination of possible options, our framework provides means to express the parameters to be taken into account (the service to be performed, the actors involved, their respective requirements, etc.) and an inference system to derive properties such as the possibility for an actor to detect potential errors (or frauds) in the computation of a variable. This inference system can be used in the design phase to check if an architecture meets the requirements of the parties or to point out conflicting requirements.

- **Privacy and discrimination**

  Actually, the interactions between personal data protection, privacy and protection against discriminations are increasingly numerous and complex. For example, there is no doubt that misuses of personal data can adversely affect privacy and self-development (for example, resulting in the unwanted disclosure of personal data to third parties, in identity theft, or harassment through email or phone calls), or lead to a loss of choices or opportunities (for example, enabling a recruiter to obtain information over the internet about political opinions or religious beliefs of a candidate and to use this information against him). It could even be suggested that privacy breaches and discriminations based on data processing are probably the two most frequent and the most serious types of consequences of personal data breaches. We have studied these interactions from a multidisciplinary (legal and technical) perspective and argued that an extended application of the application of non-discrimination regulations could help strengthening data protection [52]. We have analysed and compared personal data protection, privacy and protection against discriminations considering both the types of data concerned and the *modus operandi* (*a priori* versus *a posteriori* controls, actors in charge of the control, etc.). From this comparison, we have drawn some conclusions with respect to their relative effectiveness and argued that *a posteriori* controls on the use of personal data should be strengthened and the victims of data misuse should get compensations which are significant enough to represent a deterrence for data controllers. We have also advocated the establishment of stronger connections between anti-discrimination and data protection laws, in particular to ensure that any data processing leading to unfair differences of treatments between individuals is prohibited and can be effectively punished [29].

6.4. Network measurement, modeling and understanding

Participants: Chadi Barakat, Arnaud Legout, Ashwin Rao, Walid Dabbous, Tessema Mindaye, Mohamed Ali Kaafar, Dong Wang, Vincent Roca, Ludovic Jacquin, Byungchul Park.

The main objective of our work in this domain is a better monitoring of the Internet and a better understanding of its traffic. 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 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.

- **Checking Traffic Differentiation at the Internet Access**

  In the last few years, ISPs have been reported to discriminate against specific user traffic, especially if generated by bandwidth-hungry applications. The so-called network neutrality, advocating that an ISP should treat all incoming packets equally, has been a hot topic ever since. We propose Chkdiff, a novel method to detect network neutrality violations that takes a radically different approach from existing work: it aims at both application and differentiation technique agnosticism. We achieve this in three steps. Firstly, we perform measurements with the user's real traffic instead of using specific

application traces. Secondly, we do not assume that discrimination takes place on any particular packet field, which requires us to preserve the integrity of all the traffic we intend to test. Thirdly, we detect differentiation by comparing the performance of a traffic flow against that of all other traffic flows from the same user, considered as a whole.

Chkdiff is based on the following key ideas:
Idea 1: **Use real user traffic**. We want to test the existence of traffic discrimination for the exact set of applications run by the end user. Hence, we only consider user-generated traffic.
Idea 2: **Leave user traffic unchanged, or almost**. All methods performing active measurements send probes made of real application packets and of packets that are similar, but slightly modified, so that they do not get discriminated along their path. This is quite an assumption, as we do not know exactly what ISPs do behind the scenes. In the extreme case, ISPs could even white-list traffic generated by differentiation detecting tools. It is therefore crucial to preserve as much of the original packets as possible, as well as their original per-flow order. We will see that the modifications introduced by our tool affect only the ordering of packets, their TTL value or their IP identification field.
Idea 3: **Baseline is the entire traffic performance**. Since we do not want to make any hypothesis in advance on what kind of mechanisms - if any - are deployed, we claim that the performance of each single non-differentiated flow should present the same behaviour as that of the rest of our traffic as a whole. Differentiated flows, on the other hand, should stand out when compared to all other flows grouped together, where a large fraction of non-differentiated flows should mitigate the impact of differentiated ones.

Chkdiff is currently the subject of a collaboration with I3S around the PhD thesis of Riccardo Ravaioli (funded by the Labex UCN@Sophia). A first description of the tool is presented in [63].

- **Lightweight Enhanced Monitoring for High-Speed Networks**

  Within the collaboration with Politecnico di Bari, we worked on LEMON, a lightweight enhanced monitoring algorithm based on packet sampling. This solution targets a pre-assigned accuracy on bitrate estimates, for each monitored flow at a router interface. To this end, LEMON takes into account some basic properties of the flows, which can be easily inferred from a sampled stream, and exploits them to dynamically adapt the monitoring time-window on a per-flow basis. Its effectiveness is tested using real packet traces. Experimental results show that LEMON is able to finely tune, in real-time, the monitoring window associated to each flow and its communication overhead can be kept low enough by choosing an appropriate aggregation policy in message exporting. Moreover, compared to a classic fixed-scale monitoring approach, it is able to better satisfy the accuracy requirements of bitrate estimates. Finally, LEMON incurs a low processing overhead, which can be easily sustained by currently deployed routers, such as a CISCO 12000 device. This work is currently under submission.

- **The Complete Picture of the Twitter Social Graph**

  In this work [49], we collected the entire Twitter social graph that consists of 537 million Twitter accounts connected by 23.95 billion links, and performed a preliminary analysis of the collected data. In order to collect the social graph, we implemented a distributed crawler on the PlanetLab infrastructure that collected all information in 4 months. Our preliminary analysis already revealed some interesting properties. Whereas there are 537 million Twitter accounts, only 268 million already sent at least one tweet and no more than 54 million have been recently active. In addition, 40% of the accounts are not followed by anybody and 25% do not follow anybody. Finally, we found that the Twitter policies, but also social conventions (like the followback convention) have a huge impact on the structure of the Twitter social graph.

- **Meddle: Middleboxes for Increased Transparency and Control of Mobile Traffic**

  Mobile networks are the most popular, fastest growing and least understood systems in today's Internet ecosystem. Despite a large collection of privacy, policy and performance issues in mobile networks users and researchers are faced with few options to characterize and address them. In this work [62] we designed Meddle, a framework aimed at enhancing transparency in mobile networks and providing a platform that enables users (and researchers) control mobile traffic. In the mobile environment, users are forced to interact with a single operating system tied to their device, generally run closedsource apps that routinely violate user privacy, and subscribe to network providers that can (and do) transparently modify, block or otherwise interfere with network traffic. Researchers face a similar set of challenges for characterizing and experimenting with mobile systems. To characterize mobile traffic and design new protocols and services that are better tailored to the mobile environment, we would like a framework that allows us to intercept and potentially modify traffic generated by mobile devices as they move with users, regardless of the device, OS, wireless technology, or carrier. However, implementing this functionality is difficult on mobile devices because it requires warrantyvoiding techniques such as jail breaking to access and manipulate traffic at the network layer. Even when using such an approach, carriers may manipulate traffic once it leaves the mobile device, thus rendering some research impractical. Furthermore, researchers generally have no ability to deploy solutions and services such as prefetching and security filters, that should be implemented in the network. In this work, we designed Meddle, a framework that combines virtual private networks (VPNs) with middleboxes to provide an experimental platform that aligns the interests of users and researchers.

- **Mobile users' behavior modeling in Video on Demand systems and its implication on user privacy and caching strategies**

  In this project, we examine mobile users' behavior and their corresponding video viewing patterns from logs extracted from the servers of a large scale VoD system. We focus on the analysis of the main discrepancies that might exist when users access the VoD system catalog from WiFi or 3G connections. We also study factors that might impact mobile users' interests and video popularity. The users' behavior exhibits strong daily and weekly patterns, with mobile users' interests being surprisingly spread across almost all categories and video lengths, independently of the connection type. However, by examining the activity of users individually, we observed a concentration of interests and peculiar access patterns, which allows to classify the users and thus better predict their behavior. We also find the skewed video popularity distribution and demonstrate that the popularity of a video can be predicted using its very early popularity level. We then analyzed the sources of video viewing and found that even if search engines are the dominant sources for a majority of videos, they represent less than 10% (resp. 20%) of the sources for the highly popular videos in 3G (resp. WiFi) network. We also report that both the type of connection and the type of mobile device used have an impact on the viewing time and the source of viewing. Using our findings, we provide insights and recommendations that can be used to design intelligent mobile VoD systems and help in improving personalized services on these platforms. This work has been published in IMC 2012 [54].

- **Explicative models for Information Spreading on the web from a user profiling perspective**

Microblog services offer a unique approach to online information sharing allowing microblog users to forward messages to others. We study the process of information diffusion in a microblog service developing Galton-Watson with Killing (GWK) model, which has many implications ranging from privacy protection to experiments validation and benchmarking. We describe an information propagation as a discrete GWK process based on Galton-Watson model which models the evolution of family names. Our model explains the interaction between the topology of the social graph and the intrinsic interest of the message. We validate our models on dataset collected from Sina Weibo and Twitter microblogs. Sina Weibo is a Chinese microblog web service which reached over 100 million users as for January 2011. Our Sina Weibo dataset contains over 261 thousand tweets which have retweets and 2 million retweets from 500 thousand users. Twitter dataset contains over 1.1 million tweets which have retweets and 3.3 million retweets from 4.3 million users. The results of the validation show that our proposed GWK model fits the information diffusion of microblog service very well in terms of the number of message receivers. We show that our model can be used in generating tweets load and also analyze the relationships between parameters of our model and popularity of the diffused information. Our work is the first to give a systemic and comprehensive analysis for the information diffusion on microblog services, to be used in tweets-like load generators while still guaranteeing popularity distribution characteristics. Our paper illustrating this study will be presented in IEEE Infocom 2013 [69].

- **Tracking ICMP black holes at an Internet Scale**

  ICMP is a key protocol to exchange control and error messages over the Internet. An appropriate ICMP's processing throughout a path is therefore a key requirement both for troubleshooting operations (e.g. debugging routing problems) and for several functionalities (e.g. Path Maximum Transmission Unit Discovery, PMTUD). Unfortunately it is common to see ICMP malfunctions, thereby causing various levels of problems. In our study, we first introduce a taxonomy of the way routers process ICMP, which is of great help to understand for instance certain traceroute outputs. Secondly we introduce IBTrack, a tool that any user can use to automatically characterize ICMP issues within the Internet, without requiring any additional in-network assistance (e.g. there is no vantage point). Finally we validate our IBTrack tool with large scale experiments and we take advantage of this opportunity to provide some statistics on how ICMP is managed by Internet routers. This work has been presented in IEEE Globecom [51].

6.5. Experimental Environment for Future Internet Architecture

Participants: Walid Dabbous, Thierry Parmentelat, Fréderic Urbani, Daniel Camara, Alina Quereilhac, Shafqat Ur-Rehman, Mohamed Larabi, Thierry Turletti, Julien Tribino.

- **SFA Federation of experimental testbeds**

  We are now involved in the NOVI (E.U. STREP) project, the F-Lab (French A.N.R.) project, the FED4FIRE (E.U. IP) project and have the lead of the "Control Plane Extensions" WorkPackage of OpenLab (E.U. IP) project. Within these frameworks, as part of the co-development agreement between the Planète team and Princeton University, we have made a great deal of contributions into one of the most visible and renown implementations of 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 2012, we have focused on the maturation of the SFA specifications and the SfaWrap codebase, with several objectives in mind. Firstly, we have contributed within the GENI (N.S.F.) project to the specifications of the Version 3 of the AM-API (Aggregate Manager API), which defines the primitives that a testbed management infrastructure has to provide in order to be SFA-compliant.

Secondly, knowing that our former SFA implementation was targeting PlanetLab testbeds only, we needed on the one hand, to make generic this SFA implementation, by completely redesign and refactor its codebase, and on the other hand, we needed to support all the resources allocation strategies supported by the testbeds, namely the allocation of both 'shared' and 'exclusive' resources. As a result of this redesign and development effort, out new SFA implementation is now disseminated and started to be known, under the name of SfaWrap, and we believe that it can be used as a production-grade alternative to quickly add SFA compatibility on top of many heterogenous testbed management frameworks.

Finally, in order to allow the community of networking researchers to execute cross-testbed experiments, involving heterogeneous resources, Planète team has been instrumental in federating a set of well-known testbeds through the SfaWrap, namely PlanetLab Europe, Senslab - developed in other Inria Project-teams -, FEDERICA, the outcome of another E.U.-funded project and more recently NITOS, an OMF-enabled wireless testbed. See [96] and [97] for more details.

- **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 developpers 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 completly 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 and is currently under submission.

- **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. For more details see the Bake web page http://planete.inria.fr/software/bake/index.html

- **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.

- **Automated Deployment and Customization of Routing Overlays Across Heterogeneous Experimentation Platforms**

  During the last decades, many institutions and companies around the world have invested great effort into building new network experimentation platforms. These platforms range from simulators, to emulators and live testbeds, and provide very heterogeneous ways to access resources and to run experiments.

  Currently, a growing concern among platform owners is how to encourage researchers from different platform communities to take advantage of the resources they offer. However, one important aspect that needs to be overcome in order to appeal researchers to use as many experimentation platforms as necessary to best validate their results, is to decrease the inherent complexity to run experiments in different platforms. Even more so, to decrease the complexity of mixing resources from different platforms on a same experiment, to achieve the combination of resources best suited to the experiment needs.

  To address this concern, we developed the Network Experiment Programming Interface (NEPI) whose goal is to make easier the use of different experimentation platforms, and switch among them easily. 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. On 2010 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.

  During 2011, we 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. Also on 2011, we started work on a prototype PlanetLab backend.

  Last year, we extended NEPI to provide automated deployment and customization of routing overlays using resources from heterogeneous experimentation platforms. The main contribution of this work is to enable researchers to easily integrate different resources, such as simulated, emulated or physical nodes, on a same experiment, using a network overlay, thus addressing one of the main concerns previously mentioned.

We started by adding support to easily build routing overlays on PlanetLab, and providing the ability to customize network traffic by adding user defined filters to packets traversing the overlay tunnels [48]. We then improved this work by adding the ability to include simulated nodes from the ns-3 backend and emulated nodes from the linux containers backend into a single overlay network. We demonstrated the use of NEPI to build adn control routing overlays which incorporate resources from different on the ns-3 2012 community workshop [74].

- **Content Centric Networks Live Experimentation**

  Realistic experimentation on top of Internet-like environments is key to evaluate the feasibility of world wide deployment of CCNx, and to assess the impact of existing Internet traffic conditions on CCN traffic. However, deploying live experiments on the Internet is a difficult and error prone task, specially when performed manually.

  To address this issue, during the last year, we extended NEPI, a framework for managing network experiments, to support easy design, and automated deployment and control, of CCNx experiments on the PlanetLab testbed. Among other features, NEPI now enables the deployment of user modified CCNx sources on arbitrary PlanetLab nodes, and the creation of tunnels to enable the use of multicast FIB entries between CCNx daemons over the Internet. By supporting easy CCNx experimentation on PlanetLab, NEPI can help to explore the co-existence of CCN and TCP/IP architecture.

  This work was presented as a poster and a demo at CCNxCon 2012, the CCNx http://www.ccnx.org/ community meeting [73]. The work had a very good reception and gained NEPI some new users.

  An online tutorial and demo were also made available at NEPI's web page http://nepi.inria.fr/wiki/nepi/CCNxOnPlanetLabEurope, for dissemination purposes.

- **Smooth-transition: a new methodology for dealing with various network experiment environments**

  The smooth-transition is a new methodology, which supports various network experiment environments covering from pure simulation through realistic emulation consistently. The reproducibility in experimental network research is getting important feature for iterative experiments in short-term and long-term period. The main idea of this concept is providing the reproducibility in a broader sense. So far, we had to implement different experiments by different environment, such as simulation, application-level emulation, and link-level emulation. Whereas the smooth-transition is able to keep the context of the experiments started from a pure simulation up to a realistic emulation gradually. That means the user does not need to waste time any more for learning and following a lot of documents and manuals from each different environment. Moreover, anyone can easily start to use the testbed and to develop inside (i.e. protocol stack). Because NS3 which is the most popular and powerful network simulator has been used in this concept as an experiment engine.

  The smooth-transition employees Network Experiment Programming Interface (NEPI) to conduct all functions, such as composing scenario, node deployment, experiment control, and resource management. The core of building this concept is NS3 which has Emulation (EMU) and Direct Code Execution (DCE) modules. EMU supports to use real network devices instead of NS3 MAC and PHY layer implementations. DCE is able to launch real application on top of NS3 protocol stacks. Furthermore, real Linux kernel (currently, net-next 2.6 is available) can replace NS3 Internet protocols by its advanced mode. This concept needs back-end system covering all experiment nodes. Control and Management Framework (OMF) plays an important role as a software framework to control and manage an wireless network testbed, and all messages are exchanged by Extensible Messaging and Presence Protocol (XMPP). Nitos scheduler has been adopted as a reservation system http://nitlab.inf.uth.gr/NITlab/index.php/scheduler. The user can reserve a time slot, nodes,

and wireless channels through its web page. In addition, SFA supports that the testbed is federated with other ones of outside.

The testbed provides PCAP files as a common outcome, and this file contains captured in and out packets. However, the file size is easily over gigabytes, then it makes a very long delay to process dozens of that files. To reduce the processing time efficiently, we are using an indexing scheme for fast collecting desired packets by filtering. In particular, this scheme is very useful to find packets occurred rarely, when an detailed analysis is required for an network event, such as retransmission, intrusion detection, and node association/disassociation. The indexing information is stored in a database file, and it does not need to be modified after making the file. The size of the file is very small compared with the PCAP file, so it provides fast packet filtering permanently, even after leaving the testbed. This work, post-processing of PCAP files, is in a collaboration with Diego Dujovne and Luciano Ahumada from the Universidad Diego Portales of Chili. Especially, YoungHwan Kim, a postdoc of the Planète group, has been currently dispatched for this collaboration for fourteen weeks (September 15 2012 $\sim$ January 26 2013) in Santiago, Chile.

- **The FIT experimental platform**

  We have started, since 2011, the procedure of building a new experimental platform at Sophia-Antipolis, in the context of the FIT Equipment of Excellence project. This platform has two main goals : the first one is to enable highly controllable experiments due to its anechoic environment. These experiments can be either hybrid-experiments (as NEPI will be deployed) or federated-experiments through several testbeds. The second goal is to make resource consuming experiments (like CCNx) possible due to some powerful servers that will be installed and connected to the PlanetLab testbed. During 2012, the specifications has been defined and the procedure will continue during the next year.

- **Network Simulations on a Grid**

  We studied an hybrid approach for the evaluation of networking protocols based on the ns-3 network simulator and a Grid testbed. We analyzed the performance of the approach using a simple use case. Our evaluation shows that the scalability of our approach is mainly limited by the processor speed and memory capacities of the simulation node. We showed that by exploiting the emulation capacity of ns-3, it is possible to map complex network scenarios on grid nodes. We also proposed a basic mapping algorithm to distribute a network scenario on several node [32].

RAP Project-Team

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 investigated 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 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 a research contract. 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 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 used 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.

In 2011, we began work on bandwidth allocation in meshed networks. We have evaluated the TWIN architecture in a metropolitan area network with an original medium access control (MAC) algorithm. This algorithm was inspired by our prior work on access networks and ensures an efficient and fair allocation of bandwidth to flows between network nodes.

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 relating to this invention has been granted. This patent is owned by Orange following the terms of our contract with them. The paper [16] 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. An extended version of the paper has been accepted for publication in Journal of Optical Communication and Networking [24].

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. Some of this work is performed in collaboration with Orange Labs under the terms of our research contract. An interesting new development is the application of new coherent optical technology that allows tunable receivers as well as tunable transmitters. We are evaluating the performance of a bandwidth allocation algorithm that exploits this technology.

A wider reaching collaboration has been established under the terms of a Celtic Plus project called SASER. This project was approved by the EU in 2012 and funding has been obtained for our participation from the French authorities. The project kickoff meeting was held in November 2012. Our contribution relates to the use of TWIN to create an extended metropolitan optical network. Our partners in the corresponding work package task are Orange, Telecom Bretagne and the engineering school ENSSAT. Overall reponsibility for the work package (where alternative optical network architectures are also evaluated) is with Alcatel-Lucent Bell Labs.

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 contributes 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 also contributes to the development of efficient forwarding strategies and the elaboration of economic arguments that make CCN a viable replacement for IP. CONNECT partners are Alcatel-Lucent (lead), Orange, Inria/RAP, Inria/PLANETE, Telecom ParisTech, UPMC/LIP6.

A paper describing a proposed flow-aware approach for CCN traffic management and its performance evaluation has been presented at the conference Infocom 2012 [20]. We have reviewed the literature on cache performance (dating from early work on computer memory management) and identified a practical and versatile tool for evaluating the hit rate (proportion of requests that are satisfied from the cache) as a function of cache size and the assumed object popularity law. This approximate method was first proposed in 2002 by Che, Tung and Wang for their work on web caching. We applied this approximation to evaluate CCN caching performance taking into account the huge population and diverse popularity characteristics that make other approaches ineffective [19]. The excellent accuracy of this method over a wide range of practically relevant traffic models has been explained mathematically [18]. CONNECT ends in December 2012. We are currently defining a new project proposal that should be submitted to the ANR INFRA call in February 2013.

4.3. 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: if a node of this network has x requests to transmit, then it receives a fraction of the capacity proportional to $\log(1+x)$, the logarithm of its current load. A fluid scaling analysis of such a network is presented. We have shown that the interaction of several time scales plays an important role in the evolution of such a system, in particular its coordinates may live on very different time and space scales. As a consequence, the associated stochastic processes turn out to have unusual scaling behaviors which give an interesting fairness property to this class of algorithms. A heavy traffic limit theorem for the invariant distribution has also been proved. A generalization to the resource sharing algorithm for which the $\log$ function is replaced by an increasing function.

4.4. Algorithms: Distributed Hash Tables

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. A paper has been submitted on this subject for the case where there are at most two copies of each file [25]. An article for the general case is in preparation. 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 [11].

4.5. 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 abut 2500 different types of proteins in different proportions (from few dozens for the replication machinery up to 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 a minimal quantity for each protein type in order to ensure the smooth-running of the cell, on the other side an “overproduction policy” for all the proteins is infeasible, 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 (PPP). 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. See [12].

The theoretical environment of Poisson point processes has lead 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 hypothesis 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 elongation 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. In particular, by using a PPP representation we have been able to include the deterministic continuous phenomenon of protein degradation, which is the main protein degradation mechanism for stable proteins. We have showed moreover that this more realistic description is surprisingly identical in distribution with the classic assumption of protein degradation by means of a degrading protein (*proteosome*). We have used our model also to compare the variances resulting by choosing different hypotheses for the probability elongation, in particular we have hypothesize the protein assembly to be deterministic. This assumption is justified because of the elongation step, which consists of a large number

of elementary steps, can be described by the sum of exponential steps and the resulting distribution is well approximated by a Gaussian distribution because of the central limit theorem. Under the hypothesis of small variance of the resulting Gaussian distribution, we can assume the elongation step 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.

We have developed a C++ stochastic simulator for our general model, which has allowed the computation of variance when it was not possible to derive explicit analytic close formulas and the simulation of some extension of the actual model.

4.6. Stochastic Networks: Large Bike Sharing Systems

Participants: Christine Fricker, Hanene Mohamed, Danielle Tibi.

This is a collaboration with Nicolas Gast (EPFL) starting in December 2010. Bike sharing systems were launched by numerous cities as a part of urban transportation, for example Velib in 2007 (20 000 bikes, 1 500 stations). One of the major issues is the availability of the resources: bikes or free slots. These systems become a hot topic in Operation Research but studies on these stochastic networks are very few. To our knowledge, no theoretical study of such bike sharing systems exist taking into account the limited capacity of the stations.

We modeled this system in a symmetric case. Mean field limit gives the dynamic of a large system and the limiting stationary behavior of a single station as the system gets large. Analytical results are obtained and convergence proved in the standard model via Lyapounov functions. It allows to find the best ratio of bikes par station and to measure the improvement of incentive mechanisms, as choosing among two stations for example. Redistribution by trucks is also investigated. See [26].

Further results have been obtained for some heterogeneous systems. By mean field techniques, analytical results are obtained with Hanene Mohamed for systems with clusters (see [17]).

In a work in progress with *Danielle Tibi* , a more direct method is used when the network has a product form invariant measure by central and local limit theorem. It is a way to prove in this case the equivalence of ensembles, known in physic statistics. It applies to the simplest non homogeneous model. It gives a way to generalize the cluster case.

4.7. Random Graphs

Participant: Nicolas Broutin.

4.7.1. Connectivity in models of wireless networks

This is joint work with S. Boucheron (Paris 7), L. Devroye (McGill), N. Fraiman (McGill), and G. Lugosi (Pompeu Fabra).

The traditional models for wireless networks rely on geometric random graphs. However, if one wants to ensure that the graph be fully connected the radius of influence (hence the power necessary, and number of links) is too large to be fully scalable. Recently some models have been proposed that skim the neighbours and only retain a random subset for each node, hence creating a sparser overlay that would hopefully be more scalable. The first results on the size of the subsets which guarantee connectivity of overlay (the irrigation graph) [3] confirm that the average number of links per node is much smaller, but it remains large. These results motivate further investigations on the size of the largest connected component when one enforces a constant average degree which are in the process of being written.

4.7.2. Random graphs and minimum spanning trees

This is a long term collaboration with L. Addario-Berry (McGill), C. Goldschmidt (Oxford) and G. Miermont (ENS Lyon).

The random graph of Erdős and Rényi is one of the most studied models of random networks. Among the different ranges of density of edges, the "critical window" is the most interesting, both for its applications to the physics of phase transitions and its applications to combinatorial optimization (minimum spanning tree, constraint satisfaction problems). One of the major questions consists in determining the distribution of distances between the nodes. A limit object (a scaling limit) has been identified, that allows to describe precisely the first order asymptotics of pairwise distances between the nodes. This limit object is a random metric space whose definition allows to exhibit a strong connection between random graphs and the continuum random tree of Aldous. A variety of questions like the diameter, the size of cycles, etc, may be answered immediately by reading them on the limit metric space [2].

In a stochastic context, the minimum spanning tree is tightly connected to random graphs via Kruskal's algorithm. Random minimum spanning trees have attracted much research because of their importance in combinatorial optimization and statistical physics; however, until now, only parameters that can be grasped by local arguments had been studied. The scaling limit of the random graphs obtained in [2] permits to describe precisely the metric space scaling limit of a random minimum spanning tree [21], which identifies a novel continuum random tree which is truly different from that of Aldous.

4.7.3. Analysis of recursive partitions

This is joint work with R. Neininger (Frankfurt) and H. Sulzbach (Frankfurt/McGill).

The quadtrees are essential data structures that permit to store and manipulate geometric data by building a recursive partition of the space. In order to evaluate their performance, Flajolet and his co-authors have estimated the average cost of reporting all the data matching certain random queries. When the query does not fully specify all the fields, one talk about a partial match query. Such queries are ubiquitous, but analyzing their behaviour turns out to be intricate, and no performance guarantee was available in the form of a bound on the probability that any query would take much more time that one expects. [14] provides such guarantees by analysing the behaviour of all the queries at the same time, as a process. This yields estimates for the cost of the worst possible query (not a uniformly random one), as well as asymptotics for the variance and higher moments.

This line of research has motivated the analysis of the related combinatorial model of recursive lamination of the disk. The model had been recently introduced, but no full analysis was available. The techniques developed in the context of quadtrees have inspired a proof that the dual tree of the recursive lamination does converge to a limit tree-like metric space which is identified [23].

4.7.4. Navigation and point location in Poisson Delaunay triangulation

Nicolas Broutin has recently initiated a project with O. Devillers (Inria Sophia) and R. Hemsley (Inria Sophia) concerning the performance of local routing algorithms in plane subdivisions. Such algorithms also turn out to be important for the *point location* problem: for instance, finding the face of the subdivision which contains a query point is the first step towards inserting this point as a vertex. The aim is to prove that when the subdivision consists of the faces of a Delaunay triangulation, and when the points are random, any natural strategy which would take you closer to the aim performs well. Preliminary results about a specific routing algorithm, the cone walk, that we designed for its amenability to analysis appear in [22].

4.8. Stochastic Networks: Jackson Networks

Participant: Danielle Tibi.

Lyapounov functions and essential spectral radius of Jackson networks, joint work with I. Ignatiouk-Robert (University of Cergy-Pontoise). A family of explicit multiplicative Lyapounov 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. See [27].

SOCRATE Team

6. New Results

6.1. Flexible Radio Node

6.1.1. Radio wave propagation

The MR-FDPF (Multi-Resolution Frequency Domain Partial Flow) method is proven to be a fast and efficient method to simulate radio wave propagation. It is a deterministic model which can provide an accurate radio coverage prediction. In reality, radio channels have the nature of randomness due to e.g. moving people or air flow. Thus they can not be rigorously simulated by a pure deterministic model. However, it is believed that some statistics can be extracted from deterministic models and these statistics can be very useful to describe radio channels in reality. In [20], large scale fading statistical characteristics are extracted based on the MR-FDPF method. They are validated by comparison to both the theoretical result and measurement. The match also demonstrates that MR-FDPF is capable of simulating large scale fading.

In [2] we study Realistic Prediction of Bit error rate (BER) and adaptive modulation and coding (AMC) for Indoor Wireless Transmissions. Bit error rate is an important parameter for evaluating the performance of wireless networks. In this letter, a realistic BER for indoor wireless transmissions is predicted. The prediction is based on a deterministic radio propagation model, the MR-FDPF model, which is capable of providing accurate fading statistics. The obtained BER map can be used in many cases, e.g., adaptive modulation and coding scheme or power allocation.

In [4], we propose a modification of the MR-FDPF method that allows simulating radio propagation channels in a frequency range. The performance of the proposed MR-FDPF implementation has been analyzed based on different realistic propagation scenarios. We also analyze the possibility of applying the multi-resolution frequency domain approach to the well-known transmission-line matrix method. The proposed multi-resolution frequency domain transmission-line matrix method provides a computationally efficient way of modeling radio wave propagation in three dimensional space at multiple frequencies.

In [3], we consider the performance of coded wireless communication systems experiencing non-frequency selective fading channels in shadowed environments. The quality of service (QoS) in a wireless network is dependent on the packet error outage (PEO). We address the problem of finding a tractable expression for the coded PEO over Nakagami-m channels with shadowing, considering multilevel modulations, various block, convolutional channel coding schemes and hard decision decoding. In order to obtain the coded PEO, an inversion of the coded packet error probability (PEP) w.r.t. the signal to noise ratio (SNR) is needed. To this end, we propose an invertible approximation for the coded PEP w.r.t. the uncoded bit error probability (BEP) in Nakagami-m fading channels which is accurate for all BEPs of interest. The BEP itself depends on the average SNR and we hence make use of previous results on the inversion of the uncoded BEP w.r.t. the SNR in Nakagamim fading channels, holding for M-PSK and M-QAM signals. We were thus able to obtain a reliable closed form expression for the coded PEO in flat fading and shadowing channels

6.1.2. Power consumption

In [24], we propose the use of an existing opensource network simulator, WSNet, to evaluate the interest of using multi-mode relays in terms of energy consumption. We show that the combination of MIMO and multi-mode provides a solution to reduce global energy consumption, but that conclusions are really scenario-dependent. Moreover, we explain how a multi-mode MIMO terminal can improve these results using adaptive strategies.

the energy consumption in wireless sensor networks is studied. In order to minimize the consumed power at the analog and RF part, an energy recovering system combined with a wake-up radio is proposed for discussion. The proposed architecture has three activity levels : zero consumption, low and high energy consumption. In order to quantify the gain in terms of power consumption, a power consumption model state of the art is proposed. in [7] all radio channel models which can be used for MIMO heterogeneous network with small cells are described.

6.1.3. MIMO

In [28], we study MIMO and next generation system. For the past decade or more MIMO systems have been the subject of very intensive research. However in the past few years, these techniques have begun to be implemented in practice. In particular they have appeared in the standards for next generation systems such as LTE, 3GPP-LTE Advanced and WiMAX, as well as the latest versions of Wifi. This chapter, extracted from the book edited by the Cost Action 2100: "Pervasive Mobile and Ambient Wireless Communications", brings together the MIMO systems used in next generation systems with other work on the implementation and simulation of these systems. It also describes advances in MIMO techniques in a number of areas. The first section is divided into two sub-sections dealing first with simulators and testbeds which are used in system-level simulators to evaluate overall system capacity, as discussed in later chapters of this book. Secondly the development of terminals for next generation MIMO systems is considered, especially considering the additional RF hardware required for MIMO. Section 7.2 then discusses especially precoding techniques used in many of the recent standards to implement MIMO. In particular precoding allows the implementation of closed loop or adaptive MIMO. In next generation systems there is also much increased attention on MU-MIMO and on multi-terminal MIMO in general, including so-called "network MIMO " approaches, which appear in LTE as Coordinated Multiple Point: this is covered in Sect. 7.3. Various advanced MIMO transmission and detection approaches are covered in Sects. 7.4 to 7.6, including some interesting work on MIMO techniques involving continuous phase modulation, giving advantages in terms of peak-to-average power ratio.

6.2. Agile Radio Resource Sharing

6.2.1. Wireless Multi-hop Networks

In [6], we study energy-delay tradeoff in wireless multihop networks with unreliable links. Energy efficiency and transmission delay are very important parameters for wireless multihop networks. Numerous works that study energy efficiency and delay are based on the assumption of reliable links. However, the unreliability of channels is inevitable in wireless multihop networks. In addition, most of works focus on self-organization protocol design while keeping non-protocol system parameters fixed. While, very few works reveal the relationship between the network performance and these physical parameters, in other words, the best networks performance could be obtained by the physical parameters. This paper investigates the tradeoff between the energy consumption and the latency of communications in a wireless multihop network using a realistic unreliable link model. It provides a closed-form expression of the lower bound of the energy–delay tradeoff and of energy efficiency for different channel models (additive white Gaussian noise, Rayleigh fast fading and Rayleigh block-fading) in a linear network. These analytical results are also verified in 2-dimensional Poisson networks using simulations. The closed-form expression provides a framework to evaluate the energy–delay performance and to optimize the parameters in physical layer, MAC layer and routing layer from the viewpoint of cross-layer design during the planning phase of a network.

6.2.2. Relay and Cooperative Communications

In [16], we aim at characterizing the gain induced by using relay channels in a linear network under both capacity constraint and realistic energy model. We express a general model based on a convex optimization problem. Then, we use numerical tools to obtain results on the outer and inner bounds of the capacity of the full and half duplex relay channel. We then extend this 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 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.

In [15], we present preliminary results on achievables rates in half-duplex cooperative multiple access channels (CMAC).We show that the upper bound on the capacity of the half-duplex CMAC can be solved using convex optimization techniques. Under a Gaussian model, we study the maximal achievable rate by every node in the network. We propose a number of scenarios, encompassing existing and theoretical cooperation schemes. Using these hypotheses, we evaluate the performance of both a non-cooperative concurrent access and simple cooperative multi-hop or relaying schemes with respect to the upper bound. The performance is compared for the various scenarios, and we provide analyses of specific cases in order to illustrate how our framework may be used to answer targeted questions about the capacity of CMACs.

In [31], we aim at obtaining usable bounds on the performance of CMAC under a Gaussian model. We first show that the problem can be transformed into a convex optimization problem which makes it easily solvable using numerical tools. We propose, as a line of study, to consider the maximal achievable common rate by every node in the network. We then proceed to express closed-form bounds on the capacity region of the CMAC in that common rate scenario. We study simple cooperation schemes based on existing results in relay channels and compare them to other medium sharing approaches. In the end, we show that using the relay-channel based protocols can be efficient for some parameters, but gets less interesting in the Gaussian case if the source-destination links are good enough.

In [30], we study the optimal power allocations in CMACs, where we aim at maximizing the rate achievable by both sources simultaneously rather than the sum of achievable rates. Separating our study between the coherent and non-coherent case, we obtain closed-form expressions for the optimal power allocations w.r.t. the outer bounds of the capacity region, as well as decode-and-forward and non-cooperative inner bounds. We point out during our resolution that the general CMAC model behaves as a multiple access relay channel (MARC), where a ”virtual” relay node is introduced to represent the cooperation between the sources. This equivalent model simplifies the original power allocation problem. We finally show that the general cut-set outer bound on the capacity region of the equivalent MARC matches exactly the tightest known outer bound on the capacity region of the original CMAC.

In [17] we address the distributed power adaptation problem on the downlink for wireless cellular networks. As a consequence of uncoordinated local scheduling decisions in classical networks, the base stations produce mutual uncontrolled interference on their co-channel users. This interference is of a variable nature, and is hardly predictable, which leads to suboptimal scheduling and power control decisions. While some works propose to introduce cooperation between base stations, in this work we propose instead to introduce a model of power variations, called trajectories in the powers space, to help each base station to predict the variations of other base stations powers. The trajectories are then updated using a Model Predictive Control (MPC) to adapt transmit powers according to a trade-off between inertia (to being predictable) and adaptation to fit with capacity needs. A Kalman filter is used for the interference prediction. In addition, the channel gains are also predicted, in order to anticipate channel fading states. This scheme can be seen as a dynamic distributed uncoordinated power control for multichannel transmission that fits the concept of self-optimised and self-organised wireless networks. By using the finite horizon MPC, the transmit powers are smoothly adapted to progressively leave the current trajectory toward the optimal trajectory. We formulate the optimisation problem as the minimisation of the utility function of the difference between the target powers and MPC predicted power values. The presented simulation results show that in dynamic channel conditions, the benefit of our approach is the reduction of the interference fluctuations, and as a consequence a more accurate interference prediction, which can further lead to a more efficient distributed scheduling, as well as the reduction of the overall power consumption.

6.2.3. BAN

In [26] we present a simple Body Area Network (BAN) platform that was built to monitor the performance of a marathon athlete all along the race, meeting real-time and QoS constraints, under good transmission conditions. Data collected during the event (packet loss, signal strength) allowed us to obtain a primary knowledge about the behavior of the radio transmissions between the different links in the network. The results of this experiment and ther important disparities observed between the links point out the need to improve the transmission strategy.

6.2.4. Network coding

One of the most powerful ways to achieve transmission reliability over wireless links is to employ efficient coding techniques. In [10] investigates the performance of a transmission over a relay channel where information is protected by two layers of coding. In the first layer, transmission reliability is ensured by fountain coding at the source. The second layer incorporates network coding at the relay node. Thus, fountain coded packets are re-encoded at the relay in order to increase packet diversity and reduce energy consumption. Performance of the transmission is measured by the total number of transmissions needed until the message is successfully decoded at the destination. We show through both analytical derivations and simulations that adding network coding capabilities at the relay optimizes system resource consumption. When the source uses a random linear fountain code, the proposed two layer encoding becomes more powerful as it reduces the transmission rate over the direct link between the source and the destination.

In [27] we study the deployment of fountain codes and network coding in a wireless sensor network (WSN). A WSN is composed of sensor nodes with restricted capacities: memory, energy and computational power. The nodes are usually randomly scattered across the monitored area and the environment may vary. In the presence of fading, outage and node failures, fountain codes are a promising solution to guaranty reliability and improve transmission robustness. The benefits of fountain codes are explored based on an event-driven WSNet simulator considering realistic implementation based on standard IEEE802.15.4. Fountain codes are rateless and capable of adapting their rate to the channel on the fly using a limited feedback channel. In this thesis, we highlight the benefits brought by fountain code in terms of energy consumption and transmission delay. In addition to the traditional transmission with fountain code, we propose in this thesis to study the network coding transmission scheme where nodes are allowed to process the information before forwarding it to their neighbors. By this means, we can say that packet diversity is exploited as each individual packet is unique and contains different representations of binary data. Redundancy is thus optimized since repetitions are avoided and replaced with diversified information. This can further lead to an overall improved performance in cooperative communication where nodes are allowed to assist in relaying packets from the source the destination. We highlight in this thesis the benefits of fountain code combined to network coding and show that it leads to a reduction in transmission delay and energy consumption. The latter is vital to the life duration of any wireless sensor network.

In [9] we tackles the problem of providing end to end reliable transmissions in a randomly deployed wireless sensor network. To this aim, we investigate the simultaneous use of gradient broadcast routing (for its inherent adaptability to any network topology and its changes), fountain codes (for their universal property) and intra-flow network coding (to introduce packet diversity in redundant copies). We present the impact of the proposed strategy on a realistic network. This work permits to highlight that, compared to basic gradient broadcast routing, the strategy not only improves the reliability and the delay in the network but also clearly increases its lifetime.

6.2.5. Vehicular networks

In [22] we study a hybrid propagation model For large-scale variations caused By vehicular traffic in small cells. we present a propagation model generating time series of large-scale power variations for small-cell radio links intersected by vehicular traffic. The model combines stochastic processing and geometric computation. For each road crossing a link, a two-state process parameterized by mobility statistics represents the obstruction status. When the status is set to obstructed, a fluctuation pattern is generated. Based on previously published measurements, both mobility statistics and time series results are validated through the comparison of respectively inter-obstruction duration distributions and outage probabilities. The proposed model avoids resource consuming iterative propagation prediction while providing realistic and frequency adaptive results.

In [21], we performed measurements of large-scale variations caused by vehicular traffic in small-cell. This paper presents and characterizes large-scale variations of received power generated by vehicular traffic crossing a radio link. Measurements in the 2 GHz band for several small-cell configurations involve various transmitter heights, link distances and urban densities. Observations showed that stronger losses up to 30 dB

are due to medium to high vehicles. Lower vehicles have a smaller impact in links perpendicular to traffic, but amplitude variations and duration can reach larger values when the receiver is at cell radius limits.

6.2.6. security

In [18] we study Security Embedding on ultra wideband impulse radio(UWB-IR) Physical Layer. The main goal of this work is to incorporate security in an existing ultra wideband (UWB) network. We present an embedding method where a tag is added at the physical layer and superimposed to the UWB-impulse radio signal. The tag should be added in a transparent way so that guaranteeing compatibility with existing receivers ignoring the presence of the tag. We discuss technical details of the new embedding method. In addition, we discuss embedding strength and we analyze robustness performance. We demonstrate that the proposed embedding technique meets all the system design constraints.

In [11] we study Jamming in time-hopping ultrawide band (TH-UWB) Radio. With the great expansion of wireless communications, jamming becomes a real threat.We propose a new model to evaluate the robustness of a communication system to jamming. The model results in more scenarios to be considered ranging from the favorable case to the worst case. The model is applied to a TH-UWB radio. The performance of such a radio in presence of the different jamming scenarios is analyzed. We introduce a mitigation solution based on stream cipher that restricts the jamming problem of the TH-UWB communication to the more favorable case while preserving confidentiality.

6.2.7. Network Information Theory

Fundamental performance limits of multi-hop wireless transmissions are being investigated in [33] from a multiobjective perspective where transmission decisions (i.e. relay selection, scheduling or routing decision) modify the trade-off between capacity, reliability, end-to-end delay or network-wide energy consumption. In our previous work presented in the Inria research report RR-7799, Pareto-optimal performance bounds and network parameters have been derived for a 1-relay and 2-relay network within a MultiObjective(MO) performance evaluation framework. We show in this report that these bounds are tight since they can be reached by simple practical coding strategies performed by the source and the relays. Such strategies constitute achievable lower MO performance bounds on the real MO performance limits. More precisely, we adopt a coding strategy where the source transmits a random linear fountain code which is coupled to a network coding strategy performed by the relays. Two different network coding strategies are investigated. Practical performance bounds for both strategies are compared to the theoretical bound. We show that the theoretical bound is tight: generational distance between the practical and theoretical bound for the best strategy is only of 0.0042

In [19] we revisit the problem of non-cooperativ association of mobiles to access points using game theory. We consider in this paper games related to the association problem of mobiles to an access point. It consists of deciding to which access point to connect. We consider the choice between two access points or more, where the access decisions may depend on the number of mobiles connected to each one of the access points. We obtain new results using elementary tools in congestion and crowding games.

In [23] we study stochastic analysis of energy savings with sleep mode in Orthogonal Frequency-Division Multiple Access (OFDMA) wireless networks. The issue of energy efficiency in OFDMA wireless networks is discussed in this paper. Our interest is focused on the promising concept of base station sleep mode, introduced recently as a key feature in order to dramatically reduce network energy consumption. The proposed technical approach fully exploits the properties of stochastic geometry, where the number of active cells is reduced in a way that the outage probability, or equivalently the signal to interference plus noise distribution, remains the same. The optimal energy efficiency gains are then specified with the help of a simplified but yet realistic base station power consumption model. Furthermore, the authors extend their initial work by studying a non-singular path loss model in order to verify the validity of the analysis and finally, the impact on the achieved user capacity is investigated. In this context, the significant contribution of this paper is the evaluation of the theoretically optimal energy savings of sleep mode, with respect to the decisive role that the base station power profile plays.

6.3. Software Radio Programming Model

6.3.1. Virtual Radio Machine

In [14] we present a survey of existing prototypes dedicated to software defined radio. We propose a classification related to the architectural organization of the prototypes and provide some conclusions about the most promising architectures. This study should be useful for cognitive radio testbed designers who have to choose between many possible computing platforms. We also introduce a new cognitive radio testbed currently under construction and explain how this study have influenced the test-bed designers choices.

6.3.2. Embedded systems

In [13], we explore new area/throughput trade-offs for the Girault, Poupard and Stern authentication protocol (GPS). This authentication protocol was selected in the NESSIE competition and is even part of the standard ISO/IEC 9798. The originality of our work comes from the fact that we exploit a fixed key to increase the throughput. It leads us to implement GPS using the Chapman constant multiplier. This parallel implementation is 40 times faster but 10 times bigger than the reference serial one. We propose to serialize this multiplier to reduce its area at the cost of lower throughput. Our hybrid Chapman's multiplier is 8 times faster but only twice bigger than the reference. Results presented here allow designers to adapt the performance of GPS authentication to their hardware resources. The complete GPS prover side is also integrated in the network stack of the POW-WOW sensor which contains an Actel IGLOO AGL250 FPGA as a proof of concept.

The people involved in this axes also published in the computer science flied. For instance in [1] static vulnerability detection in java service-oriented components is studied. In [12] A lightweight Hash function family based on FCSRs is studied.

TREC Project-Team

6. New Results

6.1. Design and Performance Analysis of Wireless Networks

Participants: François Baccelli, Bartłomiej Błaszczyszyn, Chung Shue Chen, Miodrag Jovanović, Holger Paul Keeler, Mir Omid Haji Mirsadeghi, Frédéric Morlot, Tien Viet Nguyen.

CDMA/UMTS, Wireless LANs, ad hoc networks, IEEE 802.11, mesh networks, cognitive radio, Hiperlan, CSMA, TCP, MAC protocols, exponential back-off protocols, signal to interference ratio, coverage, capacity, transport capacity, admission and congestion control.

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 [17] 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-opportunist 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 [61].

6.1.1.2. Impact of propagation-loss model on the geometry and performance of cellular networks

6.1.1.2.1. Impact of Shadowing on QoS

Shadowing is believed to degrade the quality of service in wireless cellular networks. In [18] we discovered a more subtle reality. Increasing variance of the lognormal shadowing tends to “separate” the strongest (serving BS) signal from all other signals — a phenomenon observed for heavy-tailed distributions and called “single big jump principle”. In consequence, in some cases, an increase of the variance of the shadowing can significantly reduce the mean interference factor and improve some QoS metrics in interference limited systems. We exemplify this phenomenon, similar to stochastic resonance and related to the “single big jump principle” of the heavy-tailed log-nornal distribution, 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.

6.1.1.2.2. Using Poisson processes to model lattice cellular networks

In [51] we mathematically proved that a large spatially homogeneous (arbitrary, including hexagonal) network is perceived by a typical user as an equivalent (infinite) Poisson network, provided shadowing is strong enough. This justifies an almost ubiquitous Poisson assumption made in the stochastic-analytic approach to study of the quality of user-service in cellular networks.

6.1.1.2.3. Linear-Regression Estimation of the Propagation-Loss Parameters Using Mobiles' Measurements

In [35] we proposed a new linear-regression model for the estimation of the path-loss exponent and the parameters of the shadowing from the propagation-loss data collected by the mobiles with respect to their serving base stations. The model is based on the aforementioned Poisson convergence result.

6.1.1.3. Quality of Real-Time Streaming in Wireless Cellular Networks

In [50] we present a new stochastic service model with service capacity sharing and interruptions, meant to be useful for the performance evaluation and dimensioning of wireless cellular networks offering real-time streaming, like e.g. mobile TV. Our general model takes into account Markovian, multi-class process of call arrivals, arbitrary streaming time distribution, and allows for a general service (outage) policy saying which users are temporarily denied the service due to insufficient service capacity. Using Palm theory formalism, we develop expressions for several important characteristics of this model, including mean time spent in outage and mean number of outage incidents for a typical user of a given class. We also propose some natural class of least-effort-served-first service policies, for which the aforementioned expressions can be efficiently evaluated on the basis of the Fourier analysis of Poisson process. Last but not least, we show how our model can be used to analyse the quality of real-time streaming in 3GPP Long Term Evolution (LTE) cellular networks. We identify and evaluate an optimal and a fair service policy, the latter being suggested by LTE implementations, as well as propose some intermediate policies which allow to solve the optimality/fairness trade-off caused by unequal user radio-channel conditions.

6.1.1.4. Theoretically Feasible QoS in a MIMO Cellular Network Compared to the Practical LTE Performance

Our goal in [39] is to build a global analytical approach for the evaluation of the quality of service perceived by the users in wireless cellular networks which is calibrated in some reference cases. To do so, a model accounting for interference in a MIMO cellular system is firstly described. An explicit expression of users bit-rates theoretically feasible from the information theory point of view is then deduced. The comparison between these bit-rates and practical LTE performance permits to obtain the progress margins for potential evolution of the technology. Moreover, it leads to an analytical approximate expression of the system performance which is calibrated with the practical one. This expression is the keystone of a global analytical approach for the evaluation of the QoS perceived by the users in the long run of users arrivals and departures in the network. We illustrate our approach by calculating the users QoS as function of the cell radius in different mobility and interference cancellation scenarios.

6.1.1.5. Self-Optimization of Radio Resources in Cellular Networks

In [19], we surveyed the mathematical and algorithmic tools for the self-optimization of mobile cellular networks based on Gibbs' sampler. This technique allows for the joint optimization of radio resources in heterogeneous cellular networks made of a juxtaposition of macro and small cells. 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.

Three patents were filed on this line of thought under the Inria/Alcatel–Lucent joint laboratory.

6.1.1.6. Coverage in Cellular Networks

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] [20], 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 a game theoretic view of Spatial Aloha in collaboration with E. Altman and M.K. Hanawal [Inria MAESTRO] [22] This line of though is currently continued with Chandramani Singh. We also compared the performance of spatial Aloha to CSMA.

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 collaboration with Qualcomm [12], 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. Comparison of the maximal spatial throughput of Aloha and CSMA in Wireless multihop Ad-Hoc Networks

In [46] this paper we compare the spatial throughput of Aloha and Carrier Sense Multiple Access (CSMA) in Wireless multihop Ad-Hoc Networks. In other words we evaluate the gain offered by carrier sensing (CSMA) over the pure statiscal collision avoidance which is the basis of Aloha. We use a Signal-to-Interference-and-Noise Ratio (SINR) model where a transmission is assumed to be successful when the SINR is larger than a given threshold. Regarding channel conditions, we consider both standard Rayleigh and negligible fading. For slotted and non-slotted Aloha, we use analytical models as well as simulations to study the density of successful transmissions in the network. As it is very difficult to build precise models for CSMA, we use only simulations to compute the performances of this protocol. We compare the two Aloha versions and CSMA on a fair basis, i.e. when they are optimized to maximize the density of successful transmissions. For slotted Aloha, the key optimization parameter is the medium access probability, for non-slotted Aloha we tune the mean back-off time, whereas for CSMA it is the carrier sense threshold that is adjusted. Our study shows that CSMA always outperforms slotted Aloha, which in turn outperforms its non-slotted version.

6.1.2.3. Stochastic Analytic Evaluation of End-to-End Performance of Linear Nearest Neighbour Routing in MANETs with Aloha

Planar Poisson models with the Aloha medium access scheme have already proved to be very useful in studies of mobile ad-hoc networks (MANETs). However, it seems difficult to quantitatively study the performances of end-to-end routing in these models. In order to tackle this problem, in [52], we study a linear stationary route embedded in an independent planar field of interfering nodes. We consider this route as an idealization of a "typical" route in a MANET obtained by some routing mechanism. Such a decoupling allows us to obtain many numerically tractable expressions for local and mean end-to-end delays and the speed of packet progression, assuming slotted Aloha MAC and the Signal-to-Interference-and-Noise Ratio (SINR) capture condition, with the usual power-law path loss model and Rayleigh fading. These expressions show how the

network performance depends on the tuning of Aloha and routing parameters and on the external noise level. In particular we show a need for a well-tuned lattice structure of fixed relaying nodes, which helps to relay packets on long random routes in the presence of a non-negligible noise. We also consider a Poisson-line MANET model, in which nodes are located on roads forming a Poisson-line process. In this case our linear route is rigorously (in the sense of Palm theory) the typical route in this Poisson-line MANET.

6.1.3. Vehicular Ad-Hoc Networks (VANETs)

Vehicular Ad Hoc NETworks (VANETs) are special cases of MANETs where the network is formed between vehicles. VANETs are today the most promising civilian application for MANETs and they are likely to revolutionize our traveling habits by increasing safety on the road while providing value added services.

6.1.3.1. Point-to-Point, Emergency and Broadcast Communications

Our aim in [36] is to analyze the Aloha medium access (MAC) scheme in one-dimensional, linear networks, which might be an appropriate assumption for VANETs. The locations of the vehicles are assumed to follow a homegeneous Poisson point process. Assuming powerlaw mean path-loss and independent point-to-point fading we study performance metrics based on the signal-over-interference and noise ratio (SINR). In contrast to previous studies where the receivers are at a fixed distance from the transmitter, we assume here that the receivers are the nearest neighbors of the transmitters in the Poisson process and in a given direction. We derive closed formulas for the capture probability and for the density of progress of a packet sent by a given node. We compute the mean delay to send a packet transmitted at each slot until successful reception. We also evaluate an upper bound to discover the neighborhood within a given space interval. We show that we can include noise in the previous models.

6.1.4. Cognitive Radio Networks

We wrote a survey [26] 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.

6.2. Network Dynamics

Participants: Abir Benabid, Julieta Bollati, Anne Bouillard, Ana Bušić, Emilie Coupechoux, Nadir Farhi.

Queueing network, stability, inversion formula, probing, estimator, product-form, insensitivity, markov decision, max-plus algebra, network calculus.

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 in FIFO tandem networks

In cooperation with Giovanni Stea [University of Pisa, Italy], we present in [31] algorithms to compute worst-case performance upper bounds when the service policy is FIFO, using linear programming. Linear programming leads to tight bounds; however, the computation corst is too high for reasonable-size networks. We then develop approximate solution schemes to find both upper and lower delay bounds on the worst-case delay. Both of them only require to solve just one LP problem, and they produce bounds which are generally more accurate than those found in the literature. Finally, we have a conjecture on what sould be the worst-case trajectory under usual assumptions.

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 have been presented at Valuetools (invited paper, [29]).

6.2.1.3. Using arrival curves for detecting anomalies in a network

In cooperation with Aurore Junier [Inria/IRISA] and Benoît Ronot [Alcatel-Lucent], we present an on-line algorithm that performs a flow of messages analysis. More precisely, it is able to highlight hidden abnormal behaviors that existing network management methods would not detect. Our algorithm uses the notion of constraint curves, introduced in the Network Calculus theory, defining successive time windows that bound the flow. The advantage of this algorithm is that it can be performed online, and in a second version has different levels of precision. This work has been presented in [30] and a patent [57] has been submitted.

6.2.1.4. Min,plus algorithms for fast weak-KAM integrators

In cooperation with Erwan Faou [IPSO-Inria Rennes, DMA-ENS] and Maxime Zavidovique [Paris 6]. We consider a numerical scheme for Hamilton-Jacobi equations based on a direct discretization of the Lax-Oleinik semi-group. We prove that this method is convergent with respect to the time and space stepsizes provided the solution is Lipschitz, and give an error estimate. Moreover, we prove that the numerical scheme is a geometric integrator satisfying a discrete weak-KAM theorem which allows to control its long time behavior. Taking advantage of a fast algorithm for computing min-plus convolutions based on the decomposition of the function into concave and convex parts, we show that the numerical scheme can be implemented in a very efficient way. The results can be found in [49].

6.2.2. 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.

6.2.2.1. Piecewise Homogeneous Events

In collaboration with Bruno Gaujal [Inria Grenoble - Rhone-Alpes], we proposed in [15] 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 queueing 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).

6.2.2.2. Perfect Sampling of Networks with Finite and Infinite Capacity Queues

In [33], we consider open Jackson queueing networks with mixed finite and infinite buffers and analyze the efficiency of sampling from their exact stationary distribution. We show that perfect sampling is possible, although the underlying Markov chain has a large or even infinite state space. The main idea is to use a Jackson network with infinite buffers (that has a product form stationary distribution) to bound the number of initial conditions to be considered in the coupling from the past scheme. We also provide bounds on the sampling time of this new perfect sampling algorithm under hyper-stability conditions (to be defined in the paper) for each queue. These bounds show that the new algorithm is considerably more efficient than existing perfect samplers even in the case where all queues are finite. We illustrate this efficiency through numerical experiments.

6.2.3. 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).

6.2.3.1. 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 [14]. 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.

6.2.3.2. 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] [16], 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.

6.2.3.3. Bounded State Space Truncation

Markov chain modeling often suffers from the curse of dimensionality problems and many approximation schemes have been proposed in the literature that include state-space truncation. Estimating the accuracy of such methods is difficult and the resulting approximations can be far from the exact solution. Censored Markov chains (CMC) allow to represent the conditional behavior of a system within a subset of observed states and provide a theoretical framework to study state-space truncation. However, the transition matrix of a CMC is in general hard to compute. Dayar et al. (2006) proposed DPY algorithm, that computes a stochastic bound for a CMC, using only partial knowledge of the original chain. In [32], we prove that DPY is optimal for the information they take into account. We also show how some additional knowledge on the chain can improve stochastic bounds for CMC.

6.2.4. 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.: *Stochastic 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).

6.2.4.1. Density Classification on Infinite Lattices and Trees

In a joint work with N. Fatès [Inria Nancy – Grand-Est], J. Mairesse and I. Marcovici [LIAFA, CNRS and Université Paris 7] [43] 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.

6.3. 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 [25], we consider a model of diffusion where the individuals' behavior is the result of a strategic choice. We study 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 show 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 compute the contagion threshold for a general diffusion model and show the existence of (continuous and discontinuous) phase transitions. We also compute 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 show 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. Coordination in Network Security Games: a Monotone Comparative Statics Approach

Malicious softwares or malwares for short have become a major security threat. While originating in criminal behavior, their impact are also influenced by the decisions of legitimate end users. Getting agents in the Internet, and in networks in general, to invest in and deploy security features and protocols is a challenge, in particular because of economic reasons arising from the presence of network externalities. In [24], [42], we focus on the question of incentive alignment for agents of a large network towards a better security. We start with an economic model for a single agent, that determines the optimal amount to invest in protection. The model takes into account the vulnerability of the agent to a security breach and the potential loss if a security breach occurs. We derive conditions on the quality of the protection to ensure that the optimal amount spent

on security is an increasing function of the agent's vulnerability and potential loss. We also show that for a large class of risks, only a small fraction of the expected loss should be invested. Building on these results, we study a network of interconnected agents subject to epidemic risks. We derive conditions to ensure that the incentives of all agents are aligned towards a better security. When agents are strategic, we show that security investments are always socially inefficient due to the network externalities. Moreover alignment of incentives typically implies a coordination problem, leading to an equilibrium with a very high price of anarchy.

6.4. Point Processes, Stochastic Geometry and Random Geometric Graphs

Participants: François Baccelli, Bartłomiej Błaszczyszyn, Pierre Brémaud, Kumar Gaurav, Mir Omid Haji Mirsadeghi.

stochastic geometry, point process, shot-noise, Boolean model, random tessellation, percolation, stochastic comparison

6.4.1. Modeling, comparison and impact of spatial irregularity of point processes on coverage, percolation, and other characteristics of random geometric models

We develop a general approach for comparison of clustering properties of point processes. It is funded on some basic observations allowing to consider void probabilities and moment measures as two complementary tools for capturing clustering phenomena in point processes. As expected, smaller values of these characteristics indicate less clustering. Also, various global and local functionals of random geometric models driven by point processes admit more or less explicit bounds involving the void probabilities and moment measures, thus allowing to study the impact of clustering of the underlying point process. When stronger tools are needed, dcx ordering of point processes happens to be an appropriate choice, as well as the notion of (positive or negative) association, when comparison to the Poisson point process is concerned. The whole approach has been worked out in a series of papers [62], [63], [64], [65]. This year we have prepared revisions of the two latter ones, from which [65] is now accepted for the publication in Adv. Appl. Probab. We have also prepared a review article [53] for *Lecture Notes in Mathematics*, Springer.

6.4.1.1. 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 λ and μ 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 n and cnin 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 appeared in [27].

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 colloidial systems. In these models, each point (molecule, particle,$\cdots$) 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,$\cdots$) 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 [66] the generating functionals of several models of random packing processes: the classical Matérn hard-core model; its extensions, the k-Matérn models and the ∞-Matérn model, which is an example of random sequential packing process. The main new results are: 1) A sufficient condition for the ∞-Matérn model to be well-defined (unlike the other two, the ∞-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] [60], 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 continued studying a new spatial birth and death point process model where the death rate is a shot noise of the point configuration. 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. A paper on the matter will be presented at Infocom 13.

6.4.5. A population model based on a Poisson line tessellation

In [44], we introduce a new population model. Taking the geometry of cities into account by adding roads, we build a Cox process driven by a Poisson line tessellation. We perform several shot-noise computations according to various generalizations of our original process. This allows us to derive analytical formulas for the uplink coverage probability in each case.

6.4.6. Information Theory and Stochastic Geometry

In a joint work with V. Anantharam [UC Berkeley], we study the Shannon regime for the random displacement of stationary point processes. We currently investigate Multiple Access Channels.

6.4.7. Navigation on Point Processes and Graphs

The thesis of Mir Omid Mirsadeghi [6] studied optimal navigations in wireless networks in terms of first passage percolation on some space-time SINR graph. It established both “positive” and “negative” results on the associated 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: Emilie Coupechoux, Kumar Gaurav, Mathieu Leconte, Marc Lelarge.

random graphs, combinatorial optimization, local weak convergence, diffusion, network games.

6.5.1. Matchings in infinite graphs

In [13] with Charles Bordenave [CNRS-Université de Toulouse] and Justin Salez [Université Paris 7], 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.

6.5.2. Convergence of Multivariate Belief Propagation, with Applications to Cuckoo Hashing and Load Balancing

In [58], with Laurent Massoulié [Inria-MSR], 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 capacitated 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 multiple-choice hashtables. In particular, cuckoo hashing is a popular and simple way to build a hashtable where each item is only allowed to be assigned keys within a predetermined, random subset of all keys. In this context, it is important to determine the load threshold under which cuckoo hashing will succeed with high probability in building such a hashtable. The results on the density of maximum capacitated b-matchings allow to determine this threshold.

6.5.3. A new approach to the orientation of random hypergraphs

A h-uniform hypergraph $H = (V, E)$ is called (l, k)-orientable if there exists an assignment of each hyperedge e to exactly l of its vertices such that no vertex is assigned more than k hyperedges. Let $H_{n,m,h}$ be a hypergraph, drawn uniformly at random from the set of all h-uniform hypergraphs with n vertices and m edges. In [41], we determine the threshold of the existence of a (l, k)-orientation of $H_{n,m,h}$ for $k \geq 1$ and $h > l \geq 1$, extending recent results motivated by applications such as cuckoo hashing or load balancing with guaranteed maximum load. Our proof combines the local weak convergence of sparse graphs and a careful analysis of a Gibbs measure on spanning subgraphs with degree constraints. It allows us to deal with a much broader class than the uniform hypergraphs.

6.5.4. Bipartite graph structures for efficient balancing of heterogeneous loads

In [40], with Laurent Massoulié [Inria-MSR], we look into another application of the results on the asymptotic maximum size of b-matchings 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 [8] with Hamed Amini [EPFL] and 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.

6.5.6. Upper deviations for split times of branching processes

In [9], upper deviation results are obtained for the split time of a supercritical continuous-time Markov branching process. More precisely, with Hamed Amini [EPFL], we establish the existence of logarithmic limits for the likelihood that the split times of the process are greater than an identified value and determine an expression for the limiting quantity. We also give an estimation for the lower deviation probability of the split times which shows that the scaling is completely different from the upper deviations.

6.5.7. Epidemics in random clustered networks

In [54], we study a model of random networks that has both a given degree distribution and a tunable clustering coefficient. We consider two types of growth processes on these graphs: diffusion and symmetric threshold model. The diffusion process is inspired from epidemic models. It is characterized by an infection probability, each neighbor transmitting the epidemic independently. In the symmetric threshold process, the interactions are still local but the propagation rule is governed by a threshold (that might vary among the different nodes). An interesting example of symmetric threshold process is the contagion process, which is inspired by a simple coordination game played on the network. Both types of processes have been used to model spread of new ideas, technologies, viruses or worms and results have been obtained for random graphs with no clustering. In this paper, we are able to analyze the impact of clustering on the growth processes. While clustering inhibits the diffusion process, its impact for the contagion process is more subtle and depends on the connectivity of the graph: in a low connectivity regime, clustering also inhibits the contagion, while in a high connectivity regime, clustering favors the appearance of global cascades but reduces their size. For both diffusion and symmetric threshold models, we characterize conditions under which global cascades are possible and compute their size explicitly, as a function of the degree distribution and the clustering coefficient. Our results are applied to regular or power-law graphs with exponential cutoff and shed new light on the impact of clustering.

6.5.8. Leveraging Side Observations in Stochastic Bandits

The paper [37] considers stochastic bandits with side observations, a model that accounts for both the exploration/exploitation dilemma and relationships between arms. In this setting, after pulling an arm i, the decision maker also observes the rewards for some other actions related to i. We will see that this model is

suited to content recommendation in social networks, where users' reactions may be endorsed or not by their friends. We provide efficient algorithms based on upper confidence bounds (UCBs) to leverage this additional information and derive new bounds improving on standard regret guarantees. We also evaluate these policies in the context of movie recommendation in social networks: experiments on real datasets show substantial learning rate speedups ranging from 2.2x to 14x on dense networks.

6.5.9. Universality in Polytope Phase Transitions and Message Passing Algorithms

In [28], with Mohsen Bayati and Andrea Montanari [Stanford], we consider a class of nonlinear mappings F in R^N indexed by symmetric random matrices A in $R^{N\times N}$ with independent entries. Within spin glass theory, special cases of these mappings correspond to iterating the TAP equations and were studied by Erwin Bolthausen. Within information theory, they are known as 'approximate message passing' algorithms. We study the high-dimensional (large N) behavior of the iterates of F for polynomial functions F, and prove that it is universal, i.e. it depends only on the first two moments of the entries of A, under a subgaussian tail condition. As an application, we prove the universality of a certain phase transition arising in polytope geometry and compressed sensing. This solves -for a broad class of random projections- a conjecture by David Donoho and Jared Tanner.

6.5.10. Far-out Vertices In Weighted Repeated Configuration Model

In [34] we consider an edge-weighted uniform random graph with a given degree sequence (Repeated Configuration Model) which is a useful approximation for many real-world networks. It has been observed that the vertices which are separated from the rest of the graph by a distance exceeding certain threshold play an important role in determining some global properties of the graph like diameter, flooding time etc., in spite of being statistically rare. We give a convergence result for the distribution of the number of such far-out vertices. We also make a conjecture about how this relates to the longest edge of the minimal spanning tree on the graph under consideration.

URBANET Team

6. New Results

6.1. Scalable protocols for capillary networks.

Participants: Ibrahim Amadou, Quentin Lampin, Bilel Romdhani, Alexandre Mouradian, Isabelle Augé-Blum, Fabrice Valois

6.1.1. Beacon-less and opportunistic routing.

During the thesis of Ibrahim Amadou [1], we were focused on the issues of energy in WSNs through energy-efficient routing and medium access control protocols. The contributions of research work can be summarized as follows. First, we were interested on the energy issues at the routing layer for multi-hop wireless sensor networks (WSNs). We proposed a mathematical framework to model and analyze the energy consumption of routing protocols in multi-hop WSNs by taking into account the protocol parameters, the traffic pattern and the network characteristics defined by the medium channel properties, the dynamic topology behavior, the network diameter and the node density. We showed that Beacon-less routing protocol is a good candidate for energy saving in WSNs.

We investigated the performance of some existing relay selection schemes which are used by Beacon-less routing protocols. Extensive simulations were realized in order to evaluate their performance locally in terms of packet delivery ratio, duplicated packet and delay. Then, we extended the work in multi-hop wireless networks and developed an optimal solution, Enhanced Nearest Forwarding within Radius, which tries to minimize the per-hop expected number of retransmissions in order to save energy.

We presented a new Beacon-less routing protocol called Pizza-Forwarding (PF) without any assumption on the radio environment: neither the radio range nor symmetric radio links nor radio properties (shadowing, etc.) are assumed or restricted. A classical greedy mode is proposed. To overcome the hole problem, packets are forwarded to an optimal node in the two hop neighbor following a reactive and optimized neighborhood discovery.

In order to save energy due to idle listening and overhearing, we proposed to combine PF's main concepts with an energy-efficient MAC protocol to provide a joint MAC/routing protocol suitable for a real radio environment. Performance results lead to conclude to the powerful behavior of PF-MAC.

In collaboration with Orange Labs, we designed QOR, an opportunistic routing protocol for wireless sensor networks [16]. QOR first builds a stable directed acyclic logical routing structure and a prefix-based addressing plan stemming from data sinks. This addressing plan is then used to define the potential forwarders set for each source and allows a strict scheduling and an unique selection of the forwarder for each transmission thanks to a cascading acknowledgment scheme. QOR is particularly suited for sensor networks that require high delivery ratio under severe energy constraints. Extensive simulations show the benefits of QOR over an implementation of the IETF routing protocol for Lossy and Low Power networks, RPL, tailored to provide high delivery ratios. Our case studies shows that QOR saves up to 50% energy and reduces the end-to-end delay of a factor of 4 times while maintaining similar delivery ratios.

Most existing routing protocols designed for WSNs assume that links are symmetric, which is in contradiction with what is observed in the field. Indeed, many links in real-world WSNs are asymmetric. Asymmetric links can dramatically decrease the performance of routing algorithms not designed to cope with them. Quite naturally, most existing routing protocol implementations prune the asymmetric links to only use the symmetric ones. In our experience, asymmetric links are a valuable asset to improve network connectivity, capacity and overall performance [20],[2]. We therefore introduced AsymRP (Asymmetric Convergecast Routing Protocol) [21], a new routing protocol for collecting data in WSNs. AsymRP assumes 2-hop neighborhood knowledge and uses implicit and explicit acknowledgment. It takes advantage of asymmetric links to increase delivery ratio while lowering hop count and packet replication.

6.1.2. MAC and cross-layer mechanisms for QoS.

Protocols developed during the last years for Wireless Sensor Networks (WSNs) are mainly focused on energy-consumption optimization and autonomous mechanisms (e.g. self-organization, self-configuration, etc). Nevertheless, with new WSN applications appear new QoS requirements such as time constraints. Real-time applications require the packets to be delivered before a known time bound which depends on the application requirements. We particularly focused on applications which consist in alarms that are sent to the sink node (e.g. air pollution monitoring). We proposed the Real-Time X-layer Protocol (RTXP) [27], a real-time communication protocol that integrates mechanisms for both MAC and routing layers. Our proposal aims at guaranteeing an end-to-end constraint delay, while keeping good performances on other parameters, such as energy consumption. For this purpose the protocol relies on a hop-count-based Virtual Coordinate System (VCS) which classifies nodes having the same hop-count from the sink, allows forwarder selection, and gives to the nodes an unique identifier in a 2-hop neighborhood allowing deterministic medium access. This protocol has better performances than state-of-the-art protocols, in terms of time constraints and reliability, even with unreliable radio links.

In the ARESA2 project, but also in a joint collaboration with Orange Labs, we studied receiver initiated MAC protocol to compare their performance to the more classical receiver-based MAC one [17]. We proposed the Self Adapting Receiver Initiated MAC protocol (SARI-MAC), a novel asynchronous MAC protocol for energy constrained Wireless Sensor Networks. SARI-MAC self-adapts to the traffic load to meet specified Quality of Service requirements at the lowest energy cost possible. To do so, SARI-MAC relies on traffic estimation, duty-cycle adaptation and acknowledgment mechanisms. Our performance evaluation assesses that SARI-MAC meets given QoS requirements in a energy efficient manner and outperforms the state of the art protocol RI-MAC in a broad range of traffic scenarios.

For energy constrained wireless sensor networks, lifetime is a critical issue. Several medium access control protocols have been proposed to address this issue, often at the cost of poor network capacity. To address both capacity and energy issues, we proposed a novel medium sharing protocol for Wireless Sensor Networks named Cascading Tournament (CT-MAC) [15]. CT-MAC is a synchronous, localized, dynamic, joint contention/allocation protocol. Relying on cascading iterations of tournaments, CT-MAC allocates multiple time slots to nodes that compete for accessing the medium. CT-MAC offers an unprecedented trade-off between traffic delay, network capacity and energy efficiency and stands out as a solid candidate for energy constrained sensor networks that must support heterogeneous traffic loads. Our simulations show that CT-MAC significantly outperforms the state-of-the-art SCP- MAC protocol.

6.2. Characterizing urban capillary wireless networks.

Participants: Sandesh Uppoor, Diala Naboulsi, Rodrigue Domga Komguem, Anis Ouni, Alexandre Mouradian, Isabelle Augé-Blum, Hervé Rivano, Marco Fiore, Fabrice Valois

6.2.1. Properties of urban road traffic of interest to mobile networking.

The management of mobility is commonly regarded as one of the most critical issues in large-scale telecommunication networks. The problem is exacerbated when considering vehicular mobility, which is characterized by road-constrained movements, high speeds, sudden changes of movement direction and acceleration, and significant variations of these dynamics over daytime. The understanding of the properties of car movement patterns becomes then paramount to the design and evaluation of network solutions aimed at vehicular environments.

We first analyzed how the vehicular mobility in a large-scale urban region affects a cellular infrastructure intended to support on-board users. We studied the spatial and temporal distribution of traffic load induced by vehicular users, their spatial flows, their inter-arrival and residence times at cells [22].

We then studied the topological features of a network built on moving vehicles, considering the instantaneous connectivity of the system [28]. Our results evidence the spatial and temporal diversity of road traffic, stressing the importance of a correct modeling of road traffic towards the reliable performance evaluation of network

protocols. Additionally, the results outline how commonly adopted assumptions (e.g., Poisson user arrivals at the network base stations) do not hold under vehicular environments, and how the V2V-based network has low connectivity, availability, reliability and navigability properties.

6.2.2. The limits of RSSI-based localization.

Numerous localization protocols in Wireless Sensor Networks are based on Received Signal Strength Indicator. Because absolute positioning is not always available, localization based on RSSI is popular. More, no extra hardware is needed unlike solutions based on infra-red or ultrasonic. Moreover, the theory gives a RSSI as a function of distance. However, using RSSI as a distance metric involves errors in the measured values, resulting path-loss, fading, and shadowing effects. We did experimentation results from three large WSNs, each with up to 250 nodes [23]. Based on our findings from the 3 systems, the relation between RSSI and distance is investigated according to the topology properties and the radio environment. We underline the intrinsic limitations of RSSI as a distance metric, in terms of accuracy and stability. Contrary to what we assumed, collaborative localization protocol based on Spring-Relaxation algorithm can not smooth the distance-estimation errors obtained with RSSI measurements.

6.2.3. Modeling and optimization of wireless networks.

In critical real-time applications, when an event is detected, the Worst Case Traversal Time (WCTT) of the message must be bounded. However, despite this, real-time protocols for WSNs are rarely formally verified. The model checking of WSNs is a challenging problem for several reasons. First, WSNs are usually large scale so it induces state space explosion during the verification. Moreover, wireless communications produce a local broadcast behavior which means that a packet is received only by nodes which are in the radio range of the sender. Finally, the radio link is probabilistic. The modeling of those aspects of the wireless link in model checking is not straightforward and it has to be done in a way that mitigates the state space explosion problem. We are currently working on proposing a methodology adapted to WSNs, and based on Timed Automata (TA) and model-checking. First results are promising [19], but needed to be further investigated.

While the large variety of routing protocols (geographical, gradient, reactive, ...) proposed in the literature provide a set of pertinent solutions for optimizing the energy consumption for multi-hop wireless networks, they do not permit to know the conditions of use of these protocols based on parameters such as: the dynamics of topology, traffic pattern, the density and diameter of the network, the load, etc. In [12], we presented a theoretical model for evaluating the energy consumption for communication protocols taking into account both the dynamics of nodes and links, the properties of topology, the traffic pattern, the control/data packets and a realistic channel model. This model is applied successively to several protocols (GPSR, AODV, OLSR and PF) to highlight their optimum usage and it permits to conclude that Beacon-Less routing protocols are adapted for application with low traffic.

We continued developing optimization tools for building optimal solution to various problems of multi-hop wireless networks. Most of these contributions combine graph theoretical basis with Mixed Integer Linear Programming techniques, and are valuable for understanding the extremal behaviors of the systems and guide the development of efficient architectures and protocols. In this sense, we have considered a new edge coloring problem to model call scheduling optimization issues in wireless mesh networks: the proportional coloring [6]. It consists in finding a minimum cost edge coloring of a graph which preserves the proportion given by the weights associated to each of its edges. We show that deciding if a weighted graph admits a proportional coloring is pseudo-polynomial while determining its proportional chromatic index is NP-hard. We then give lower and upper bounds for this parameter that can be computed in pseudo-polynomial time. We finally identify a class of graphs and a class of weighted graphs for which the proportional chromatic index can be exactly determined.

Dealing with wireless mesh network, we have investigated the fundamental trade-off between transmitting energy consumption and network capacity [24]. The results on this trade-off have been computed using MILP models solved with column generation techniques. The main contribution relies in the ability to consider a realistic SINR model of the physical layer with a continuous power control and discrete transmission rate

selection at each node. In order to model these functionalities, a strong formulation (in the sense that the linear relaxation gives relevant lower bounds) of the rate selection is introduced.

The behavior of beaconless geographic forwarding protocols for wireless sensor networks has also been modeled [9]. A realistic physical layer is taken into account by combining MILP models with simulation based inputs on the number of required retransmissions for realizing a transmission. The model is then able to compute energy efficient routings and allows for understanding the most efficient relay selection schemes, denoted Furthest Forward within Reliable neighbors (FFRe).

6.3. Solutions for cellular networks.

Participants: Anis Ouni, Fabrice Valois, Hervé Rivano, Marco Fiore

6.3.1. Content downloading through a vehicular network.

We considered a system that leverages vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication to transfer large contents to users on-board moving cars. This paradigm is intended to relieve the cellular infrastructure from the high load that such downloads would induce, once vehicles are widely equipped with infotainment devices.

We first characterized the theoretical performance limits of such a vehicular content downloading system by modeling the downloading process as an optimization problem, and maximizing the overall system throughput. Our approach allows us to investigate the impact of different factors, such as the roadside infrastructure deployment, the vehicle-to-vehicle relaying, and the penetration rate of the communication technology, even in presence of large instances of the problem [7]. We then evaluated practical protocols for vehicular downloading, devising solutions for the selection of relay vehicles and data chunks at the Road Side Units (RSUs), and evaluating them in real-world road topologies, under different infrastructure deployment strategies [8].

Our results show that V2V transfers can significantly increase the download rate of vehicular users in urban/suburban environments, and that such a result holds throughout diverse mobility scenarios, RSU placements and network loads. Also, they highlight the existence of two operational regimes at different penetration rates and the importance of an efficient, yet 2-hop constrained, V2V relaying.

6.3.2. Toward green mesh and cellular networks.

On the one hand, a promising technique for minimizing the transmission power of cellular networks seems to be a dramatic densification of micro-cells coverage. On the other hand, increasing the number of micro-cells multiplies the energy consumed by the cells whatever their state, idle, transmitting or receiving. For a sustainable deployment of such micro-cell infrastructures and for a significant decrease of the overall energy consumption, an operator needs to be able to switch off cells when there are not absolutely needed. The densification of the cells induces the need for an autonomic control of the on/off state of cells. This has motivated a preliminary investigation on exploiting within the micro-cellular settings the manifold results of duty cycles for Wireless Sensor Networks where switching nodes on and off is done in a distributed or localized manner while coverage and connectivity properties are maintained [29].

Focusing on broadband wireless mesh networks based on OFDMA resource management, and considering a realistic SINR model of the physical layer with a continuous power control and discrete transmission rate selection at each node, we have investigated the trade-off between transmission energy consumption and network capacity [24]. 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. We also highlight that power control and rate selection are not expandable to an optimal system configuration.

6.4. Miscellaneous security issues in capillary networks.

Participants: Ochirkhand Erdene-Ochir, Fabrice Valois, Marco Fiore

6.4.1. Resiliency in routing protocols.

Within the ARESA2 project, we defined the notion of resiliency for routing protocols in wireless sensor networks and we applied it to several routing strategies to provide an understandable taxonomy [3]. Efforts have been made to compare routing protocols according to their resiliency in wireless multi-hop sensor networks in the presence of packet dropping malicious insiders. In [13], we proposed a new taxonomy of routing protocols obtained by applying our resiliency metric. Several resiliency enhancing methods such as introducing a random behavior to the classical routing protocols and a new data replication method based on the distance information have been evaluated as well. Simulation results demonstrate the effectiveness of the proposed approach.

6.4.2. Verifying the positions announced by mobile nodes.

A growing number of ad hoc networking protocols and location-aware services require that mobile nodes learn the position of their neighbors. However, such a process can be easily abused or disrupted by adversarial nodes. In absence of a-priori trusted nodes, the discovery and verification of neighbor positions presents challenges that have been scarcely investigated in the literature.

We proposed a fully-distributed cooperative solution that is robust against independent and colluding adversaries. Results show that our protocol can thwart more than 99% of the attacks under the best possible conditions for the adversaries, with minimal false positive rates [5].

A centralized solution was also developed, that leverages anonymous position beacons from vehicles, and the cooperation of nearby cars collecting and reporting the beacons they hear. Such information allows an authority to verify the locations announced by vehicles, or to infer the actual ones if needed [18].