RESEARCH CENTER

FIELD
Networks, Systems and Services, Distributed Computing

Activity Report 2013

Section New Results

Edition: 2014-03-20

ACES Project-Team

5. New Results

5.1. Self-describing objects

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

Coupled objects enable basic integrity checking for physical objects, and use cases were demonstrated for security and logistics applications. In these applications, high reliability in the RFID reading infrastructure is assumed for the system to work. This suggest another idea for coupled objects: using control data structures distributed over the physical objects in order to improve the reliability of RFID reading protocols. This is the purpose of the Pervasive_RFID project, in collaboration with the IETR which is described in more details below 7.1.2 .

Another development in the line of the coupled objects principles are self-describing objects. While previous works enabled integrity checking over a set of physical objects, these mechanisms were limited in two aspects: expressiveness and autonomy. More precisely, coupled objects support the detection of special conditions (such as a missing element), but not the characterization of these conditions (such as describing the problem, identifying the missing element). Moreover, this compromises the autonomous feature of coupled objects, which would depend on external systems for analyzing these special conditions. Self-describing objects are an attempt to overcome these limitations, and to broaden the application perspectives of autonomous RFID systems.

The principle is to implement distributed data structure over a set of RFID tags, enabling a complex object (made of various parts) or a set of objects belonging to a given logical group to "self-describe" itself and the relation between the various physical elements. Some applications examples includes waste management, assembling and repair assistance, prevention of hazards in situations where various products / materials are combined etc. The key property of self-describing objects is, like for coupled objects, that the vital data are self-"hosted" by the physical element themselves (typically in RFID chips), not an external infrastructure like most RFID systems. This property provides the same advantages as in coupled objects, namely high scalability, easy deployment (no interoperability dependence/interference), and limited risk for privacy.

However, given the extreme storage limitation of RFID chips, designing such systems is difficult:

- data structures must be very frugal in terms of space requirements, both for the structure and for the coding.
- Data structures must be robust and able to survive missing or corrupted elements if we want to ensure the self-describing property for a damaged or incorrect object.

An application of self-describing objects has been proposed in for waste management, in the context of the bin that think project 7.1.1 . A generic graph structure applicable to RFID systems for supporting self-describing objects is proposed in Arnab Sinha's thesis document (to be defended in April 2014).

5.2. Pervasive support for Smart Homes

Participants: Andrey Boytsov, Michele Dominici, Bastien Pietropaoli, Sylvain Roche, Frederic 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 *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.

5.2.1. A multi-level context computing architecture

Computing context is a major subject of interest in smart spaces such as smart homes. Contextual data are necessary for services to adapt themselves to the context and to be as efficient as possible. Contextual data may be obtained via augmented appliances capable of communicating their state and a bunch of sensors. It becomes more and more real with the development of the Internet of Things. Unfortunately, the gathered data are not always directly usable to understand what is going on and to build services on them. In order to address this issue, we studied a multi-level context computing architecture divided in four layers:

- *Exploitation layer*: the highest layer, it exploits con- textual data to provide adapted services
- *Context and situation identification layer*: this is what analyzes ongoing situations and potentially predicts future situations
- *Perception layer*: it offers a first layer of abstraction for small pieces of context independent of deployed sensors
- *Sensing layer*: it mainly consists of the data gathered by sensors

In this architecture, every layer is based on the results of its underlying layers. In 2013, we studied several methods that enable the building of such levels of abstractions (see figure 2). The first level of abstraction coming to mind when describing what people are doing in a Home is high level abstractions such as "cooking". Those activities are then the highest level abstraction we want our system to be able to identify.

We proposed to use plan recognition algorithms to analyze sequences of actions and thus predict future actions of users. It is, in our case, adapted to identify ongoing activities and predict future ones. There exist different plan recognition algorithms. However, one interested us particularly, PHATT introduced by Goldman, Geib and Miller. In order to understand how PHATT is working, it is important to understand the hierarchical task network (HTN) planning problem which is "inverted" by the algorithm to perform plan recognition. It consists in automatically generating a plan starting from a set of tasks to execute and some constraints. In our case, we are able to predict future situations depending of the previously observed situations. To give an example, if we want to predict that the situation dinner will occur soon, it is sufficient to have observed situations such as cooking and/or setting the table. The performances of PHATT have been evaluated by Andrey Andrey Boytstov and Frédéric Weis. These results will be published in 2014.

5.2.2. Propagation of BFT

Context-aware applications have to sense the environment in order to adapt themselves and provide with contextual services. This is the case of Smart Homes equipped with sensors and augmented appliances. However, sensors can be numerous, heterogeneous and unreliable. Thus the data fusion is complex and requires a solid theory to handle those problems. For this purpose, we adopted the belief functions theory (BFT). The aim of the data fusion, in our case, is to compute small pieces of context we call context attributes. Those context attributes are diverse and could be for example the presence in a room, the number of people in a room or even that someone may be sleeping in a room. Since the BFT requires a substantial amount of computations, we proposed to reduce as much as possible the number of evidence required to compute a context attribute. Moreover, the number of possible worlds, *i.e.* the number of possible states for a context attribute, is also an

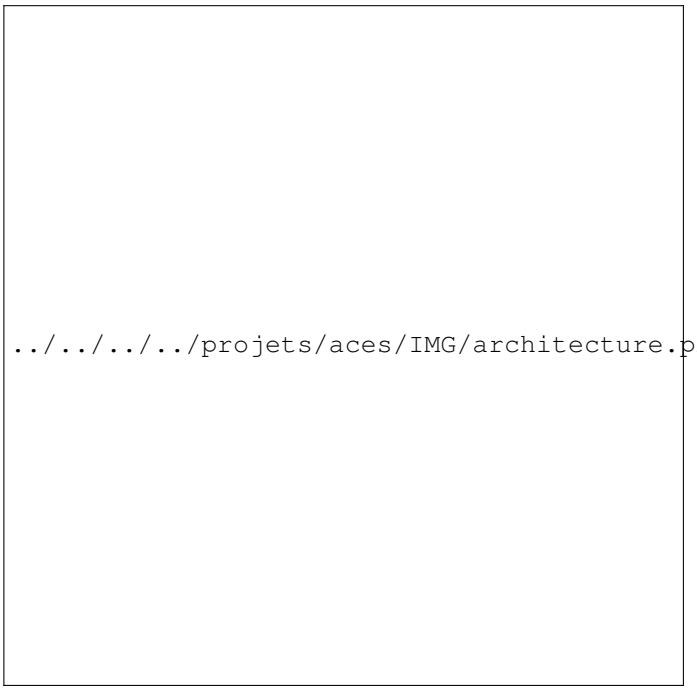

Figure 2. Multi-level context computing architecture

important source of computation. Thus, reducing the number of possible worlds we are working on is also important.

It is especially problematic when working on embedded systems, which may be the case when trying to observe context in smart homes. Thus, with this objective in mind, we observed that some context attributes could be used to compute others. By doing this, the number of gathered and combined evidence for each context attribute could be drastically reduced. This principle is illustrated by Figure 3 : the sets of possible worlds for "Presence" and "Posture" are seen as subsets of "Sleeping". So we proposed and implemented a method to propagate BFT through a set of possible states for a context attribute.

Figure 3. Propagation of Belief Functions Theories

5.2.3. Definition of virtual sensors

In our multi-level architecture, the sensor measures may be imperfect for multiple reasons. The most annoying reasons when deploying a system are biases and noisy measures. It requires fine tuning each type the system is deployed in a new environment. In order to prevent from doing this work again and again at levels where models are hard to build, we proposed to add a new sublayer to the sensing layer (see Figure 2): virtual sensors. Instead of modifying high level models, we created sensor abstractions such as motion sensor, sound sensor, temperature sensor, etc. It is particularly convenient when working with typed data such as temperature or sound level. It is possible to use different brands of sensors for sensors of the same type. Thus, those sensors, even if they are measuring the same physical event, can return very different data due to their range, sensibility, voltage, etc. By creating abstraction of sensors, it is possible to build models directly from typed data simplifying even more the building of models as those data have are understandable by humans. Those virtual sensors are built very simply from common heuristics and can be used for ias and noise compensation, Data aggregation and Meta-data generation.

It is also possible in these virtual sensors to implement fault and failure detection mechanisms using the BFT. It enables the detection of fault in the case of sensors of the same type. At higher level, those mechanisms will detect inconsistency between sensors of different types which is not of the same utility. Thus, those virtual virtual sensors, without disabling any features in our architecture, bring more stability for our models. Moreover, by keeping the virtual sensors very simple, they are easy to adapt and tune in a new environment and the overhead in terms of computation is reduced to the minimum and does not really impact the global system performance. Finally, the fine tuning part is always reduced to this level of our architecture and nothing else has to be changed when we move the system from one environment to another.

ADAM Project-Team

6. New Results

6.1. Self-Adaptive Software Systems

Participants: Russel Nzekwa, Romain Rouvoy [correspondant], Lionel Seinturier.

The design of self-adaptive and autonomic software systems raises many challenges. In his PhD thesis, Russel Nzekwa [12] proposes a new result with the CORONA framework that enables to build flexible autonomic systems. CORONA relies on an architectural description language which reifies the structure of the control system architecture. CORONA enables the flexible integration of non-functional-properties during the design of autonomic systems. It also provides tools for checking conflicts in the architecture of autonomic systems. Finally, the traceability between the design and the runtime implementation is carried out through the code generation of skeletons from architectural descriptions of control systems. The work on CORONA goes toward the long term objective of setting up an integrated design and programming solution for self-adaptive systems, where feedback control loops play the central role as first class elements.

6.2. Energy Management in Software Systems

Participants: Rémi Druilhe, Laurence Duchien, Lionel Seinturier [correspondant].

Energy management and saving is a concern that spans the entire domain of information and communication technologies and sciences. Recently it has been recognized that to improve its efficiency, energy has to be managed, not only at the hardware level, but also at the level of software systems, especially in distributed environments. In his PhD thesis, Rémi Druilhe [11] proposes a new result with the HOMENAP system for networked digital home environments. This work is the result of a collaboration with Orange Labs. HOMENAP takes into account three main properties: heterogeneity, dynamicity and quality of service. HOMENAP proposes an autonomic decision-making system to deal with the placement of digital services on networked devices. Based on the observation of relevant events, the system takes the decision to modify the distribution of digital services on devices in order to preserve a defined tradeoff between energy efficiency and quality of service. HOMENAP participates to the long term objective of dealing with energy as a main steering factor for self-optimizing software systems.

6.3. Automated Software Repair

Participant: Martin Monperrus [correspondant].

Automated software repair aims at assisting developers in order to improve the quality of software systems, for example by recommending some repair actions to fix bugs. In [15], we present some major results in this direction by mining fix transactions of existing software repositories. From the empirical study of 14 software repositories containing 89,993 versioning transactions, we show that we can learn a probability distribution of repair actions. We show that certain distributions over repair actions can result in an infinite time (in average) to find a repair shape while other fine-tuned distributions enable to find a repair shape in hundreds of repair attempts. We now aim at going beyond this empirical study and theoretical analysis by exploring how to use this learned knowledge for new software systems.

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 2013 have in particular accounted for the increasingly connected networking environment, as envisioned by the Future Internet, and further focused on one of its major components that is the Internet of Things, which allows connecting the physical with the digital word. In more detail, our research has 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);
- Revisiting service-oriented computing toward the Future Internet, in particular dealing with the composition of highly heterogeneous services while ensuring quality of service (§ 6.3);
- Service oriented middleware for the ultra large scale future mobile Internet of Things (§ 6.4);
- Abstractions for enabling domain experts to easily compose applications on the Internet of Things (§ 6.5);
- Lightweight streaming middleware for the Internet of Things (§ 6.6); and
- Dynamic decision networks for decision-making in self-adaptive systems (§ 6.7).

6.2. Emergent Middleware

Participants: Emil Andriescu, Amel Bennaceur, Valérie Issarny.

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. At the application layer, components may exhibit disparate data types and operations, and may have distinct business logics. At the middleware layer, they may rely on different communication standards, which define disparate data representation formats and induce different architectural constraints. Finally, at the network layer, data may be encapsulated differently according to the network technology in place.

A wide range of approaches have thus been proposed to address the interoperability challenge, as surveyed in [26]. However, solutions that require performing changes to the systems are usually not feasible since the systems to be integrated may be built by third parties (e.g., COTS —Commercial Off-The-Shelf— components or legacy systems); no more appropriate are 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; correct mediators guarantee that the components interact without errors (e.g., deadlocks) and reach their termination successfully. Therefore, formal approaches are used to synthesize mediators automatically.

We posit that interoperability should neither be achieved by defining yet another middleware nor yet another ontology but rather by exploiting existing middleware together with knowledge encoded in existing domain ontologies to synthesize and implement mediators automatically. In [2], we have introduced the notion of *emergent middleware* for realizing mediators, which was initiated as part of the FP7 FET IP CONNECT project. Our work during the year 2013 has more specifically focused on the further elaboration of a comprehensive approach to mediator synthesis, including dealing with interoperability across protocol layers.

Mediator synthesis for emergent middleware: We focus on functionally-compatible components, i.e., components that at some high level of abstraction require and provide compatible functionalities, but are unable to interact successfully due to mismatching interfaces and behaviors. To address these differences without changing the components, mediators that systematically enforce interoperability between functionally-compatible components by mapping their interfaces and coordinating their behaviors are required. Our approach for the automated synthesis of mediators is performed in several steps.

The first step is interface matching, which identifies the semantic correspondence between the actions required by one component and those provided by the other. We incorporate the use of ontology reasoning within constraint solvers, by defining an encoding of the ontology relations using arithmetic operators supported by widespread solvers, and use it to perform interface matching efficiently. For each identified correspondence, we generate an associated matching process that performs the necessary translations between the actions of the two components' interfaces. The second step is the synthesis of correct-by-construction mediators. To do so, we analyze the behaviors of components so as to generate the mediator that combines the matching processes in a way that guarantees that the two components progress and reach their final states without errors. The synthesised mediator is the most general component that ensures freedom of both communication mismatches and deadlock in the composition of the components [15]. The last step consists in making the synthesized mediator concrete by incorporating all the details about the interaction of components. To do so, we compute the translation functions necessary to reconcile the differences in the syntax of the input/output data used by each component and coordinate the different interaction patterns that can be used by middleware solutions.

We refer the interested reader to [7] for a complete description of the approach. Our contribution primarily lies in handling interoperability from the application to the middleware layer in an integrated way. The mediators we synthesize act as: (i) translators by ensuring the meaningful exchange of information between components, (ii) controllers by coordinating the behaviors of the components to ensure the absence of errors in their interaction, and (iii) middleware by enabling the interaction of components across the network so that each component receives the data it expects at the right moment and in the right format.

Automated mediation for cross-layer protocol interoperability: Existing approaches to interoperability are restricted to solving either application heterogeneity when the underlying middlewares are compatible, or solving middleware heterogeneity at each protocol layer separately. In real world scenarios, this does not suffice: application and middleware boundaries are ill-defined and solutions to interoperability must consider them in conjunction. We have been studying the case of cross-layer interoperability where protocol mediation is performed between protocol stacks, rather than between protocol layers separately. Such interoperability approaches are appropriate for systems that rely on complex protocol stacks, where application and middleware layers are tightly coupled.

Systems relying on tightly coupled protocol stacks exchange complex messages that consist of a composition of heterogeneous data formats. To enable interoperation, complex messages from one system must be translated into a different complex format that another system accepts such that the two can interact. While Off-The-Shelf and third party message parsers are widely available for simple message formats (i.e., message formats corresponding to a single protocol layer), complex message formats are typically unique since they are the result of a protocol binding. Protocol binding represents the connection between one protocol and another to create a new communication flow. Some middleware protocols recommend or restrict to certain types of default binding (e.g., HTTP provides an extensive set of rules for binding, such as Content-Encoding and Content-Type). However, real systems are often designed following a custom binding mechanism, restricting the application of automated mediation solutions. This problem occurs primarily because complex message formats cannot be easily interpreted.

Many solutions address this composition issue by introducing Domain Specific Languages that can be used by experts to specify parsers for complex message formats. Yet, whenever messages have a more complicated syntax, providing their DSL descriptions becomes difficult as well. Further, such approaches are not future proof as more protocols are expected to emerge, which will not be accounted for by DSLs that are defined according to known message formats. An alternative is to generate parsers based on the composition of third-party parsers that are usually included with protocol implementations. However, third-party parsers cannot be used unless the protocol binding rules are identified by an expert, further allowing to implement the bridge between one parser's output data and the other parser's input data. To this end, we designed an approach for generating composed parsers that can process complex messages, accompanied by a formal mechanism for defining complex message formats based on existing data formats. Our approach relies on user-provided parser composition rules, which reflect the binding requirements of complex message formats.

We posit that our method is more efficient than implementing complex parsers, defining them using DSLs, or directly implementing the binding of protocols. Furthermore, with this solution, we support the automated synthesis of mediators at the application layer using the mapping-based approach discussed above, by automatically generating an abstract representation of the application data exchanged by the interoperating components.

6.3. Service-oriented Computing in the Future Internet

Participants: Georgios Bouloukakis, Nikolaos Georgantas, Valérie Issarny, Ajay Kattepur.

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

Extensible Service Bus for the Future Internet: XSB is an abstract service bus that deals effectively with the cross-integration of heterogeneous interaction paradigms [17]. Inside the XSB, the CS, PS and TS paradigms are modeled as abstract base connectors. Their *space coupling* semantics are represented with programming interfaces used by applications (APIs) and corresponding application interface description languages (IDLs). Their behavioral semantics are formally specified in terms of LTS (Labeled Transition Systems). We formally verify the correctness of these behavioral specifications with respect to *time coupling* and *concurrency* properties expressed in LTL temporal logic. This allows stating the correctness of the connector models with respect to the semantics that they must have. This further enables identifying the behavioral semantics of the XSB connector derived from the interconnection of base connectors. More specifically, in order to identify the time coupling and concurrency semantics of XSB and construct a converter among the base connectors, we build upon the formal method of *protocol conversion via projections*[10]. According to this method, conversion between two different protocols is possible if both protocols can be projected (where projection is an abstraction defined as a set of transformations on the protocol LTS) to a *functionally sufficient* common *image protocol*. Then, the end-to-end protocol of the interconnection of the two protocols is this image protocol.

We have implemented our XSB solution into an extensible development and execution platform for application and middleware designers. Using this platform, they can easily develop composite applications: they only need to build descriptions for the constituent services and directives for data mapping among them. Our platform then deals with reconciling among the heterogeneous interaction paradigms and protocols of the services by employing *binding components (BCs)* that adapt between the native middleware of the services and the XSB bus protocol. The XSB itself is implemented on top of an existing ESB substrate. Support for new middleware platforms, new ESB substrates, or even new interaction paradigms can be incorporated in a facilitated way thanks to the provided XSB architectural framework.

[10] Lam, S.S.: Protocol Conversion. IEEE Trans. Softw. Eng. 14(3) (1988) 353–362.

QoS composition and analysis of heterogeneous choreographies: Leveraging on the functional interoperability across interaction paradigms offered by the XSB, we study the Quality of Service (QoS) performance of choreographies [21]. QoS dependency plays an important role in the service oriented system lifecycle, including discovery, runtime selection, replacement and contractual guarantees. Consequently, QoS composition among choreographed devices should tackle multi-dimensional probabilistic metrics combined with message passing constraints imposed at design-time. We make use of an algebraic QoS composition model that is applied at the interaction paradigm level to study the composition of QoS metrics, and the subsequent tradeoffs. While traditional QoS composition analysis has been done purely at the application level, analyzing the effect of middleware interactions allows us to study CS, PS and TS based device compositions. This produces interesting insights such as selection of a particular system and its middleware during design-time, or end-to-end QoS expectation/guarantees during runtime. Our formulation also allows for runtime reconfiguration, in order to optimally produce design time QoS expectations. Such flexible reconfiguration policies are crucial in the case of large scale choreographies with high variability in runtime performance of participating devices.

Further, we study the effect of time/space coupling on the latency of successful transactions across the XSB connector [20]. XSB models the message passing among peers through generic `post` and `get` operations, that represent peer behavior with both tight (CS) and loose (PS/TS) time/space coupling. The heterogeneous *lease* and *timeout* behaviors of these operations severely affect latency and success rates of messages passed either synchronously or through callbacks. By precisely studying the timing thresholds using timed automata models, we verify conditions for accurate message transactions with XSB connectors. This offers choreography designers the ability to set these timing thresholds (bottom-up) or select a particular interaction paradigm (top-down) for runtime enaction.

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

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

The Internet of Things (IoT) is characterized by an increasing number of Things embedding sensing, actuating, processing, and communication capacities. A considerable portion of those Things will be *mobile* Things, which come with several advantages yet lead to unprecedented challenges. The most critical challenges, that are directly inherited from, yet amplify, today's Internet issues, lie in handling i) the large scale of users and mobile Things, ii) providing interoperability across the heterogeneous Things, and iii) overcoming the unknown dynamic nature of the environment, due to the mobility of an ultra-large number of Things.

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

In accordance with SOA, MobIoT comprises *Discovery*, *Composition & Estimation*, and *Access* components, yet modifies their internal functionalities. In more detail, the Discovery component enables Thing-based service registration (for Things to advertise hosted services) and look-up (for Things to retrieve remote services of interest). In order to handle the ultra large number of mobile Things and their services in the IoT, the component revisits the Service-Oriented discovery and introduces *probabilistic discovery* to provide, not *all*, but only a sufficient *subset of services that can best approximate* the result that is being sought after [18], [11]. Furthermore, the Composition & Estimation component (C&E) provides automatic composition of Thing-based services. This capacity is of interest in the case where no service can perform a required measurement/action task directly (based on its atomic functionalities). Thing-based service composition executes in three phases: i) *expansion*, where composition specifications are automatically identified; ii) *mapping*, where actual service instances (running services) are selected based on their functionalities and the physical attributes of their hosts; and iii) *execution*, where the services are accessed and the composition specifications are executed.

Thing-based service composition revisits Service-Oriented composition by executing seamlessly with no involvement from developers or end users. Last but not least, the Access component provides an easy to use interface for developers to sample sensors/actuators while abstracting sensor/actuator hardware specifications. Additionally, it revisits Service-Oriented access by executing access to services transparently and wrapping access functionalities internally. Thus, it alleviates that burden from users, initially in charge of this task. The Access component supports access to remote services and to locally hosted services.

6.5. Composing Applications in the Internet of Things

Participants: Aness Bajia, Pankesh Patel, Animesh Pathak, Françoise Sailhan.

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. An ideal application development abstraction of the IoT will allow (domain expert) developers to intuitively specify the rich interactions between the extremely large number of disparate devices in the future Internet of Things. The goal of our research is then to propose a suitable application development framework, where our work this year covered the two 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 [22]. 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 evaluation based on two realistic scenarios shows that the use of our techniques/framework succeeds in improving productivity in the IoT application development process. More details of our approach can be found in [8].

Integrating support for non-functional requirements while programming IoT applications: Given that devices and networks constituting the IoT are prone to failure and consequent loss of performance, it is natural that IoT applications are expected to encounter and tolerate several classes of faults - something that still largely remains within the purview of low-level-protocol designers. As part of our work on the MURPHY project (§ 7.1.1.1), we are addressing this issue by proposing: i) a set of abstractions that can be used during macroprogramming to express fault tolerance requirements, and ii) a runtime system that employs adaptive fault tolerance (AFT) to provide fault tolerance to the sensing application. Complementary to this, we have proposed task mapping algorithms to satisfy those requirements through a constraint programming approach [19]. Through evaluations on realistic application task graphs, we show that our constraint programming model can effectively capture the end-to-end requirements and efficiently solves the combinatorial problem introduced.

We have continually incorporating our research results in the above areas into *Srijan* (§ 5.6), 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. Lightweight Streaming Middleware for the Internet of Things

Participants: Benjamin Billet, Valérie Issarny.

The Internet of Things (IoT) is a promising concept toward pervasive computing as it may radically change the way people interact with the physical world. One of the challenges raised by the IoT is the in-network continuous processing of data streams presented by Things, which must be investigated urgently because it affects the future data models of the IoT. This cross-cutting concern has been previously studied in the context of Wireless Sensor and Actuator Networks (WSAN) given the focus on the acquisition and in-network processing of sensed data. However, proposed solutions feature heterogeneous technologies that are difficult to integrate and complex to use, which represents a hurdle to their wide deployment. In addition, new types of smart sensors are emerging due to technological advances (e.g., Oracle SunSpot), enabling the implementation of complex processing tasks directly into the network, without using proxies or sending every data to the cloud. There is thus a need for a distributed middleware solution for data stream management that leverages existing WSAN work, while integrating it with today's Web technologies in order to improve the flexibility and the interoperability of the future IoT. Toward that goal, we have been developing Dioptase, a Data Stream Management System for the IoT, which aims to integrate the Things and their streams into today's Web by presenting sensors and actuators as services. The middleware specifically provides a way to describe complex fully-distributed stream-based mashups and to deploy them dynamically, at any time, as task graphs, over available Things of the network, including resource-constrained ones. To this end, Dioptase enables task graphs to be composed of Thing-specific tasks (directly implemented on the Thing) and dynamic tasks that communicates using data streams. Dynamic tasks are then described in a lightweight DSL, which is directly interpreted by the middleware and provides specific primitives to manipulate data streams.

As part of the design of Dioptase, we have been investigating dedicated task mapping. Task mapping, which basically consists of mapping a set of tasks onto a set of nodes, is a well-known problem in distributed computing research. However, as a particular case of distributed systems, the Internet of Things (IoT) poses a set of renewed challenges, because of its scale, heterogeneity and properties traditionally associated with WSAN, shared sensing, continous processing of data streams and real time computing. To handle IoT features, we present a formalization of the task mapping problem that captures the varying consumption of resources and various constraints (location, capabilities, QoS) in order to compute a mapping that guarantees the lifetime of the concurrent tasks inside the network and the fair allocation of tasks among the nodes (load balancing). It results in a binary programming problem for which we provide an efficient heuristic that allows its resolution in polynomial time. Our experiments show that our heuristic: (i) gives solutions that are close to optimal and (ii) can be implemented on reasonably powerful Things and performed directly within the network, without requiring any centralized infrastructure.

6.7. Dynamic Decision Networks for Self-Adaptive Systems

Participants: Amel Belaggoun, Nelly Bencomo, Valérie Issarny, Peter Sawyer.

Different modeling techniques have been used to model requirements and decision-making of self-adaptive systems [25]. Important successful techniques based on goal models have been prolific in supporting decision-making according to partial and total fulfillment of functional (goals) and non-functional requirements (softgoals). The final decision about what strategy to use is based on a utility function that takes into account the weighted sum of the different effects of the non-functional requirements. Such solutions have been used both at design and run time including our own solutions using runtime goal models. Different modeling techniques have been used to model requirements and decision-making of self-adaptive systems [25]. Important successful techniques based on goal models have been prolific in supporting decision-making according to partial and total fulfillment of functional (goals) and non-functional requirements (softgoals). The final decision about what strategy to use is based on a utility function that takes into account the weighted sum of the different effects of the non-functional requirements. Such solutions have been used both at design- and run-time including our own solutions using runtime goal models.

We have enriched the decision-making supported by goal models with the use of Bayesian Dynamic Decision Networks (DDNs) [12]. Our novel approach supports reasoning about partial satisfaction of soft-goals using probabilities and uses machine learning. When using DDNs, we introduce new ways to tackle uncertainty based on probabilities that can be updated based on runtime evidence. We have reported the results of the

application of the approach on two different cases, one of them being the case of dynamic reconfiguration of a remote data mirroring network that must spread data among servers while minimizing costs and loss of data. Our early results suggest the decision-making process of self-adaptive systems can be improved by using DDNs.

This work has been developed under the umbrella of the Marie Curie Project Requirements@run.time (§ 7.2.1.4). The main results achieved during the year 2013 are:

- A Bayesian-based technique to support the decision making of self-adaptive systems [14]. DDN-based approaches adopt probabilistic methods (i.e., Bayesian methods) and decision theory to assess the consequences of uncertainty. Using the approach, suitable choices to satisfice functional requirements of the system are identified from a range of alternative decisions and their expected utilities. Satisficement of NFRs is modeled using conditional probabilities given the design decisions. Preferences over decisions are modeled using weights associated with pairs of design alternatives and NFRs, and used when computing the expected utilities of the architectural design alternatives. The decision taken by the DDN is that with the highest expected utility. The approach offers the benefits of machine learning.
- A formal Bayesian definition of surprise as the basis for quantitative analysis to measure degrees of uncertainty and deviation of self-adaptive systems from normal behavior [13]. Specifically, a Bayesian surprise quantifies how new evidence affects assumptions of the world (properties in the models). A “surprising" event may provoke a large divergence between the beliefs distributions prior and posterior to that event. As such and depending on how big or small this divergence is, the running system may decide to either: (i) dynamically adapt accordingly, or (ii) temporarily avoid any action of adaptation and flag up the fact that a potential abnormal situation has been found. While doing (ii) we are offering a specific implementation of the RELAX language previously developed by Bencomo and her co-authors.

LOGNET Team

6. New Results

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

Participants: Giang Ngo Hoang [contact], Luigi Liquori, Hung Nguyen Chan [VIELINA, Vietnam].

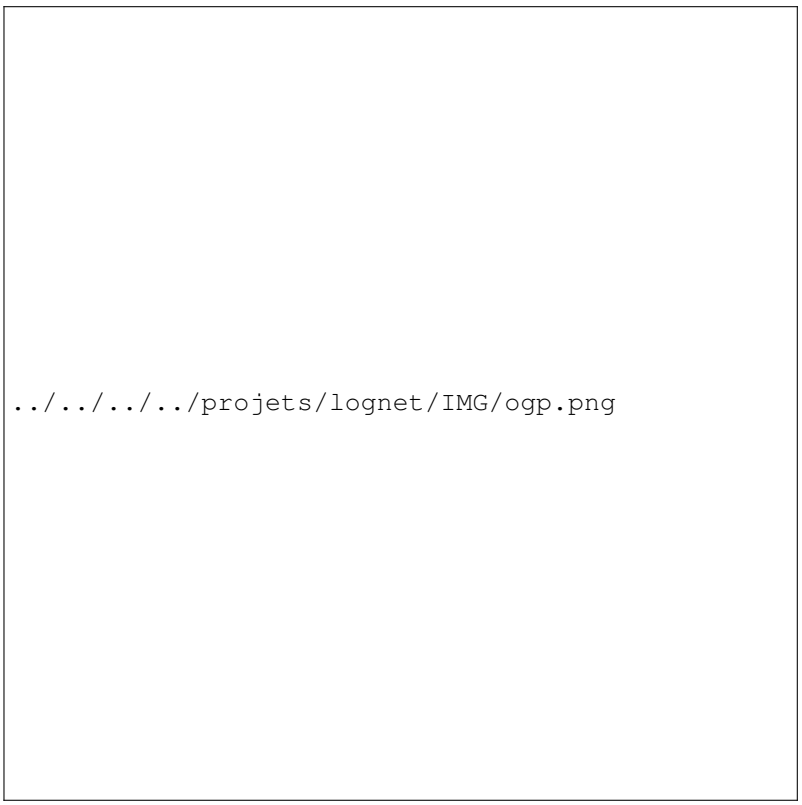

Figure 11. An Overlay Gateway Protocol Topology

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 Hets-Nets 2012 [8]: it has been also accepted to ACM SAC 2013 [36] and a long version has been accepted to the International Conference ICDCN 2014 [32].

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

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

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 [11] and the full version has been accepted to the conference [30].

6.3. 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 [12] and a journal version in [26].

6.4. 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 [31] and [28] and a full version has been invited and will appear as book chapter to [33].

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

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

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.

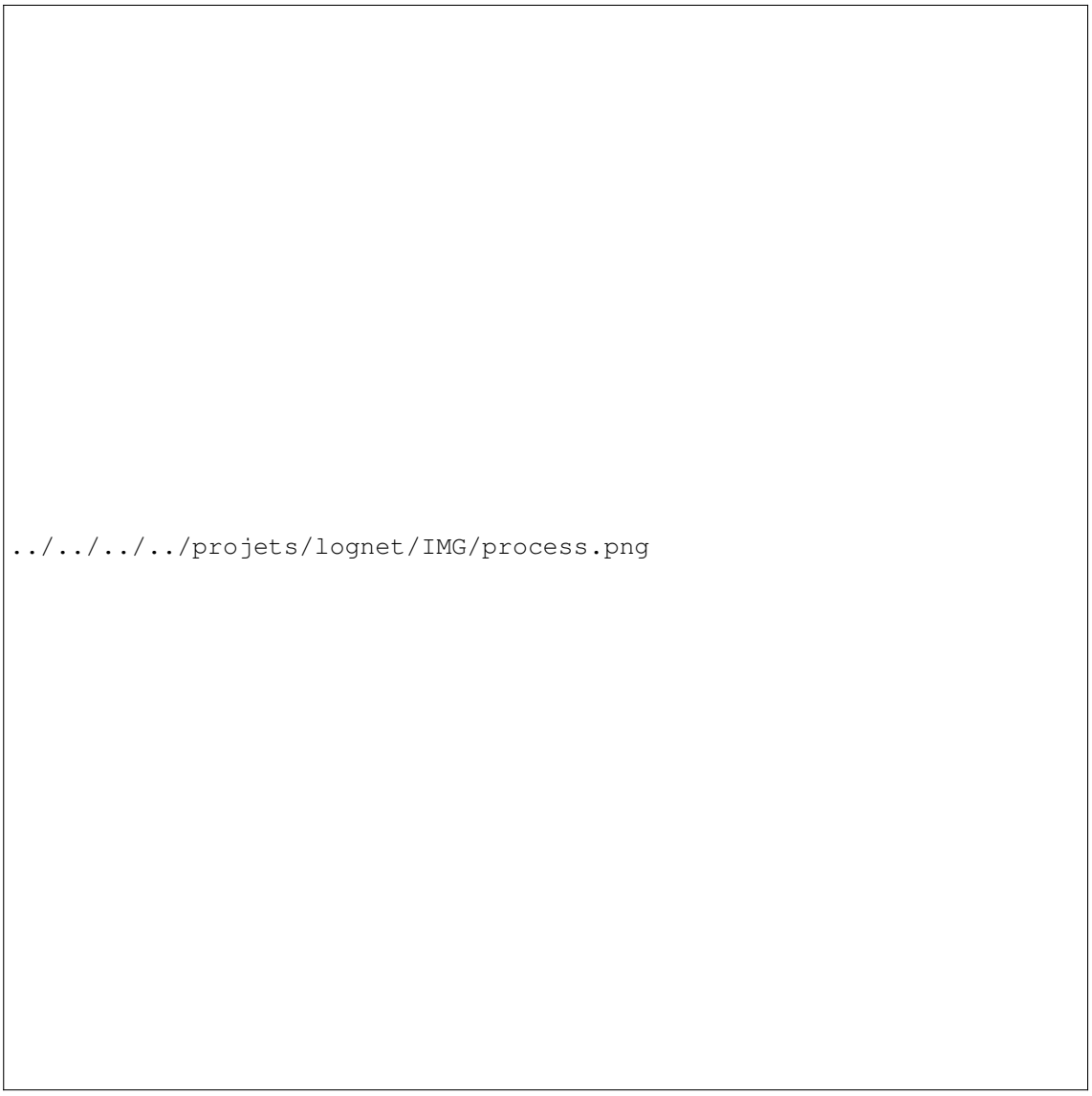

Figure 12. Process of designing a robust trust and reputation system

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 [22] and an improved software and a full paper are in preparation in 2014.

6.6. A Scalable Communication Architecture for Advanced Metering Infrastructure

Participants: Giang Ngo Hoang [contact], Luigi Liquori, Hung Nguyen Chan [VIELINA, Vietnam].

Advanced Metering Infrastructure (AMI), seen as foundation for overall grid modernization, is an integration of many technologies that provides an intelligent connection between consumers and system operators. One of the biggest challenge that AMI faces is to scalable collect and manage a huge amount of data from a large number of customers. In our paper, we address this challenge by introducing a mixed peer-to-peer (P2P) and client-server communication architecture for AMI in which metering data is aggregated and processed distributedly at multiple levels and in a tree-like manner. Through analysis we show that the architecture is featured with load scalability, resiliency with failure and partly self-organization. The

experiments performed in large scale French Grid5000 platform [G5k] shows the communication efficiency in the proposed architecture. A technical report will be submitted to an international conference [37].

6.7. 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 .

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. This work has been published in the ACM workshops [13] and [29] and a long version has been invited and appear in the Journal of Logic and Computation [27].

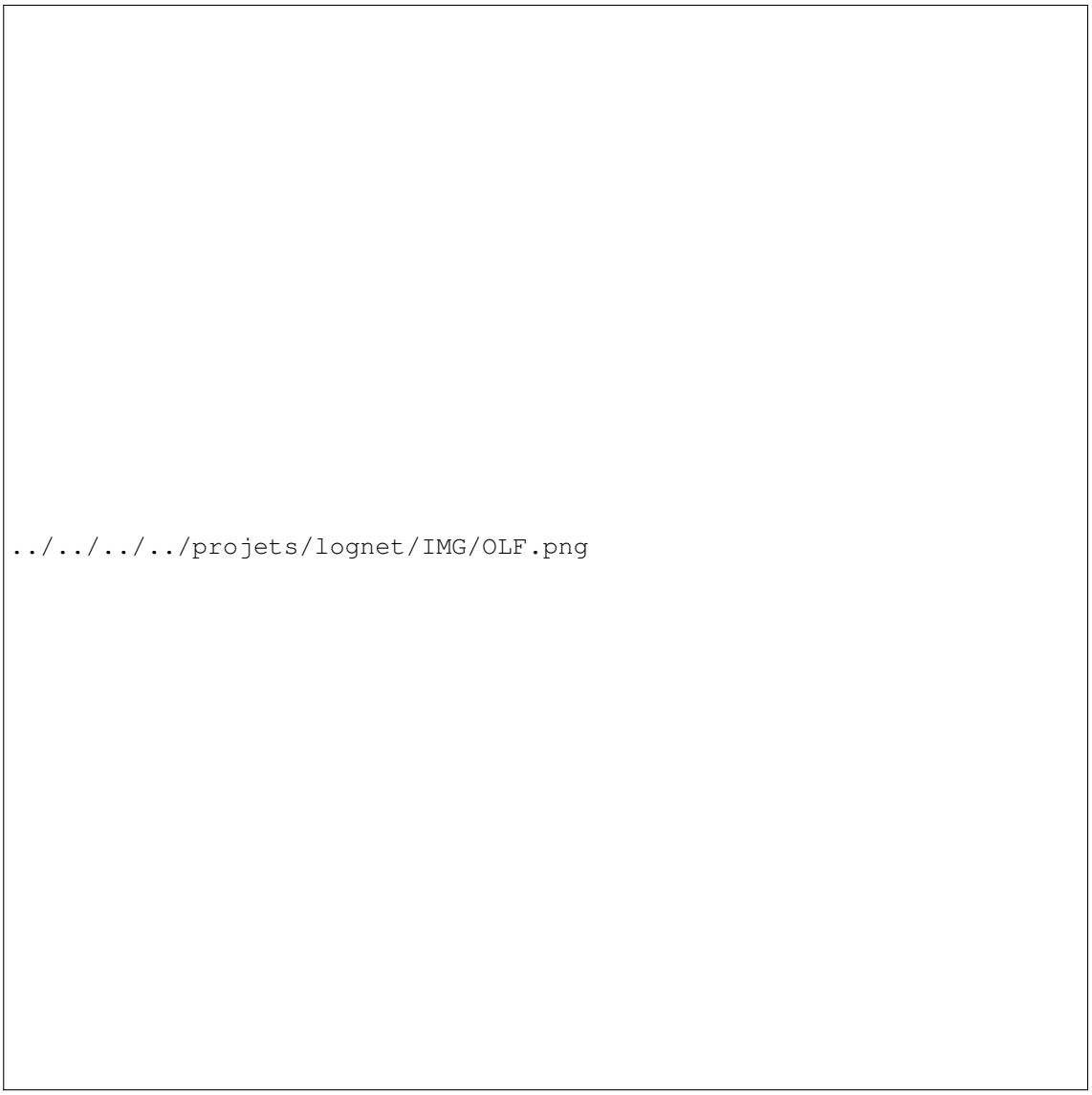

Figure 13. Some rule of the Open Logical Framework

ASAP Project-Team

6. New Results

6.1. Models and abstractions for distributed systems

6.1.1. Randomized loose renaming in O(loglog n) time

Participant: George Giakkoupis.

Renaming is a classic distributed coordination task in which a set of processes must pick distinct identifiers from a small namespace. In [24], we consider the time complexity of this problem when the namespace is linear in the number of participants, a variant known as loose renaming. We give a non-adaptive algorithm with $O(\log\log n)$ (individual) step complexity, where n is a known upper bound on contention, and an adaptive algorithm with step complexity $O((\log\log k)^2)$, where k is the actual contention in the execution. We also present a variant of the adaptive algorithm which requires $O(k\log\log k)$*total* process steps. All upper bounds hold with high probability against a strong adaptive adversary. We complement the algorithms with an $\Omega(\log\log n)$ expected time lower bound on the complexity of randomized renaming using test-and-set operations and linear space. The result is based on a new coupling technique, and is the first to apply to non-adaptive randomized renaming. Since our algorithms use $O(n)$ test-and-set objects, our results provide matching bounds on the cost of loose renaming in this setting.

This work was done in collaboration with Dan Alistarh, James Aspnes, and Philipp Woelfel.

6.1.2. An O(sqrt n) space bound for obstruction-free leader election

Participant: George Giakkoupis.

In [32] we present a deterministic obstruction-free implementation of leader election from $O(\sqrt{n})$ atomic $O(\log n)$-bit registers in the standard asynchronous shared memory system with n processes. We provide also a technique to transform any deterministic obstruction-free algorithm, in which any process can finish if it runs for b steps without interference, into a randomized wait-free algorithm for the oblivious adversary, in which the expected step complexity is polynomial in n and b. This transformation allows us to combine our obstruction-free algorithm with the leader election algorithm by Giakkoupis and Woelfel (2012), to obtain a fast randomized leader election (and thus test-and-set) implementation from $O(\sqrt{n})O(\log n)$-bit registers, that has expected step complexity $O(\log^* n)$ against the oblivious adversary. Our algorithm provides the first sub-linear space upper bound for obstruction-free leader election. A lower bound of $\Omega(\log n)$ has been known since 1989 (Styer and Peterson, 1989). Our research is also motivated by the long-standing open problem whether there is an obstruction-free consensus algorithm which uses fewer than n registers.

This work was done in collaboration with Maryam Helmi, Lisa Higham, and Philipp Woelfel.

6.1.3. Broadcast in recurrent dynamic systems

Participants: Michel Raynal, Julien Stainer.

This work [50] proposes a simple broadcast algorithm suited to dynamic systems where links can repeatedly appear and disappear. The algorithm is proved correct and a simple improvement is introduced, that reduces the number and the size of control messages. As it extends in a simple way a classical network traversal algorithm (due to A. Segall, 1983) to the dynamic context, the proposed algorithm has also pedagogical flavor.

This work has been done in collaboration with Jiannong Cao and Weigang Wu.

6.1.4. Computing in the presence of concurrent solo executions

Participants: Michel Raynal, Julien Stainer.

In a wait-free model any number of processes may crash. A process runs solo when it computes its local output without receiving any information from other processes, either because they crashed or they are too slow. While in wait-free shared-memory models at most one process may run solo in an execution, any number of processes may have to run solo in an asynchronous wait-free message-passing model. This work [47] is on the computability power of models in which several processes may concurrently run solo. We introduced a family of round-based wait-free models, called the d-solo models, $1 \leq d \leq n$, where up to d processes may run solo. Then we gave a characterization of the colorless tasks that can be solved in each d-solo model. We also introduced the (d, ϵ)-solo approximate agreement task, which generalizes ϵ-approximate agreement, and proves that (d, ϵ)-solo approximate agreement can be solved in the d-solo model, but cannot be solved in the $(d + 1)$-solo model. We also studied the relation linking d-set agreement and (d, ϵ)-solo approximate agreement in asynchronous wait-free message-passing systems. These results establish for the first time a hierarchy of wait-free models that, while weaker than the basic read/write model, are nevertheless strong enough to solve non-trivial tasks.

This work was done in collaboration with Maurice Herlihy and Sergio Rajsbaum.

6.1.5. Relating message-adversaries and failure detectors

Participants: Michel Raynal, Julien Stainer.

A message adversary is a daemon that suppresses messages in round-based message-passing synchronous systems in which no process crashes. A property imposed on a message adversary defines a subset of messages that cannot be eliminated by the adversary. It has recently been shown that when a message adversary is constrained by a property denoted TOUR (for tournament), the corresponding synchronous system and the asynchronous crash-prone read/write system have the same computability power for task solvability. In this work [39] we introduced new message adversary properties (denoted SOURCE and QUORUM), and shown that the synchronous round-based systems whose adversaries are constrained by these properties are characterizations of classical asynchronous crash-prone systems (1) in which processes communicate through atomic read/write registers or point-to-point message-passing, and (2) enriched with failure detectors such as Ω and Σ. Hence these properties characterize maximal adversaries, in the sense that they define strongest message adversaries equating classical asynchronous crash-prone systems. They consequently provide strong relations linking round-based synchrony weakened by message adversaries with asynchrony restricted with failure detectors. This not only enriches our understanding of the synchrony/asynchrony duality, but also allows for the establishment of a meaningful hierarchy of property-constrained message adversaries.

6.1.6. A hierarchy of agreement problems from simultaneous consensus to set agreement

Participants: Michel Raynal, Julien Stainer.

In this work [38] we investigated the relation linking the s-simultaneous consensus problem and the k-set agreement problem in wait-free message-passing systems. To this end, we defined the (s, k)-SSA problem which captures jointly both problems: each process proposes a value, executes s simultaneous instances of a k-set agreement algorithm, and has to decide a value so that no more than sk different values are decided. We also introduced a new failure detector class denoted $Z_{s,k}$, which is made up of two components, one focused on the "shared memory object" that allows the processes to cooperate, and the other focused on the liveness of (s, k)-SSA algorithms. A novelty of this failure detector lies in the fact that the definition of its two components are intimately related. We designed a $Z_{s,k}$-based algorithm that solves the (s, k)-SSA problem, and shown that the "shared memory"-oriented part of $Z_{s,k}$ is necessary to solve the (s, k)-SSA problem (this generalizes and refines a previous result that showed that the generalized quorum failure detector Σ_k is necessary to solve k-set agreement). We finally, investigated the structure of the family of (s, k)-SSA problems and introduced generalized (asymmetric) simultaneous set agreement problems in which the parameter k can differ in each underlying k-set agreement instance. Among other points, it shows that, for $s, k > 1$, (a) the $(sk, 1)$-SSA problem is strictly stronger that the (s, k)-SSA problem which is itself strictly stronger than the $(1, ks)$-SSA problem, and (b) there are pairs (s_1, k_1) and (s_2, k_2) such that $s_1k_1 = s_2k_2$ and (s_1, k_1)-SSA and (s_2, k_2)-SSA are incomparable.

6.2. Large-scale and user-centric distributed systems

6.2.1. FreeRec: An anonymous and distributed personalization architecture

Participants: Antoine Boutet, Davide Frey, Arnaud Jégou, Anne-Marie Kermarrec, Heverson Borba Ribeiro.

FreeRec is an anonymous decentralized peer-to-peer architecture designed to bring personalization while protecting the privacy of its users [17], [30], [44]. FreeRec's decentralized approach makes it independent of any entity wishing to collect personal data about users. At the same time, its onion-routing-like gossip-based overlay protocols effectively hide the association between users and their interest profiles without affecting the quality of personalization. The core of FreeRec consists of three layers of overlay protocols: the bottom layer, rps, consists of a standard random peer sampling protocol ensuring connectivity; the middle layer, PRPS, introduces anonymity by hiding users behind anonymous proxy chains, providing mutual anonymity; finally, the top clustering layer identifies for each anonymous user, a set of anonymous nearest neighbors. We demonstrate the effectiveness of FreeRec by building a decentralized and anonymous content dissemination system. Our evaluation by simulation, our PlanetLab experiments, and our probabilistic analysis show that FreeRec effectively decouples users from their profiles without hampering the quality of personalized content delivery.

6.2.2. HyRec: A hybrid recommender system

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

The ever-growing amount of data available on the Internet calls for personalization. Yet, the most effective personalization schemes, such as those based on collaborative filtering (CF), are notoriously resource greedy. HyRec is an online cost-effective scalable system for CF personalization. HyRec relies on a hybrid architecture, offloading CPU-intensive recommendation tasks to front-end client browsers, while retaining storage and orchestration tasks within back-end servers. HyRec has been fully implemented and extensively evaluated on several workloads from MovieLens and Digg. We convey the ability of HyRec to significantly reduce the operation costs of the content provider by up to 70% and drastically improve the scalability by up to 500%, with respect to a centralized (or cloud-based recommender approach), while preserving the quality of the personalization. We also show that HyRec is virtually transparent to the users and induces only 3% of the bandwidth consumption of a P2P solution.

6.2.3. Social market

Participants: Davide Frey, Arnaud Jégou, 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) solves 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 ala 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. In addition to our previous publication on this topic, our recent work led to a paper that appeared in TCS [20].

6.2.4. Privacy-preserving P2P collaborative filtering

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

The huge amount of information available at any time in our connected society calls for a mechanism to filter it efficiently. Recommendation systems provide such a mechanism by personalizing the information displayed for each user. However, the collection of personal information by recommendation systems threatens the privacy of users. We address the two needs for recommendation and privacy through a peer-to-peer user-based collaborative filtering system. Recommendation is done ala GOSSPLE by building an overlay network which connects users with similar interests via clustering and random peer sampling. This overlay network is then used to make recommendations based on what similar users liked. Users' privacy is protected in two ways. Users are protected from a Big Brother adversary by the peer-to-peer design of the system in which profiles are stored only by their owners. Users are protected from other malicious users who would try to learn the content of their profiles by our landmark-based cosine similarity measure. It indirectly computes the similarity of two users by comparing their respective similarities with a set of randomly generated profiles, called landmarks. Thus, users can compute their similarity without revealing their profile, contrarily to the regular cosine similarity when used in a peer-to-peer system.

6.2.5. Gossip protocols for renaming and sorting

Participants: George Giakkoupis, Anne-Marie Kermarrec.

In [33] we devise efficient gossip-based protocols for some fundamental distributed tasks. The protocols assume an n-node network supporting point-to-point communication, and in every round, each node exchanges information of size $O(\log n)$ bits with (at most) one other node. We first consider the *renaming* problem, that is, to assign distinct IDs from a small ID space to all nodes of the network. We propose a renaming protocol that divides the ID space among nodes using a natural push or pull approach, achieving logarithmic round complexity with ID space $\{1, \cdots, (1+\epsilon)n\}$, for any fixed $\epsilon > 0$. A variant of this protocol solves the *tight* renaming problem, where each node obtains a unique ID in $\{1, \cdots, n\}$, in $O(\log^2 n)$ rounds. Next we study the following *sorting* problem. Nodes have consecutive IDs 1 up to n, and they receive numerical values as inputs. They then have to exchange those inputs so that in the end the input of rank k is located at the node with ID k. Jelasity and Kermarrec (2006) suggested a simple and natural protocol, where nodes exchange values with peers chosen uniformly at random, but it is not hard to see that this protocol requires $\Omega(n)$ rounds. We prove that the same protocol works in $O(\log^2 n)$ rounds if peers are chosen according to a non-uniform power law distribution.

This work has been done in collaboration with Philipp Woelfel.

6.2.6. Adaptive streaming

Participants: Ali Gouta, Anne-Marie Kermarrec.

HTTP Adaptive Streaming (HAS) is gradually being adopted by Over The Top (OTT) content providers. In HAS, a wide range of video bitrates of the same video content are made available over the internet so that clients' players pick the video bitrate that best fit their bandwidth. Yet, this affects the performance of some major components of the video delivery chain, namely CDNs or transparent caches since several versions of the same content compete to be cached. We investigated the benefits of a Cache Friendly HAS system (CF-DASH), which aims to improve the caching efficiency in mobile networks and to sustain the quality of experience of mobile clients. We presented a set of observations we made on large number of clients requesting HAS contents [34], [35]. Then, we evaluated CF-dash based on trace-driven simulations and testbed experiments. Our validation results are promising. Simulations on real HAS traffic show that we achieve a significant gain in hit-ratio that ranges from 15% up to 50%.

Work was done in collaboration with Yannick Le Louedec, Zied Aouini and Diallo Mamadou.

6.2.7. DynaSoRe: Efficient in-memory store for social applications

Participant: Arnaud Jégou.

Social network applications are inherently interactive, creating a requirement for processing user requests fast. To enable fast responses to user requests, social network applications typically rely on large banks of cache servers to hold and serve most of their content from the cache. The objective of this work is to build a memory cache system for social network applications that optimizes data locality while placing user views across the system. We call this system DynaSoRe (Dynamic Social stoRe). DynaSoRe storage servers monitor access traffic and bring data frequently accessed together closer in the system to reduce the processing load across cache servers and network devices. Our simulation results considering realistic data center topologies show that DynaSoRe is able to adapt to traffic changes, increase data locality, and balance the load across the system. The traffic handled by the top tier of the network connecting servers drops by 94% compared to a static assignment of views to cache servers while requiring only 30% additional memory capacity compared to the whole volume of cached data.

This work was conducted in collaboration with Xiao Bai, Flavio Junqueira, and Vincent Leroy. The product of this collaboration led to the publication of a paper at the Middleware 2013 conference [26].

6.2.8. Adaptive metrics on distributed recommendation systems

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

Current distributed recommendation systems are metric based. This means that recommendation quality depends on a single user comparison function. This is a simple solution that cannot cover the particularities of each system. Classically computing intensive data-mining methods have been used in the field of recommendation. However, they are not proper in distributed scenarios due to the lack of a global vision and the existing restrictions in terms of computing power. In this project, we study how to provide and model ad-hoc similarity metrics that can be automatically adapted to a different number of scenarios. We study our solution from two different points of view: recommendation and performance. In the first, we evaluate the capacity of data mining technics to give users relevant recommendations. Second, by exploring the performance of different approaches in order to obtain relevant recommendations we plan to study the trade-off between relevant recommendations and computational cost.

6.2.9. Cliff-Edge Consensus: Agreeing on the precipice

Participants: Michel Raynal, François Taïani.

In this project, we worked on a new form of consensus that allows nodes to agree locally on the extent of crashed regions in networks of arbitrary size. One key property of our algorithm is that it shows local complexity, i.e. its cost is independent of the size of the complete system, and only depends on the shape and extent of the crashed region to be agreed upon. In [40], we motivate the need for such an algorithm, formally define this new consensus problem, propose a fault-tolerant solution, and prove its correctness.

This work was done in collaboration with Geoff Coulson and Barry Porter.

6.2.10. Clustered network coding

Participants: Fabien André, Anne-Marie Kermarrec, Konstantinos Kloudas, Alexandre Van Kempen.

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

In this work, we propose a novel erasure code system, tailored for networked archival systems. The efficiency of our approach relies on a combination of the use of random codes coupled with a clever yet simple clustered placement strategy. Our repair protocol leverages network coding techniques to reduce by 50% the amount of data transferred during maintenance, as several cluster files are repaired simultaneously. We demonstrate both through an analysis and extensive experimental study conducted on a public testbed that our approach dramatically decreases both the bandwidth overhead during the maintenance process and the time to repair data lost upon failure.

This has been done in collaboration with Erwan le Merrer, Nicolas, Le Scouarnec and Gilles Straub.

ATLANMOD Project-Team

6. New Results

6.1. 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 this year, Atlanmod has continued working actively on this research area. The main contributions are the following:

- In the context of the ARTIST FP7 project, the work has started on reusing (and extending accordingly) MoDisco and several of its components to provide the Reverse Engineering support required within the project. More particularly, the MoDisco Model Discovery + Model Understanding two-step approach is being promoted as an important part of the ARTIST migration methodology and process [35] [19]. Work has also been performed, in the context of the TEAP FUI project dealing with Enterprise Architecture, on how to design and implement a model driven federation approach from heterogeneous data sources (e.g. Excel files, databases, etc.) directly inspiring from these same MoDisco principles [20].
- 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 different components of the system, therefore appropriate techniques must be provided for the discovery and evolution of changing business rules. In [39], [25], [26], we describe a MDRE framework and tool aiming at extracting business rules out of COBOL source code. In [27], we describe a Model-based process and tool to extract business rules, expressed as OCL integrity constraints, from relational databases. In these works, 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. A thesis financed by IBM to advance the research on this topic has been completed this year
- In a web context, JSON has become a very popular lightweigth format for data exchange. JSON is human readable and easy for computers to parse and use. However, JSON is schemaless. Though this brings some benefits (e.g. flexibility in the representation of the data) it can become a problem when consuming and integrating data from different JSON services since developers need to be aware of the structure of the schemaless data. We believe that a mechanism to discover (and visualize) the implicit schema of the JSON data would largely facilitate the creation and usage of JSON services. For instance, this would help developers to understand the links between a set of services belonging to the same domain or API. In this sense, we have proposed a model-based approach to generate the underlying schema of a set of JSON documents [22].

6.2. Security

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

In 2013, we have tackled the aforementioned problems with respect to three key information system components: networks of firewalls, relational database systems and content management systems.

- Firewalls are a key element in network security. They are in charge of filtering the traffic of the network in compliance with a number of access-control rules that enforce a given security policy. In [33] we have described a model-driven reverse engineering approach able to extract the security policy implemented by a set of firewalls in a working network, easing the understanding, analysis and evolution of network security policies. In [17] we have extended this method to cope with a more complex and specific scenario, i.e, the management of stateful packet filtering.
- A similar approach have been successfully used to extract AC information from relational database systems. Concretely, in [32] we contribute a security metamodel and a reverse engineering process that combines standard database access-control rules with the fine-grained access control provided by triggers and stored procedures. The extraction of this comprehensive model helps security experts to visualize and manipulate database security policies in a vendor-independent manner.
- Out-of-the-box Web Content Management Systems (WCMSs) are the tool of choice for the development of millions of enterprise web sites. However, little attention has been brought to the analysis of how developers use the content protection mechanisms provided by WCMSs, in particular, Access-control (AC). We have proposed in [34] a metamodel tailored to the representation of WCMS AC policies, easing the analysis and manipulation tasks by abstracting from vendor-specific details.

6.3. Collaborative development

In the field of Domain-Specific Languages (DSLs), we have focused on the improvement of the DSLs definition process. When developing DSMLs, the participation of end-users is normally limited to providing domain knowledge and testing the resulting language prototypes. Language developers, which are perhaps not domain experts, are therefore in control of the language development and evolution. This may cause misinterpretations which hamper the development process and the quality of the DSML. Thus, it would be beneficial to promote a more active participation of end-users in the development process of DSMLs. While current DSML workbenches are mono-user and designed for technical experts, we have presented a process and tool support for the example-driven, collaborative construction of DSMLs based on Collaboro in order to engage end-users in the creation of their own languages [23], [24].

6.4. MDE Scalability

As Model-Driven Engineering (MDE) is increasingly applied to larger and more complex systems, additional research and development is imperative in order to enable MDE to remain relevant with industrial practice. In [31] we attempt to provide a research roadmap for scalability in MDE and outline directions for work in this emerging research area. As a first result in this roadmap, in [37] we show that rule-based languages like ATL have strong parallelization properties. Parallelization is indeed one of the traditional ways of making computation systems scalable. We describe the implementation of a parallel transformation engine for the current version of the ATL language and experimentally evaluate the consequent gain in scalability. Finally in [28] we compare the improved scalability of the ATL transformation engine with other engines in the community by addressing the task of generating and analyzing very large flow graphs.

6.5. Model Quality

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

- A benchmark that facilitates the comparison between the plethora of tools that provide some kind of quality assurance for models. Similarly to what it is done in many other domains, a common set of test benchmarks that new tools can rely on to experiment and evaluate themselves could speed up the advance in the field. Our proposal can be found [30]

- Validation of the feasibility to apply this kind of techniques in industrial settings based on two case studies [12] and [36]
- Advanced on the verification of model transformations using SMT solvers (instead of SAT or CSP-based approaches commonly used before), with some encouraging results [21] and, related to this, [13]
- A method to build models using instance-level information in terms of examples and counterexamples (gathering requirements using these instance scenarios is usually better from a stakeholder's point of view than trying to explain us general rules about the business). So far existing approaches have often focused on the generation of static models from such instance-level information but have omitted the inference of OCL business rules that could complement the static models and improve the precision of the software specification. We propose an approach to automating such inference [29]. The basic idea is based on an incorporation of the problem solving mechanism and getting user feedback: Candidates are generated by a problem solving, and irrelevant ones are eliminated using the user feedback on generated counterexamples and examples. Our approach is realized with the support tool InferOCL and has been applied on several user cases, indicating a possibility to apply this solution prototype in practice.

CIDRE Project-Team

6. New Results

6.1. Intrusion Detection

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

In 2013, we continue to strengthen our research efforts around intrusion detection parameterized by a security policy.

In [33], we have proposed a language for specifying and composing fine-grained information flow policies. The language used a XML-syntax and has a formal semantic. BSPL enables to precisely specify the expected behavior of applications relatively to their sensitive pieces of information. More precisely it permits to specify where a piece of data owned by an application is allowed to disseminate: in which files or processes.

In [25], we have experimented the previous language (BSPL). We have developed a policy manager for android devices. The manager is able to check the consistency of a policy and to compose two consistent policies. We have also proposed a semi-automatic method for computing information flow policies of applications. We have thus computed some examples of policies and shown that these policies are rich enough to permit benign execution of an application without raising useless alerts and sufficiently restrictive to detect malicious actions induced by a malware.

In [40], we have proposed a new data-structure called System Flow Graph (or SFG in short) that offers a compact representation of how pieces of data flow inside a system. For a given application, the system flow graph describes its external behavior. We have shown that this new data structure suits to represent malware behavior and permits to give an early diagnostic in case of intrusion.

In [36] we have collaborated with Mathieu Jaume from Université de Paris 6 describes a formal framework to draw a correspondence between two types of policy definitions - policies that are defined by properties over states of a system and those that are described by properties over executions of a system.

In [34] and in C.Hauser's PhD desertion, we have extended previous work on kBlare (an IDS that detect illegal flows of information at the kernel level) so as to follow information flows at the network level. To that end, a set of nodes administrated by a single entity can be configured according to a distributed security policy expressed in terms of legal information flows. The different operating systems (kBlare) at each node cooperate by tagging each network packet with a tag that describes the information content of the payload. This way, it is possible to detect illegal information flow of information at the network level. This can be used to detect attacks against confidentiality or integrity of the overall system.

6.1.2. Terminating-Insensitive Non-Interference Verification based on an Information Flow Control

In 2010-2011, we started an informal collaboration with colleagues from CEA LIST laboratory. In 2012, this collaboration has turn into a reality by the funding of a PhD student (Mounir Assaf). This PhD thesis is about the verification of security properties of programs written in an imperative language with pointer aliasing (a subset of C language) by techniques borrowed from the domain of static analysis. One of the property of interest for the security field is called Terminating-Insensitive Non-Interference. Briefly speaking, when verified by a program, this property ensures that the content of any secret variable can not leak into public ones (for any terminating execution). However, this property is too strict in the sense that a large number of programs although perfectly secure are rejected by classical analyzers.

In 2013, Mounir Assaf has studied novel approaches that combine static and dynamic information flow monitoring. These approaches are promising since they enable permissive (accepting a large subset of executions) yet sound (rejecting all insecure executions) enforcement of non-interference. We have investigated a dynamic information flow monitor for a language supporting pointers. Our flow-sensitive monitor relies on prior static analysis in order to soundly enforce non-interference. We have also proposed a program transformation that preserves the behavior of initial programs and soundly inlines our security monitor. This program transformation enables both dynamic and static verification of non-interference in a language supporting pointers. This work has been published in [27] and [45].

6.1.3. Visualization of Security Events

The studies that were performed last year clearly showed that there was an important need for technologies that would allow analysts to handle in a consistent way the various types of log files that they have to study in order to detect intrusion or to perform forensic analysis. Consequently, we proposed this year ELVis, a security-oriented log visualization system that allows the analyst to import its log files and to obtain automatically a relevant representation of their content based on the type of the fields they are made of. First, a summary view is proposed. This summary displays in an adequate manner each field according to its type (i.e. categorical, ordinal, geographical, etc.). Then, the analyst can select one or more fields to obtain some details about it. A relevant representation is then automatically selected by the tool according to the types of the fields that were selected.

ELVis [35] has been presented in VizSec 2013 (part of Vis 2013) in October in Atlanta. A working prototype is currently being tuned in order to perform field trials with our partners in DGA-MI. Next year, we are planing to perform research on how various log files can be combined in the same representation. In the PANOPTESEC project, we will also perform some research on visualization for security monitoring in the context of SCADA systems.

6.2. Privacy

6.2.1. Geoprivacy

With the advent of GPS-equipped devices, a massive amount of location data is being collected, raising the issue of the privacy risks incurred by the individuals whose movements are recorded. In [31], we focus on a specific inference attack called the de-anonymization attack, by which an adversary tries to infer the identity of a particular individual behind a set of mobility traces. More specifically, we propose an implementation of this attack based on a mobility model called Mobility Markov Chain (MMC). A MMC is built out from the mobility traces observed during the training phase and is used to perform the attack during the testing phase. We design several distance metrics quantifying the closeness between two MMCs and combine these distances to build de-anonymizers that can re-identify users in an anonymized geolocated dataset. Experiments conducted on real datasets demonstrate that the attack is both accurate and resilient to sanitization mechanisms such as downsampling. This paper has received the IEEE best student paper award at the conference TrustCom 2013.

In [30], we propose to adopt the MapReduce paradigm in order to be able to perform a privacy analysis on large scale geolocated datasets composed of millions of mobility traces. More precisely, we design and implement a complete MapReduce-based approach to GEPETO. GEPETO (for GEoPrivacy-Enhancing TOolkit) is a flexible software that can be used to visualize, sanitize, perform inference attacks and measure the utility of a particular geolocated dataset. The main objective of GEPETO is to enable a data curator (e.g., a company, a governmental agency or a data protection authority) to design, tune, experiment and evaluate various sanitization algorithms and inference attacks as well as visualizing the following results and evaluating the resulting trade-off between privacy and utility. Most of the algorithms used to conduct an inference attack (such as sampling, k-means and DJ-Cluster) represent good candidates to be abstracted in the MapReduce formalism. These algorithms have been implemented with Hadoop and evaluated on a real dataset. Preliminary results show that the MapReduced versions of the algorithms can efficiently handle millions of mobility traces.

6.2.2. Privacy-enhanced Social Networks

In [38], we have proposed a systematic methodology for evaluating the quality of the privacy proposed by a social networking platform. It is based on an analysis grid organizing a correspondence between a number of design features and properties having an impact on privacy, and a level of distribution. For each property, we consider three possible distribution levels: centralized, decentralized and fully decentralized. For security properties, in particular, we have defined those distribution levels with the help of three different attacker models: an attacker has the ability to compromise either one entity in the system, a pre-defined subset of entites in the system, or the whole set of peers in the system. We argument on the idea that the more powerful the attacker model needed to compromise a property for all users in the system, the higher the privacy level linked to this property. A formal evaluation tool based on lattice structures is then proposed to compare social network systems based on this analysis grid. An example evaluation is also provided, with the thorough analysis of several well-known systems of various kinds, notably leading to the conclusion that some privacy-oriented social networking architectures, presented by their authors as fully distributed, showed centralized characteristics for many privacy-related properties.

6.2.3. Privacy Enhancing Technologies

The development of NFC-enabled smartphones has paved the way to new applications such as mobile payment (m-payment) and mobile ticketing (m-ticketing). However, often the privacy of users of such services is either not taken into account or based on simple pseudonyms, which does not offer strong privacy properties such as the unlinkability of transactions and minimal information leakage. In [26], we introduce a lightweight privacy-preserving contactless transport service that uses the SIM card as a secure element. Our implementation of this service uses a group signature protocol in which costly cryptographic operations are delegated to the mobile phone.

6.2.4. Privacy and Web Services

We have proposed [55] a new model of security policy based for a first part on our previous works in information flow policy and for a second part on a model of Myers and Liskov. This new model of information flow serves web services security and allows a user to precisely define where its own sensitive pieces of data are allowed to flow through the definition of an information flow policy. A novel feature of such policy is that they can be dynamically updated, which is fundamental in the context of web services that allow the dynamic discovery of services. We have also presented an implementation of this model in a web services orchestration in BPEL (Business Process Execution Language).

6.2.5. Privacy-preserving Ad-hoc Routing

6.2.5.1. Proactive Protocol

In [39], we have proposed a *proactive* ad hoc routing protocol that preserves the anonymity of the source and of the destination of the packet flows, and assures the unlinkability of flows between any pair of participants to local observers and to global attackers to a lesser extend. Our solution is based on OLSR and combines Bloom filters and ephemeral identifiers. More specifically, the routing process allows any node to discover the topology of the ad hoc network. Once such a topology is known, a source node can establish beforehand a path to reach any destination node. To conceal the identity of the source and destination nodes, the path may not be the shortest ones nor terminate at the destination node. Then, by including the ephemeral public identifiers of the intermediate nodes into a Bloom filter, the source node is able to specify the nodes that have to rebroadcast packets. Thus, when receiving a packet, a node has simply to check, using its ephemeral private identifier, whether it has to rebroadcast the packet, without knowing the source, the destination, nor the previous and next hop.

6.2.5.2. Reactive Protocol

In [42], we have proposed a classification of privacy preserving properties that ensure privacy in ad hoc network routing. We also proposed a taxonomy of adversary's model to analyse existing privacy preserving ad hoc routing protocols. To improve these protocols and to try address all privacy preserving properties,

we proposed NoName [42], a novel privacy-preserving ad hoc routing protocol. Based on trapdoor, virtual switching and partially disjoint multipath using Bloom filter, NoName ensures anonytmity of the source, of the destination and of intermediate nodes. It also ensures unlinkability between source and message and between destination and message.

In [43], we have proposed another anonymous *proactive* ad hoc routing protocol, called APART, based on Gentry's fully homomorphic cryptography. Even though this technology is currently quite inefficient from a computational perspective, especially for an application in ad-hoc networks, the protocol APART is merely a proof of concept showing that an anonymous proactive protocol is possible thanks to it. The main idea is that each node maintains a routing table that contains only encrypted data. When a source node want to communicate with a destination node, it cooperates with its neighbors to discover the node that is the next hop to the destination node. This is done in such a way that the source node does not know the entry in its routing table that corresponds to the destination, and the next hop does only know that it has to rebroadcast the messages coming from that source.

6.2.6. Right to be forgotten

The right to be forgotten has become an investigation topic in itself within the field of privacy protection. In [46], we present the joint research project funded by the ministry of justice between our team and researchers in law and sociology, in order to examine the current state, in society and in technology, of the notion of a right to be forgotten, to identify the forthcoming computing tools capable of implementing the notion, and to evaluate the relevance of an autonomous legislation to define it and regulate it. In association with this study and in the light of the identified state-of-the-art, we have proposed in [47] a new technique to implement a right to be forgotten in the manner of a degradation of the quality of published data in time, associated with a fully distributed ephemeral publication technology. We show how this technique could fit various use cases in geosocial networks.

6.3. Trust

Digital reputation mechanisms have indeed emerged as a promising approach to cope with the specificities of large scale and dynamic systems. Similarly to real world reputation, a digital reputation mechanism expresses a collective opinion about a target user based on aggregated feedback about his past behavior. The resulting reputation score is usually a mathematical object (*e.g.* a number or a percentage). It is used to help entities in deciding whether an interaction with a target user should be considered. Digital reputation mechanisms are thus a powerful tool to incite users to behave trustworthily. Indeed, a user who behaves correctly improves his reputation score, encouraging more users to interact with him. In contrast, misbehaving users have lower reputation scores, which makes it harder for them to interact with other users. To be useful, a reputation mechanism must itself be accurate against adversarial behaviors. Indeed, a user may attack the mechanism to increase his own reputation score or to reduce the reputation of a competitor. A user may also free-ride the mechanism and estimate the reputation of other users without providing his own feedback. From what has been said, it should be clear that reputation is beneficial in order to reduce the potential risk of communicating with almost or completely unknown entities. Unfortunately, the user privacy may easily be jeopardized by reputation mechanisms which is clearly a strong argument to compromise the use of such a mechanism. Indeed, by collecting and aggregating user feedback, or by simply interacting with someone, reputation systems can be easily manipulated in order to deduce user profiles. Thus preserving user privacy while computing robust reputation is a real and important issue that we address in our work [51], [23].

6.4. Other Topics Related to Security and Distributed Computing

6.4.1. Network Monitoring and Fault Detection

Monitoring a system consists in collecting and analyzing relevant information provided by the monitored devices, so as to be continuously aware of the system state (situational awareness). However, the ever growing complexity and scale of systems makes both real time monitoring and fault detection a quite tedious task. Thus

the usually adopted option is to focus solely on a subset of information states, so as to provide coarse-grained indicators. As a consequence, detecting isolated failures or anomalies is a quite challenging issue. We propose in [24], [44] 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 [12], we consider the setting of large scale distributed systems, in which each node needs to quickly process a huge amount of data received in the form of a stream that may have been tampered with by an adversary (*i.e.*, data items ordering can be manipulated by an omniscient adversary [13]). In this situation, a fundamental problem is how to detect and quantify the amount of work performed by the adversary. To address this issue, we propose AnKLe (for Attack-tolerant eNhanced Kullback- Leibler divergence Estimator), a novel algorithm for estimating the KL divergence of an observed stream compared to the expected one. AnKLe combines sampling techniques and information-theoretic methods. It is very efficient, both in terms of space and time complexities, and requires only a single pass over the data stream. Experimental results show that the estimation provided by AnKLe remains accurate even for different adversarial settings for which the quality of other methods dramatically decreases. Considering n as the number of distinct data items in a stream, we show that AnKLe is an (ε, δ)-approximation algorithm with a space complexity $\widetilde{\mathcal{O}}(\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2})$ 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 [21] 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 the continuation of [53] which proposed an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, in [22], we analyze 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 Uniform Sampling in Dynamic Systems

In [21], we consider the problem of achieving uniform node sampling in large scale systems in presence of a strong adversary. We first propose an omniscient strategy that processes on the fly an unbounded and arbitrarily biased input stream made of node identifiers exchanged within the system, and outputs a stream that preserves Uniformity and Freshness properties. We show through Markov chains analysis that both properties hold

despite any arbitrary bias introduced by the adversary. We then propose a knowledge-free strategy and show through extensive simulations that this strategy accurately approximates the omniscient one. We also evaluate its resilience against a strong adversary by studying two representative attacks (flooding and targeted attacks). We quantify the minimum number of identifiers that the adversary must insert in the input stream to prevent uniformity. To our knowledge, such an analysis has never been proposed before.

MYRIADS Project-Team

6. New Results

6.1. Dependable Cloud Computing

Participants: Roberto-Gioacchino Cascella, Stefania Costache, Florian Dudouet, Eugen Feller, Filippo Gaudenzi, Yvon Jégou, Ancuta Iordache, David Margery, Christine Morin, Anne-Cécile Orgerie, Guillaume Pierre, Nikos Parlavantzas, Yann Radenac, Matthieu Simonin, Cédric Tedeschi.

6.1.1. Multi-data Center and Multi-cloud

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

Participants: Roberto-Gioacchino Cascella, Stefania Costache, 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 [39] [50], 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 [51]. In 2013, we designed the CIMI inspired REST-API for VEP 2.0 with support for Constrained Execution Environment (CEE), advance reservation and scheduling service, and support for SLAs [55], [54] [56]. We integrated support for delegated certificates and developed test scripts to integrate the Virtual Infrastructure Network (VIN) service. VEP 1.1 was slightly modified to integrate the usage control (Policy Enforcement Point (PEP)) solution developed by CNR. The CEE management interface was developed during 2013 and is available through the graphical API as well as through the RESTful API.

6.1.1.2. Towards a distributed cloud inside the backbone

Participants: Anne-Cécile Orgerie, Cédric Tedeschi.

The DISCOVERY proposal currently in phase of construction and lead by Adrien Lèbre from ASCOLA team, and currently on leave at Inria aims at designing a distributed cloud, leveraging the resources we can find in the network's backbone. [3]

In this context, and in collaboration with ASCOLA and ASAP teams, we started the design of an overlay network whose purpose is to be able, with a limited cost, to locate geographically-close nodes from any point of the network. The basis for this overlay is described as part of a recent research report [44].

6.1.1.3. Multi-cloud application deployment in ConPaaS

Participants: Guillaume Pierre, Yann Radenac.

We extended ConPaaS to support application deployment over multiple clouds. There are two main reasons for this: first, it is a necessary mechanism to allow application migration from one cloud to another, without any service interruption. Second, for some applications it may be useful to execute over multiple clouds on a permanent basis, for reliability reasons for example. The main challenges to address were ensuring full network connectivity between resources acquired in multiple clouds. We addressed these issues by integrating the IPOP virtual network in ConPaaS. Second, we designed protocols to ensure application and data migration without any service interruption during the migration.

6.1.2. Scalability of Snooze Self-healing Cloud Management System

Participants: Eugen Feller, Yvon Jégou, David Margery, Christine Morin, Anne-Cécile Orgerie, Matthieu Simonin.

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

We evaluated the scalability and resilience of Snooze IaaS management system [26]. Unlike existing systems, for scalability, ease of configuration, and high availability, Snooze is based on a self-organizing and self-healing hierarchical architecture of system services [36], [27], [27]. In Snooze hierarchy, each compute server is managed by a local controller that interacts with one of the group managers to which it is dynamically assigned and the set of group managers is coordinated by a group leader elected among them. We performed an extensive scalability study of Snooze across over 500 servers of the Grid'5000 experimentation testbed. We evaluated the Snooze self-organizing and self-healing hierarchy with thousands of system services. The results show that the resource consumption of the Snooze system services is bounded both during the hierarchy construction and system operation. We also show that Snooze prototype implementation is robust enough to manage thousands of servers and hundreds of VMs. Moreover, its autonomic behavior allows to achieve high availability in the presence of a large number of simultaneous system services failures. Indeed, as long as at least two group managers remain operational the system remains alive. We also demonstrated the application deployment scalability across hundreds VMs on the example of a Hadoop MapReduce application. We participated in the Scale Challenge organized in the framework of the ACM/IEEE CC-Grid 2013 conference [26] and won the second prize.

6.1.3. Application Performance Modeling in Heterogeneous Cloud Environments

Participants: Ancuta Iordache, Guillaume Pierre.

Heterogeneous cloud platforms offer many possibilities for applications for make fine-grained choice over the types of resources they execute on. This opens for example opportunities for fine-grain control of the tradeoff between expensive resources likely to deliver high levels of performance, and slower resources likely to cost less. We designed a methodology for automatically exploring this performance vs. cost tradeoff when an arbitrary application is submitted to the platform. Thereafter, the system can automatically select the set of resources which is likely to implement the tradeoff specified by the user. A publication on this topic is currently in preparation.

6.1.4. Flexible SLA & SLO Management

Participants: Stefania Costache, Christine Morin, Nikos Parlavantzas.

Merkat is a market-based, SLO-driven, PaaS system for private clouds. Merkat 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. Merkat provides users the flexibility to adapt controllers to their application types, and it can support diverse application types and performance goals. Merkat is implemented in Python and uses OpenNebula for virtual machine operations.

We evaluated Merkat in simulation and we analyzed the behavior of the system for multiple user types [23]. Furthermore, we deployed Merkat on Grid'5000 and EDF's tested and tested it with applications representative to EDF [22]. Results showed that: (i) the system provides flexible support for different application types (static and malleable) and different SLOs (deadline and performance); (ii) the system provides good user satisfaction achieving acceptable performance degradation, compared to existing centralized solutions. Furthermore, we extended Merkat to manage different clusters and run MPI applications on them. We also submitted a survey on evolution of resource management systems for shared virtualized computing infrastructures to an international journal. This work was carried out in the framework of Stefania Costache's PhD thesis [11].

6.2. Heterogeneous Resource Management

Participants: Eliya Buyukkaya, Djawida Dib, Eugen Feller, Tran Ngoc Minh, Christine Morin, Nikos Parlavantzas, Guillaume Pierre.

6.2.1. Cross-resource scheduling in heterogeneous cloud environments

Participants: Eliya Buyukkaya, Tran Ngoc Minh, Guillaume Pierre.

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

6.2.2. Maximizing private cloud provider profit in cloud bursting scenarios

Participants: Christine Morin, Djawida Dib, Nikos Parlavantzas.

Current PaaS offerings either provide no support for SLA guarantees or provide limited support targeting a restricted set of application types. To overcome this limitation, we are developing an open, SLA-driven PaaS system, called Meryn, that aims at providing SLA guarantees to diverse application types while maximizing the PaaS provider profit. Meryn supports cloud bursting and applies a decentralized protocol for selecting cloud resources, trying to minimize the cost of running applications without affecting their agreed quality properties. We have performed a preliminary evaluation of Meryn [24] and worked on optimising the system and performing further experiments on the Grid5000 testbed. This work is part of Djawida Dib's PhD thesis.

6.2.3. Data life-cycle management in clouds

Participants: Eugen Feller, Christine Morin.

Infrastructure as a Service (IaaS) clouds provide a flexible environment where users can choose and control various aspects of the machines of interest. However, the flexibility of IaaS clouds presents unique challenges for storage and data management in these environments. Users use manual and/or ad-hoc methods to manage storage and data in these environments. FRIEDA is a Flexible Robust Intelligent Elastic Data Management framework that employs a range of data management strategies approaches in elastic environments. In the context of the DALHIS associate team [4], we evaluated the importance of this framework on multiple cloud testbeds. Our evaluation showed that storage planning needs to be performed in coordination with compute planning and the specific configuration of virtual machine had a strong impact on the application (e.g., some applications performed better on small instances than large instances) [40].

6.3. Energy-efficient Resource Infrastructures

Participants: Alexandra Carpen-Amarie, Bogdan Florin Cornea, Ismael Cuadrado Cordero, Djawida Dib, Eugen Feller, Yunbo Li, Christine Morin, Anne-Cécile Orgerie, Guillaume Pierre.

6.3.1. Energy-efficient IaaS clouds

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

Energy consumption has always been a major concern in the design and cost of data centers. The wide adoption of virtualization and cloud computing has added another layer of complexity to enabling an energy-efficient use of computing power in large-scale settings. Among the many aspects that influence the energy consumption of a cloud system, the hardware-component level is one of the most intensively studied. However, higher-level factors such as virtual machine properties, their placement policies or application workloads may play an essential role in defining the power consumption profile of a given cloud system. In this work, we explored the energy consumption patterns of Infrastructure-as-a-Service (IaaS) cloud environments under various synthetic and real application workloads. For each scenario, we investigated the power overhead triggered by different types of virtual machines, the impact of the virtual cluster size on the energy-efficiency of the hosting infrastructure and the tradeoff between performance and energy consumption of MapReduce virtual clusters through typical cloud applications [21].

6.3.2. Energy-aware IaaS-PaaS co-design

Participants: Alexandra Carpen-Amarie, Djawida Dib, Guillaume Pierre, Anne-Cécile Orgerie.

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

The wide adoption of the cloud computing paradigm plays a crucial role in the ever-increasing demand for energy-efficient data centers. Driven by this requirement, cloud providers resort to a variety of techniques to improve energy usage at each level of the cloud computing stack. However, prior studies mostly consider resource-level energy optimizations in IaaS clouds, overlooking the workload-related information locked at higher levels, such as PaaS clouds. We argue that cross-layer cooperation in clouds is a key to achieving an optimized resource management, both performance and energy-wise. To this end, we claim there is a need for a cooperation API between IaaS and PaaS clouds, enabling each layer to share specific information and to trigger correlated decisions. We identified the drawbacks raised by such co-design objectives and discuss opportunities for energy usage optimizations, and plan to start the research to address these issues in 2014.

6.3.3. *Performance and energy-efficiency evaluation of Hadoop deployment models*

Participants: Eugen Feller, Christine Morin.

The exponential growth of scientific and business data has resulted in the evolution of the cloud computing and the MapReduce parallel programming model. Cloud computing emphasizes increased utilization and power savings through consolidation while MapReduce enables large scale data analysis. The Hadoop framework is the most popular open source software implementing the MapReduce model. In our work, we evaluated Hadoop performance in two modes – the traditional model of collocated data and compute services and separated mode where the services are deployed on separate services. The separation of data and compute services provides more flexibility in environments where data locality might not have a considerable impact such as virtualized environments and clusters with advanced networks. In this work, we also conducted an energy efficiency evaluation of Hadoop on physical and virtual clusters in different configurations. The experiments were performed on the Grid'5000 experimentation testbed. To enable virtual machine management, the Snooze cloud stack developed by the Myriads project-team was used. Our extensive evaluation shows that: (1) performance on physical clusters is significantly better than on virtual clusters; (2) performance degradation due to separation of the services depends on the data to compute ratio; (3) application completion progress correlates with the power consumption and power consumption is heavily application specific [28].

6.3.4. *Energy consumption models and predictions for large-scale systems*

Participant: Christine Morin.

Responsible, efficient and well-planned power consumption is becoming a necessity for monetary returns and scalability of computing infrastructures. While there is a variety of sources from which power data can be obtained, analyzing this data is an intrinsically hard task. In our work, we described a generic approach to analyze large power consumption datasets collected from computing infrastructures. As a first step, we proposed a data analysis pipeline that can handle the large-scale collection of energy consumption logs, apply sophisticated modeling to enable accurate prediction, and evaluate the efficiency of the analysis approach. We presented the analysis of a power consumption data set collected over a 6-month period from two clusters of the Grid'5000 experimentation platform used in production. We used Hadoop with Pig to handle the large volume of data. Our data processing generated a summary of the data that provides basic statistical aggregations, over different time scales. The aggregate data was then analyzed as a time series using sophisticated modeling methods with R statistical software. We exploited time series to detect outliers and derive hourly and daily power consumption predictive models. We demonstrated the accuracy of the predictive models and the efficiency of the data processing performed on a 55-node cluster at NERSC [34]. Energy models from such large dataset can help in understanding the evolution of consumption patterns, predicting future energy trends, and providing basis for generalizing the energy models to similar large-scale systems.

6.3.5. *Simulating Energy Consumption of Wired Networks*

Participant: Anne-Cécile Orgerie.

Predicting the performance of applications, in terms of completion time and resource usage for instance, is critical to appropriately dimension resources that will be allocated to these applications. Current applications, such as web servers and Cloud services, require lots of computing and networking resources. Yet, these resource demands are highly fluctuating over time. Thus, adequately and dynamically dimension these resources is challenging and crucial to guarantee performance and cost-effectiveness. In the same manner, estimating the energy consumption of applications deployed over heterogeneous cloud resources is important in order to provision power resources and make use of renewable energies. Concerning the consumption of entire infrastructures, some studies show that computing resources represent the biggest part in Cloud's consumption, while others show that, depending on the studied scenario, the energy cost of the network infrastructure that links the user to the computing resources can be bigger than the energy cost of the servers. In this work, we aim at simulating the energy consumption of wired networks which receive little attention in the Cloud computing community even though they represent key elements of these distributed architectures. To this end, we are contributing to the well-known open-source simulator ns3 by developing an energy consumption module named ECOFEN.

6.3.6. *Simulating the impact of DVFS within SimGrid*

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

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

6.4. Unconventional Models for Large Computations and Platforms

Participants: Marko Obrovac, Christine Morin, Cédric Tedeschi.

6.4.1. *Chemical computing at large scale*

Participants: Marko Obrovac, Cédric Tedeschi.

One of the commonly cited problem when dealing with chemistry-inspired computing is its lack of experimental validation. The DHT-based runtime developed recently, in the framework of Marko Obrovac's PhD thesis [13], has been deployed over the Grid'5000 platform with promising results. This runtime is now mature enough for being considered as a viable candidate to underlie a distributed workflow engine [32].

6.4.2. *Template workflows*

Participants: Christine Morin, Cédric Tedeschi.

In the framework of the DALHIS associate team [5], we plan to combine the high-level template workflow language TIGRES [6], developed by our partner team from Lawrence Berkeley National Lab (LBL) with the workflow management system developed in the team [17]. This work started with the development of a parser of TIGRES.

6.5. Experimental Platforms

Participants: Alexandra Carpen-Amarie, Maxence Dunnewind, Nicolas Lebreton, Julien Lefeuvre, David Margery, Eric Poupart.

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

6.5.1. Energy measurement

Participants: David Margery, Maxence Dunnewind, Nicolas Lebreton.

In the context of the ECO_2Clouds project, the BonFIRE infrastructure was updated. At the hardware level power distribution units that report electricity usage for each outlet were installed. At the software layer, a probe reporting energy sources used was configured. This probe gets its information from RTE, the French Electricity transport network, and allows publication of CO_2 metrics for each machine in the testbed. Moreover, access to these metrics was abstracted through the general API to access BonFIRE.

6.5.2. Deployment of IaaS management system

Participant: Alexandra Carpen-Amarie.

The Grid'5000 platform has become one of the most complete testbeds for designing or evaluating large-scale distributed systems, playing an essential role in enabling experimental research at all levels of the Cloud Computing stack and providing configurable cloud platforms similar to commercially available clouds.

However, the complexity of managing the deployment and tuning of large-scale private clouds emerged as a major drawback. Typically, users study specific cloud components or carry out experiments involving applications running in cloud environments. A key requirement in this context is seamless access to ready-to-use cloud platforms, as well as full control of the deployment settings.

To address these needs, we developed a set of deployment tools for open-source IaaS environments, capable of installing and tuning fully-functional clouds on the Grid'5000 testbed [20]. The deployment tools support four widely-used IaaS clouds, namely OpenNebula, CloudStack, Nimbus and OpenStack.

They rely on the concept of extensible engines for defining experiments. Such engines implement all the stages of an experiment: physical node reservations in Grid'5000, environment deployment, configuration and experiment execution. We designed generic engines for nodes reservation and deployment according to a set of requirements specified in a cloud configuration file. Thus, these engines do not require any prior knowledge of lower-level Grid'5000 tools, allowing the user to easily achieve multi-site Grid'5000 deployments based on multiple environments.

6.5.3. BonFIRE

Participants: Maxence Dunnewind, David Margery, Eric Poupart.

The project was reviewed in December 2013 during CloudCom 2013 in Bristol and rated Excellent. The main achievement this year is the introduction of a reservation system for resources on the BonFIRE platform.

6.5.4. Fed4FIRE

Participants: Julien Lefeuvre, Nicolas Lebreton, David Margery.

In Fed4FIRE, two key technologies have been adopted as common protocols to enable experimenter to interact with testbeds. SFA, to provision resources, and OMF to control them. Here, we contributed to a proposal to secure usage of OMF and to a design to allow using BonFIRE through SFA.

REGAL Project-Team

5. New Results

5.1. Introduction

In 2013, we focused our research on the following areas:

- *Distributed algorithms for dynamic and large networks.*
- *Management of distributed data.*
- *Performance and robustness of Systems Software in multicore architectures.*

5.2. Distributed algorithms for dynamic networks

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

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.

5.2.1. 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 [46], [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. In 2013, we studied Ω, the eventual leader election failure detector. Ω ensures that, eventually, each process in the system will be provided by an unique leader, elected among the set of correct processes in spite of crashes and uncertainties. It is known to be weakest failure detector to solve agreement protocols such as Paxos. Then, a number of eventual leader election protocols were suggested. Nonetheless, as far as we are aware of, no one of these protocols tolerates a free pattern of node mobility. In [27] we propose a new protocol for this scenario of dynamic and mobile unknown networks.

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

5.2.2. 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. 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 and gathering.

5.2.3. 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. In 2013, we have proposed a new distributed algorithm suitable for scale-free random topologies which model some complex real world networks [37], [52].

5.2.4. Peer certification.

In a distributed system, the certification of transactions makes it possible to circumscribe malicious behaviors. Certification requires the use of a trusted third party which must be centralized to guarantee safety. At a large scale, however, centralized certification represents a bottleneck and a single point of attack or failure.

We proposed two decentralized approaches towards certifying transactions with a high probability of success. The first approach replicates transactions over multiple peers and retains identical results from a qualified majority to certify that a service has been carried out for a given client at a given time [30]. The second approach uses distributed reputations to identify trusted nodes and use them as game referees to detect and prevent cheating [57].

5.3. Management of distributed data

Participants: Pierpaolo Cincilla, Guthemberg Da Silva Silvestre, Raluca Diaconu, Jonathan Lejeune, Mesaac Makpangou, Olivier Marin, Sébastien Monnet, Dang Nhan Nguyen, Burcu Külahçioglu Özkan, Karine Pires, Masoud Saeida Ardekani, Thomas Preud'Homme, Pierre Sens, Marc Shapiro, Véronique Simon, Julien Sopena, Gaël Thomas, Mathieu Valero, Mudit Verma, Marek Zawirski.

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

- where to store or replicate the data, in order to ensure that it is available quickly and remains persistent despite failures and disconnections;
- how many copies are needed to face dynamically-changing demand (load) and offer (elasticity);
- how to parallelize writes and hence how to ensure consistency between replicas;
- tradeoffs between synchronised, consistent but slow updates, and fast but weakly-consistent ones;
- when and how to move data to computation, or computation to data, in order to improve response time while minimizing storage or energy usage;
- etc.

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

5.3.1. Long term durability

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

Well sized systems rarely loose data, still, data may be lost: the more the time passes, the greater is the risk of loss. It is thus necessary to study data durability on a long term. To do so, we have implemented an efficient simulator, we can simulate a 100 node system over years within several hours. We have observe that a given system with a given replication mechanism can store a certain amount of data above which the loss rate would be greater than an "acceptable"/fixed threshold. This amount of data can be used as a metric to compare replication strategies. We have studied the impact of the data distribution layout upon the loss rate. The way the replication mechanism distribute the data copies among the nodes has a great impact. If node contents are very correlated, the number of available sources to heal a failure is low. On the opposite, if the data copies are shuffled among the nodes, many source nodes may be available to heal the system, and thus, the system losses less pieces of data. We are also studying the impact of other parameters, like the replication degree or the way a new storer node is chosen.

5.3.2. Adaptative replication

Different pieces of data have different popularity: some data are stored but never accessed while other pieces are very "hot" and are requested concurrently by many clients. This implies that different pieces of data with different popularity should have a different number of copies to efficiently serve the requests without wasting resources. Furthermore, for a given piece of data, the popularity may vary drastically among time. It is thus important that the replication mechanism dynamically adapt the number of replicas to the demand. In the context of the ODISEA2 FUI project, we have made two main contributions. First, we have studied the popularity distribution and evolution of live video streams (Karine Pires thesis). Second, we have designed replication mechanisms able to gracefully adapt the replication degree to the demand, one based on bandwidth reservation, and one using statistical learning (Guthemberg Silvestre thesis).

5.3.3. 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: under standard assumptions (data store accesses are not known in advance, and transactions may access arbitrary objects in the

data store), it is impossible to have both SI and GPR. Our impossibility result is based on a novel decomposition of SI which proves that, like serializability, SI is expressible on plain histories. These results are published at the Euro-Par conference [42].

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

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

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

Non-Monotonic Snapshot Isolation (NMSI) is the first strong consistency criterion to allow implementations with all four properties. We also present a practical implementation of NMSI called Jessy, which we compare experimentally against a number of well-known criteria. Our measurements show that the latency and throughput of NMSI are comparable to the weakest criterion, read-committed, and between two to fourteen times faster than well-known strong consistencies. This was published in the Symp. on Reliable Distr. Sys. (SRDS) [43].

5.3.4. Distributed Transaction Scheduling

Parallel transactions in distributed DBs incur high overhead for concurrency control and aborts. Our Gargamel system proposes an alternative approach by pre-serializing possibly conflicting transactions, and parallelizing non-conflicting update transactions to different replicas. This system provides strong transactional guarantees. In effect, Gargamel partitions the database dynamically according to the update workload. Each database replica runs sequentially, at full bandwidth; mutual synchronisation between replicas remains minimal. 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.

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.

5.3.5. 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).

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? What are the upper and lower bounds of CRDTs? We co-authored an innovative approach to these questions, to be published at Principles of Programming Languages (POPL) 2014 [29].

We are currently developing an extreme-scale CRDT platform called SwiftCloud; see Section 4.2 .

5.3.6. Mixing commutative and non-commutative updates: reservations

Asynchronous updates are desirable because they ensure the system is available, fast and scalable. CRDTs are asynchronous, but cannot guarantee strong invariants, such as ensuring that a shared counter never goes negative. To solve this problem, we define a novel hybrid model that supports both synchronous and asynchronous updates, "red-blue-purple" consistency. The RPB model classifies updates into commutative, partially-commutative and non-commutative, and distinguishes the (global) states where partially-commutative operations can safely run asynchronously. We use reservation techniques to ensure operation in such states. A reservation promises, to a cache that holds it, that the system is in a state that allows the cache server to perform purple updates asynchronously. Reservations ensure that data is in a known state by caching both data and access permissions over data to make updates. This approach strengthens the safety guarantees in addition to eventual consistency [40].

5.4. Performance and Robustness of Systems Software in Multicore Architectures

Participants: Koutheir Attouchi, Harris Bakiras, Antoine Blin, Florian David, Bertil Folliot, Lokesh Gidra, Julia Lawall, Jean-Pierre Lozi, Gilles Muller [correspondent], Dang Nhan Nguyen, Thomas Preud'Homme, Suman Saha, Peter Senna Tschudin, Marc Shapiro, Julien Sopena, Gaël Thomas, Mudit Verma.

5.4.1. Managed Runtime Environments

Today, multicore architectures are becoming ubiquitous, found even in embedded systems, and thus it is essential that managed runtime environments can scale on multicore processors. We have found that two major scalability bottlenecks are the implementation of highly contented locks and of garbage collectors. On a multicore, a single lock can overload the bus because the cache line that contains the lock bounces between the cores, eliminating all the 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. This work initially targeted C code but is now being adapted to the needs of Java applications in the PhD of Florian David. For garbage collectors, as the memory is physically distributed among a set of memory controllers, a collection saturates the bus when the collector threads access remote memory. This saturation prevents the garbage collector from scaling with the number of cores, making the garbage collector a major bottleneck of managed runtime environments on multicore hardware. As part of the PhD of Lokesh Gidra, we have identified memory placement schemes that decrease the number of remote memory accesses during a collection in OpenJDK 7, thus preventing the bottleneck caused by bus saturation [36].

5.4.2. System software robustness

A widely recognized problem in the area of finding bugs in API usage in systems code is to know what APIs are expected and to identify contexts where these expectations are not satisfied. Indeed, systems code, such as an operating systems kernel, is typically voluminous, amounting to millions of lines of code, and uses many different highly specialized APIs, making it impossible for most developers to keep the usage protocols of all of them in mind. To address this issue, we have developed an approach to inferring API function usage protocols from software, relying on knowledge of common code structures (Software – Practice and Experience [26]). Building on this experience, we have developed an approach to finding resource-release omission faults in

systems code that leverages information local to a single function [44]. This approach permits finding hundreds of faults in Linux kernel code as well as a variety of other systems software, with a low rate of false positives. Finally, we have initiated an effort on understanding the range and scope of the oops reports collected in the recently revived Linux kernel oops repository [59].

Beyond finding faults in existing code, we have also considered how systems code is constructed. Specifically, in the context of Linux device drivers, we have identified the notion of a *gene*, as a sequence of code fragments that express a particular device or operating system functionality. We have performed an initial partial sequencing of the genes making up the probe functions of Linux platform drivers [45]. Relatedly, in the context of a Merlion collaboration grant with David Lo of Singapore Management University, we have considered the problem of recommending APIs to developers. We propose one approach based on the set of libraries used by other software having similar properties [47], and a second approach based on the set of libraries used to implement related feature requests [48].

5.4.3. Domain-specific languages for systems software

A challenge in the management of a datacenter is the placement of application replicas, both to avoid a single point of failure and to limit communication costs. We have proposed a novel approach, BtrPlace [23], based on the use of a domain-specific language to express constraints derived from properties of the application and of the datacenter, and the use of a constraint solver to efficiently resolve these constraints. Simulations show that BtrPlace is able to repair a configuration involving 5000 servers after a server failure in 3 minutes.

While the use of domain-specific languages such as that of BtrPlace can ease programming, it is well known that developing, and especially maintaining, a domain-specific language over time is time-consuming and challenging. This is particularly the case when the domain-specific language provides domain-specific verifications, as the code implementing these verifications has to be maintained along with the rest of the language implementation. Furthermore, new domain-specific languages typically must evolve frequently, as the language developer comes to better understand the range and scope of the domain. To address these issues, we have proposed a methodology for domain-specific language implementation development for C-like domain-specific languages [19], based on the use of rewriting rules implemented using Coccinelle. We apply this approach to our previously developed domain specific language z2z for developing network gateways, and find that the resulting language implementation is more concise and easier to extend with new language features.

SCORE Team

5. New Results

5.1. Evaluation and Design of Collaborative Editing Algorithms

Participants: Mehdi Ahmed-Nacer, Luc André, Claudia-Lavinia Ignat, Stéphane Martin, Gérald Oster, Pascal Urso.

Since the Web 2.0 era, the Internet is a huge content editing place in which users contribute to the content they browse. Users do not just edit the content but they collaborate on this content. Such shared content can be edited by thousands of people. However, current consistency maintenance algorithms seem not to be adapted to massive collaborative updating involving large amount of contributors and a high velocity of changes. This year we continued our work on the evaluation of existing collaborative editing approaches and on the design of new algorithms that overcome limitations of state of the art ones. Moreover, we started to work on experimental user studies for understanding the real-time requirements for collaborative editing and grounding a theory for the effect of real-time constraints in collaborative work [26].

We also run experiments to compare the merge automatically obtained by collaborative editing algorithms – CRDTs, OTs and the world-wide used diff3 – to the merge validated by the user. We obtain automatically such results exploiting the massively available distributed version control systems histories of open-source software. We use these results to improve an existing collaborative editing algorithm and obtain result statistically better than the existing ones (including diff3 used in major DVCS systems) [9].

In existing collaborative editing algorithms shared data is usually fragmented into fixed granularity atomic elements that can only be added or removed. Coarse-grained data leads to the possibility of conflicting updates while fine-grained data requires more metadata. In [11] we offer a solution for handling an adaptable granularity for shared data that overcomes the limitations of fixed-grained data approaches. Our solution relies on a novel commutative replicated data type (CRDT) for sequences of text that assigns unique identifiers to substrings of variable length contrary to existing CRDTs that assign unique identifiers to fixed size elements of the text (i.e. characters or lines). This offers the possibility to define coarse grained elements when they are created and refine them when needed. This greatly reduces the memory consumption since a smaller memory overhead is needed to store metadata (identifiers). Moreover, we show using simulations that overall performances of our algorithms are superior to existing ones.

We proposed a new concurrency control algorithm, based on conflict-free data types. It is built on the ideas previously developed for synchronous collaboration, extending them to support asynchronous collaboration. Our solution also includes the necessary information for providing comprehensive awareness information to users. The evaluation of our algorithm shows that comparing our solution with traditional solutions in collaborative editing, the conflict resolution strategy proposed in this paper leads to results closer to the ones expected by users [10].

5.2. Decentralized monitoring of orchestration execution

Participants: Mohamed Aymen Baouab, Olivier Perrin, Claude Godart.

Cross-organizational service-based processes are increasingly adopted by different companies when they cannot achieve goals on their own. The dynamic nature of these processes poses various challenges to their successful execution. In order to guarantee that all involved partners are informed about errors that may happen in the collaboration, it is necessary to monitor the execution process by continuously observing and checking message exchanges during runtime. This allows a global process tracking and evaluation of process metrics. Complex event processing can address this concern by analysing and evaluating message exchange events, to the aim of checking if the actual behaviour of the interacting entities effectively adheres to the modelled business constraints. In our recent work (Aymen Baouab thesis [1]), we presented an approach for

decentralized monitoring of cross-organizational choreographies. We have defined a hierarchical propagation model for exchanging external notifications between the collaborating parties. We also proposed 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 [12].

5.3. Optimization and security of business processes in SaaS contexts

Participants: Claude Godart, Elio Goettelmann, Samir Youcef.

Globalization and the increase of competitive pressures created the need for agility in business processes, including the ability to outsource, offshore, or otherwise distribute its once-centralized business processes or parts thereof. While hampered thus far by limited infrastructure capabilities, the increase in bandwidth and connectivity and decrease in communication cost have removed these limits. This is even more true with the advent of cloud, particularly in its "Service as a software" dimension. To adapt to such a context, there is a growing need for the ability to fragment one's business processes in an agile manner, and be able to distribute and wire these fragments so that their combined execution recreates the function of the original process. Our work focuses on solving some of the core challenges resulting from the need to dynamically restructure enterprise interactions. Restructuring such interactions corresponds to the fragmentation of intra and inter enterprise business process models. It describes how to identify, create, and execute process fragments without loosing the operational semantics of the original process models. In addition, this fragmentation is complicated by the constraints of quality of service, in particular the execution time and the cost, and of security, especially privacy. During the year, we consider this problem at two levels: the design of privacy-aware process models, and the process scheduling optimization. We developed a methodology to integrate privacy concerns in the design of a business process before distribution in the cloud. Based on a risk analysis, the result of the design is a set of process (re-)modelling actions, a set of constraints on process fragments assignments to clouds, and a set of constraints for cloud selection based on cloud properties [19]. We developed bi-criteria strategies for business processes scheduling in cloud environments with execution time and cost constraints, augmented with fairness metrics, and taking into account the availability of human resources, a critical point in business processes [14], [15], [3].

5.4. Large Scale Coordination of Crowdsourcing Activities

Participants: François Charoy, Karim Benouaret, Raman Valliyur-Ramalingam, Alexandre Roux d Anzi.

As a follow-up of our work on coordination of large scale processes that we have investigated in the domain of crisis management [4], [5], we have studied a new application domain for BPM, crowdsourcing. In order to make cities smarter, it would be interesting to design a platform where citizens are given an opportunity to be effectively connected to the governing bodies in their location and to contribute to the general well being. We have developed CrowdSC, a crowdsourcing framework designed for smarter cities. We have shown that it is possible to combine data collection, data selection and data assessment crowdsourcing activities in a crowdsourcing process to achieve sophisticated goals in a predefined context. Depending on the executing strategy of this process, different kinds of outcomes can be produced. We have conducted an experimental study that evaluates these process outcomes depending on different execution strategies [2], [13].

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.

Participants: Sylvain Contassot-Vivier, Thomas Jost.

The PhD thesis of Thomas Jost, co-supervised by S. Contassot-Vivier and Bruno Lévy (Alice INRIA team) since January 2010, dealt with specific algorithms for GPUs, in particular linear solvers [32]. He also worked on the use of GPUs within clusters of workstations via the study of a solver of non-linear problems [30], [33], [29]. The defense of this thesis was initially planned in January 2013 but Thomas decided at the end of 2012 to stop his PhD and to leave for industry.

6.1.2. Development methodologies for parallel programming of clusters.

Participants: Sylvain Contassot-Vivier, Jens Gustedt, Stéphane Vialle.

We have conducted a particular effort in merging and synthesizing our respective experiences of parallel programming of clusters (homogeneous, heterogeneous, hybrid). This has lead to two book chapters [19] and [34] (to appear).

6.1.3. Combining locking and data management interfaces.

Participants: Jens Gustedt, Stéphane Vialle, Soumeya Leila Hernane, Rodrigo Campos-Catelin.

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. The thesis of Soumeya Hernane [12] has been defended in 2013. It 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 5.2.1 . In 2013, work has demonstrated its efficiency for a large variety of platforms, see [22]. By using the example of dense matrix multiplication, we show that ORWL permits to reuse existing code for the target architecture, namely open source library ATLAS, Intel's compiler specific MKL library or NVidia's CUBLAS library for GPUs. ORWL assembles local calls into these libraries into efficient functional code, that combines computation on distributed nodes with efficient multi-core and accelerator parallelism.

Additionally, during the internship of Rodrigo Campos-Catelin, a detailed instrumentation of the ORWL library has been undertaken, and a new, less expensive strategy for cyclic FIFOs has been tested. This work will be continued with a master thesis at the university of Buenos Aires that will summarize and extend the results that were achieved during the internship.

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

6.1.4. Discrete and continuous dynamical systems.

Participants: Sylvain Contassot-Vivier, Marion Guthmuller.

The continuous aspect of dynamical systems has been intensively studied through the development of asynchronous algorithms for solving PDE problems. In past years, we have focused our studies on the interest of GPUs in asynchronous algorithms [29]. Also, we have investigated 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. The SimGrid environment has been used to validate and evaluate load balancing strategies in parallel iterative algorithms on large scale systems [28].

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 [31]. The expected results of that work may provide a very interesting tool for studying dynamical systems expressed under the form of a distributed application.

6.2. Transparent Resource Management for Clouds

Participants: Julien Gossa, Rajni Aron, Stéphane Genaud, Étienne Michon, Marc-Eduard Frîncu.

6.2.1. Provisioning strategies.

Our main achievement was the design of one comprehensive provisioning meta-strategy. This meta-strategy only use one parameter as a deadline given by the user. Contrary to other deadline-based provisioning strategies, our meta-strategy is able to combine any provisioning strategy in order to optimize the cost while meeting the deadline. This is achieved through simulation of cost and makespan of every available strategy thanks to SCHIaaS5.4.3 . It allows to apply the most inexpensive strategy as long as possible, before progressively switching to more expensive strategy when the deadline becomes closer.

The next step is to asses this meta-strategy among an important set of applications and platforms, both in real environments and simulation. The data are currently gathered and analyzed, and we should be able to draw conclusions soon.

6.2.2. User workload analysis.

We have conducted one broad study about workflows execution on the cloud, from both the theoretical and experimental point of view. In this study, we tried to discover causalities between the characteristics of workflows and the performances of provisioning strategies. We concluded that, except very peculiar cases, no causality can be identified. That is why we decided to make use of simulation to predict the strategies performances.

This predictive process is now integrated as a module of our cloud broker. It can be invoked by a user to help him decide which strategy should be used before any actual resource leasing.

We are now convinced that workload analysis is not a suitable approach because of its lack of generality.

6.2.3. Experimentations.

Given the very large consumption of CPU hours, the above work was supported mostly by simulation. We have assessed the gap between the performances of real executions on a private cloud and simulation. The latter proved to be very accurate, predicting almost perfectly the cost and makespan of every strategy on a wide range of workloads.

However, we have also shown that the simulation can be very sensitive to user defined input parameters (such as task runtimes) and may be mislead in borderline cases. Identifying the pitfalls and limitations of the simulation is very important and should end up in recommendations for a wise interpretation of simulation results.

We have also extended the range of experimentations to assess our simulator. First, we have extended the set of simulations with new applications, mostly workflows that are both generated and real applications (i.e. Montage). Second, we have conducted intensive experimentations on new platforms (i.e. Bonfire). The experimental data we have recently gathered in both cases is to be analyzed to further validate our approach.

6.3. Experimental methodologies for the evaluation of distributed systems

This year, M. Quinson defended his Habilitation on the experimental methodologies of distributed systems [13]. This concludes 10 years of research on this topic (including the elements presented in this section), and paves the road of future research.

6.3.1. Simulation and dynamic verification

6.3.1.1. MPI simulation

Participants: Martin Quinson, Paul Bédaride, Marion Guthmuller.

We continued our long-term effort toward the simulation of HPC application within SimGrid. We slightly increased the API coverage of our reimplementation of MPI on top of SimGrid, and proposed a new model of the network performance for MPI applications on top of Ethernet TCP networks. This model combines the advantages of flow-based networks for large data transfers as previous SimGrid network models, but also leverage algorithmic performance models extending the classical LogP models. As shown in [16], these models greatly improve the realism of MPI simulations, enabling the prediction of the performance of a non-trivial application in great details.

6.3.1.2. Dynamic verification and SimGrid

Participants: Marion Guthmuller, Martin Quinson, Gabriel Corona.

This year, our work toward the verification of liveness properties within SimGrid became fully functional thanks to the PhD work of M. Guthmuller. This relies on a system-level introspection mechanism allowing the model checker to finely explore the state of the verified programs. This is mandatory to detect the execution cycles that constitute the counter examples to liveness properties. This introspection mechanism is also used to implement a new reduction mechanism that can mitigate the state space explosion problem. A publication presenting these results is currently under review.

6.3.1.3. SimGrid framework improvement

Participants: Paul Bédaride, Martin Quinson, Gabriel Corona.

We rolled out a new major version of the SimGrid framework to our users. It contains both the HPC network models used to improve the prediction of MPI applications and all of our developments toward the dynamic verification of distributed applications. We also improved further the usability of our framework, that is now properly integrated within the Debian Linux distribution.

The next release is already underway, with a proper integration of the work from our partners on virtual machines and with a full reimplementation of the simulation kernel in C++ for a better modularity.

6.3.1.4. Formal Verification of Distributed Algorithms

Participants: Esteban Campostrini, Martin Quinson, Stephan Merz.

M. Quinson co-advised an internship with S. Merz (project-team Veridis) on the formal verification of distributed algorithm. The goal was to push further the PlusCal algorithmic language and its compiler to TLA$^+$ on which we are working since several years within the Veridis team.

We wanted to explore some hard problem raised by the verification of distributed protocol, such as how to represent timeout errors in verification settings where the time is not present. We think that this could be modeled somehow similarly to fairness properties, but more work is needed in this topic for a definitive answer.

6.3.2. Experimentation on testbeds and production facilities, emulation

6.3.2.1. Distem improvements: scalability and matrix-based inter-nodes latencies

Participants: Ahmed Bessifi, Emmanuel Jeanvoine, Lucas Nussbaum.

(For context, see sections 3.3 and 5.3 .)

Following our PDP'13 publication[18], we focused on improving Distem's scalability. First, on the Distem engine side, we parallelized the startup of physical nodes and virtual nodes, and added support for BTRFS snapshots to enable starting a very large number of virtual nodes with their own filesystems. Second, during the internship of Ahmed Bessifi we investigated several networking issues causing problems with large-scale experiments (over 4000 virtual nodes). The resulting improvements to ARP parameters tunings were integrated in Distem 0.8, and enabled network-intensive experiments with up to 8000 virtual nodes. We plan to publish those results in early 2014.

In the context of the AEN HEMERA project, we worked with Trong-Tuan Vu (EPI DOLPHIN, Inria Lille Nord Europe) to add support for specifying inter-nodes latencies using a matrix. This is especially useful for experiments on load-balancing and locality.

6.3.2.2. Evaluating load balancing HPC runtimes with Distem

Participants: Joseph Emeras, Emmanuel Jeanvoine, Lucas Nussbaum.

(For context, see sections 3.3 and 5.3 .)

We aim at demonstrating the suitability of Distem to evaluate Exascale and Cloud runtime environments providing load balancing and fault tolerance features. In that context, we reproduced some experiments published at CCGrid'2013 on Charm++ load balancers. Preliminary results are promising, and we hope that this will lead to collaborations with runtime developers.

A publication presenting how Distem to test HPC runtimes (scalability, fault tolerance and load balancing capabilities) is in the works.

6.3.2.3. Further improvements to XPFlow

Participants: Tomasz Buchert, Lucas Nussbaum, Jens Gustedt.

(For context, see sections 3.3 and 5.6 .)

We strengthened our XPFlow experiment control system using several sets of experiments, including experiments on the OpenStack IaaS Cloud stack on hundreds of Grid'5000 nodes.

A publication describing XPFlow was submitted to CCGrid'2014[21].

6.3.2.4. Further improvements to Kadeploy

Participants: Luc Sarzyniec, Emmanuel Jeanvoine, Lucas Nussbaum.

(For context, see sections 3.3 and 5.5 .)

We continued the development of Kadeploy:

- The support for multi-partition images was added;
- The communication interface between the Kadeploy server and the Kadeploy client was completely rewritten to use a REST API;
- A test framework, integrated with Inria's Continuous Integration facility, was added.

Two new Kadeploy releases were published during 2013, including those changes.

6.3.2.5. Grid'5000

Participants: Sébastien Badia, Luc Sarzyniec, Émile Morel, Lucas Nussbaum.

(For context, see sections 3.3 and 5.7 .)

The team continued to support Grid'5000. Highlights of 2013 include:

- Lucas Nussbaum is now a member of the *Bureau* and *Comité d'Architectes* of GIS Grid'5000. In the context of the *Comité d'Architectes*, he led the writing on several internal documents (on possible evolutions of the testbed).
- An article describing Grid'5000's support for experiments on IaaS Clouds[15] was published at the *Testing The Cloud* workshop.
- A new cluster, *graphite*, was installed in Nancy.

6.3.3. Convergence and co-design of experimental methodologies

6.3.3.1. Practical study on combining experimental methodologies

Participants: Maximiliano Geier, Lucas Nussbaum, Martin Quinson.

During an internship, we explored how simulation, emulation and experimentation on Grid'5000 could be combined in practice. Starting with a simple question on a particular system, we used a representative framework for each methodology: SimGrid for simulation, Distem for emulation and Grid'5000 for experimentation, and described our experiments using the workflow logic provided by the XPFlow tool. We identified a set of pitfalls in each paradigm that experimenters may encounter regarding models, platform descriptions and others. We proposed a set of general guidelines to avoid these pitfalls. We showed these guidelines may lead to accurate simulation results. Finally, we provided some insight to framework developers in order to improve the tools and thus facilitate this convergence.

The results of this work were published at the *WATERS* workshop[17].

6.3.3.2. Organization of an event on reproducible research

Participant: Lucas Nussbaum.

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

ALPINES Team

6. New Results

6.1. Integral equations on multi-screens

We developed a new functional framework for the study of scalar wave scattering by objects, called multi-screens, that are arbitrary arrangements of thin panels of impenetrable materials. From a geometric point of view, multi-screens are a priori non-orientable non-Lipschitz surfaces. We use our new framework to study boundary integral formulations of the scattering by such objects.

6.2. Second-kind Galerkin boundary element method for scattering at composite objects

In the context of scattering of time-harmonic acoustic waves at objects composed of several homogeneous parts with different material properties, a novel second-kind boundary integral formulation for this scattering problem was proposed in [X. Claeys, A single trace integral formulation of the second kind for acoustic scattering, Report 2011-14, SAM, ETH Zürich]. We recasted it into a variational problem set in L2 and investigated its Galerkin boundary element discretization from a theoretical and algorithmic point of view. Empiric studies demonstrate the competitive accuracy and superior conditioning of the new approach compared to a widely used Galerkin boundary element approach based on a first-kind boundary integral formulation.

6.3. Instability phenomenon for a rounded corner in presence of a negative material

We studied a 2D transmission problem between a positive material and a negative material. In electromagnetics, this negative material can be a metal at optical frequencies or a negative metamaterial. We highlighted an unusual instability phenomenon in some configurations: when the interface between the two materials presents a rounded corner, it can happen that the solution depends critically on the value of the rounding parameter. To prove this result, we provided an asymptotic expansion of the solution, when it is well-defined, in the geometry with a rounded corner. Then, we demonstrated that the asymptotic expansion is not stable with respect to the rounding parameter. We also conducted obtained numerical experiments with finite element methods to validate these results.

6.4. Parallel design and performance of direction preserving preconditioners

In the context of preconditioned iterative methods, our work has focused on so called direction preserving preconditioners. In [9] we consider the parallel design and performance of nested filtering factorization (NFF), a multilevel parallel preconditioning technique for solving large sparse linear systems of equations by using iterative methods. NFF is based on a recursive decomposition that requires first to permute the input matrix, which can have an arbitrary sparsity structure, into a matrix with a nested block arrow structure. This recursive factorization is a key feature in allowing NFF to have limited memory requirements and also to be very well suited for hierarchical parallel machines. NFF is also able to preserve some directions of interest of the input matrix A. Given a set of vectors T which represent the directions to be preserved, the preconditioner M satisfies a right filtering property $MT = AT$. This is a property which has been exploited in different contexts, as multigrid methods [Brandt et al., 2011, SIAM J. Sci. Comput.], semiseparable matrices [Gu et al, 2010, SIAM J. Matrix Anal. Appl.], incomplete factorizations [Wagner, 1997, Numer. Math] , or nested factorization [Appleyard and Cheshire, 1983, SPE Symposium on Reservoir Simulation]. It is well known that for difficult problems with heterogeneities or multiscale physics, the iterative methods can converge very slowly, and this is often due to the presence of several low frequency modes. By preserving the directions

corresponding to these low frequency modes in the preconditioner, their effect on the convergence is alleviated and a much faster convergence is often observed. NFF can be seen as an extension of nested factorization that can be used for matrices with arbitrary sparsity structure and for which the computation can be performed in parallel. While the algebra of NFF has been introduced previously [Grigori et al, 2010, Inria tech. report], we relate the arithmetic complexity of NFF to the depth of recursion of its decomposition, and with our data distribution and implementation, we estimate its arithmetic and communication complexity. We also discuss the convergence of NFF on a set of matrices arising from the discretization of a boundary value problem with highly heterogeneous coefficients on three-dimensional grids. Our results show that on a $400 \times 400 \times 400$ regular grid, the number of iterations with NFF increases slightly while increasing the number of subdomains up to 2048. In terms of runtime performance on Curie, a Bullx system formed by nodes of two eight-core Intel Sandy Bridge processors, NFF scales well up to 2048 cores and it is 2.6 times faster than the domain decomposition preconditioner Restricted Additive Schwarz (RAS) as implemented in PETSc http://www.mcs.anl.gov/petsc/. The choice of the filtering vectors plays an important role in direction preserving preconditioners. There are problems for which we have prior knowledge of the near kernel of the input matrix, and this is indeed the case for the problems tested in this paper. They can also be approximated by using techniques similar to the ones used in deflation, however we do not discuss further this option here.

6.5. New resuls in communication avoiding algorithms for sparse linear algebra

In the context of sparse linear algebra algorithms, our recent results focus on two operations, incomplete LU factorization preconditioners and sparse matrix-matrix multiplication.

In [12] we present a communication avoiding ILU0 preconditioner for solving large linear systems of equations by using iterative Krylov subspace methods. Recent research has focused on communication avoiding Krylov subspace methods based on so called s-step methods. However there was no communication avoiding preconditioner available yet, and this represents a serious limitation of these methods. Our preconditioner allows to perform s iterations of the iterative method with no communication, through ghosting some of the input data and performing redundant computation. It thus reduces data movement by a factor of 3s between different levels of the memory hierarchy in a serial computation and between different processors in a parallel computation. To avoid communication, an alternating reordering algorithm is introduced for structured and unstructured matrices, that requires the input matrix to be ordered by using a graph partitioning technique such as k-way or nested dissection. We show that the reordering does not affect the convergence rate of the ILU0 preconditioned system as compared to k-way or nested dissection ordering, while it reduces data movement and should improve the expected time needed for convergence. In addition to communication avoiding Krylov subspace methods, our preconditioner can be used with classical methods such as GMRES or s-step methods to reduce communication.

In [6] we consider a fundamental problem in combinatorial and scientific computing, the sparse matrix-matrix multiplication problem. Obtaining scalable algorithms for this operations is difficult, since this operation has a poor surface to volume ratio, that is a poor data re-use. We consider that the input matrices are random, corresponding to Erdos-Renyi random graphs. We determine new lower bounds on communication for this case, in which we assume that the algorithm is sparsity independent, where the computation is statically partitioned to processors independent of the sparsity structure of the input matrices. We show in this paper that existing algorithms for sparse matrix-matrix multiplication are sub-optimal in their communication costs, and we obtain new algorithms which are communication optimal, communicating less than the previous algorithms and matching new lower bounds.

6.6. New resuls in communication avoiding algorithms for dense linear algebra

In the context of dense linear algebra algorithms, we have focused on two operations, LU factorization and rank revealing QR factorization.

In [4] we present block LU factorization with panel rank revealing pivoting (block LU_PRRP), a decomposition algorithm based on strong rank revealing QR panel factorization. Block 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, n is the number of columns of the matrix, and for simplicity we assume that n is a multiple of b. We also assume throughout all the paper that $2 \leq b \leq n$. For example, if the size of the panel is $b = 64$, and $\tau = 2$, then $(1+2b)^{(n/b)-1} = (1.079)^{n-64} \ll 2^{n-1}$, 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 block LU_PRRP factorization does only $O(n^2 b)$ additional floating point operations compared to GEPP.

We also present block CALU_PRRP, a version of block LU_PRRP that minimizes communication, and 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. Block 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 $2^{n(H+1)-1}$. Block CALU_PRRP is also more stable in practice and is resistant to pathological cases on which GEPP and CALU fail.

We have also introduced CARRQR (paper submitted to SIAM Journal on Matrix Analysis and Applications), a communication avoiding rank revealing QR factorization with tournament pivoting. We show that CARRQR reveals the numerical rank of a matrix in an analogous way to QR factorization with column pivoting (QRCP). Although the upper bound of a quantity involved in the characterization of a rank revealing factorization is worse for CARRQR than for QRCP, our numerical experiments on a set of challenging matrices show that this upper bound is very pessimistic, and CARRQR is an effective tool in revealing the rank in practical problems. Our main motivation for introducing CARRQR is that it minimizes data transfer, modulo polylogarithmic factors, on both sequential and parallel machines, while previous factorizations as QRCP are communication sub-optimal and require asymptotically more communication than CARRQR. Hence CARRQR is expected to have a better performance on current and future computers, where commmunication is a major bottleneck that highly impacts the performance of an algorithm.

6.7. Scalable Schwarz domain decomposition methods

Domain decomposition methods are, alongside multigrid methods, one of the dominant paradigms in contemporary large-scale partial differential equation simulation. A lightweight implementation [8] of a theoretically and numerically scalable preconditioner was developped in the context of overlapping methods. The performance of this work is assessed by numerical simulations executed on thousands of cores, for solving various highly heterogeneous elliptic problems in both 2D and 3D with billions of degrees of freedom. Such problems arise in computational science and engineering, in solid and fluid mechanics.
For example, in 3D, the initial highly heterogeneous problem of 74 million unknowns is solved in 200 seconds on 512 threads. Using 16384 threads, the problem is now made of approximately 2.3 billions unknowns, and it is solved in 215 seconds, which yields an efficiency of $\approx 90\%$. In 2D, the initial problem of 695 million unknowns is solved in 175 seconds on 512 threads. Using 16384 threads, the problem is now made of approximately 22.3 billions unknowns, and it is solved in 187 seconds, which yields an efficiency of $\approx 96\%$.

6.8. Schur domain decomposition methods

We have introduced spectral coarse spaces for the BDD and FETI methods in [5]. These coarse spaces are specifically designed for the two-level methods to be scalable and robust with respect to the coefficients in the equation and the choice of the decomposition. We achieve this by solving generalized eigenvalue problems on the interfaces between subdomains to identify the modes which slow down convergence. Theoretical bounds for the condition numbers of the preconditioned operators which depend only on a chosen threshold and the maximal number of neighbours of a subdomain were proved. For FETI there are two versions of the two-level method: one based on the full Dirichlet preconditioner and the other on the, cheaper, lumped preconditioner. Some numerical tests confirm these results.

6.9. Non conforming domain decomposition methods

We have designed and analyzed a new non-conforming domain decomposition method, named the NICEM method, based on Schwarz-type approaches that allows for the use of Robin interface conditions on non-conforming grids. The method is proven to be well posed. The error analysis is performed in 2D and in 3D for P1 elements. Numerical results in 2D illustrate the new method. This work is in collaboration with C. Japhet and Y. Maday.

6.10. Quadratic finite elements with non-matching grids for the unilateral boundary contact

We analyze in [3] a numerical model for the Signorini unilateral contact, based on the mortar blue method, in the quadratic finite element context. The mortar frame enables one to use non-matching grids and brings facilities in the mesh generation of different components of a complex system. The convergence rates we state here are similar to those already obtained for the Signorini problem when discretized on conforming meshes. The matching for the unilateral contact driven by mortars preserves then the proper accuracy of the quadratic finite elements. This approach has already been used and proved to be reliable for the unilateral contact problems even for large deformations. We provide however some numerical examples to support the theoretical predictions with FreeFem++ (http://www.freefem.org/ff++).

AVALON Team

6. New Results

6.1. Energy efficiency of large scale distributed systems

Participants: Ghislain Landry Tsafack Chetsa, Mohammed El Mehdi Diouri, Jean-Patrick Gelas, Olivier Glück, Laurent Lefèvre, François Rossigneux.

6.1.1. Analysis and Evaluation of Different External and Internal Power Monitoring Devices for a Server and a Desktop Machine

Large-scale distributed systems (e.g., datacenters, HPC systems, clouds, large-scale networks, etc.) consume and will consume enormous amounts of energy. Therefore, accurately monitoring the power and energy consumption of these systems is increasingly more unavoidable. The main novelty of this contribution [15] is the analysis and evaluation of different external and internal power monitoring devices tested using two different computing systems, a server and a desktop machine. Furthermore, we also provide experimental results for a variety of benchmarks which exercise intensively the main components (CPU, Memory, HDDs, and NICs) of the target platforms to validate the accuracy of the equipment in terms of power dispersion and energy consumption. We highlight that external wattmeters do not offer the same measures as internal wattmeters. Thanks to the high sampling rate and to the different measured lines, the internal wattmeters allow an improved visualization of some power fluctuations. However, a high sampling rate is not always necessary to understand the evolution of the power consumption during the execution of a benchmark.

6.1.2. Your Cluster is not Power Homogeneous

Future supercomputers will consume enormous amounts of energy. These very large scale systems will gather many homogeneous clusters. We analyze the power consumption of the nodes from different homogeneous clusters during different workloads. As expected, we observe that these nodes exhibit the same level of performance. However, we also show that different nodes from a homogeneous cluster may exhibit heterogeneous idle power energy consumption even if they are made of identical hardware. Hence, we propose an experimental methodology to understand such differences. We show that CPUs are responsible for such heterogeneity which can reach 20% in terms of energy consumption. So energy aware (Green) schedulers must take care of such hidden heterogeneity in order to propose efficient mapping of tasks. To consume less energy, we propose an energy-aware scheduling approach taking into account the heterogeneous idle power consumption of homogeneous nodes [20]. It shows that we are able to save energy up to 17% while exploiting the high power heterogeneity that may exist in some homogeneous clusters.

6.1.3. Energy Consumption Estimations of Fault Tolerance protocols

Energy consumption and fault tolerance are two interrelated issues to address for designing future exascale systems. Fault tolerance protocols used for checkpointing have different energy consumption depending on parameters like application features, number of processes in the execution and platform characteristics. Currently, the only way to select a protocol for a given execution is to run the application and monitor the energy consumption of different fault tolerance protocols. This is needed for any variation of the execution setting. To avoid this time and energy consuming process, we propose an energy estimation framework [16], [17], [7]. It relies on an energy calibration of the considered platform and a user description of the execution setting. We evaluate the accuracy of our estimations with real applications running on a real platform with energy consumption monitoring. Results show that our estimations are highly accurate and allow selecting the best fault tolerant protocol without pre-executing the application.

6.1.4. Energy Consumption Estimations of Data Broadcasting

Future supercomputers will gather hundreds of millions of communicating cores. The movement of data in such systems will be very energy consuming. We address the issue of energy consumption of data broadcasting in such large scale systems. To this end, in [19], [7], we propose a framework to estimate the energy consumed by different MPI broadcasting algorithms for various execution settings. Validation results show that our estimations are highly accurate and allow to select the least consuming broadcasting algorithm.

6.1.5. A Smart-Grid Based Framework for Consuming Less and Better in Extreme-Scale Infrastructures

As they will gather hundreds of million cores, future exascale supercomputers will consume enormous amounts of energy. Besides being very important, their power consumption will be dynamic and irregular. Thus, in order to consume energy efficiently, powering such systems will require a permanent negotiation between the energy supplier and one of its major customers represented by exascale platforms. We have designed SESAMES [18], [53], a smart and energy-aware service-oriented architecture manager that proposes energy-efficient services for exascale applications and provides an optimized reservation scheduling. The new features of this framework are the design of a smart grid and a multi-criteria green job scheduler. Simulation results show that with the proposed multi-criteria job scheduler, we are able to save up to 2.32 % in terms of energy consumption, 24.22 % in terms of financial cost and reduce up to 7.12 % the emissions of CO_2.

6.1.6. Clustered Virtual Home Gateway (vHGW)

This result is a joint work between Avalon team (J.P. Gelas, L. Lefevre) and Addis Abeba University (M. Tsibie and T. Assefa). The customer premises equipment (CPE), which provides the interworking functions between the access network and the home network, consumes more than 80% of the total power in a wireline access network. In the GreenTouch initiative (cf Section 7.3), we aim at a drastic reduction of the power consumption by means of a passive or quasi-passive CPE. Such approach requires that typical home gateway functions, such as routing, security, and home network management, are moved to a virtual home gateway (vHGW) server in the network. In our first prototype virtual home gateways of the subscribers were put in LXC containers on a unique GNU/Linux server. The container approach is more scalable than separating subscribers by virtual machines. We demonstrated a sharing factor of 500 to 1000 virtual home gateways on one server, which consumes about 150 W, or 150 to 300 mW per subscriber. Comparing this power consumption with the power of about 2 W for the processor in a thick client home gateway, we achieved an efficiency gain of 5-10x. The prototype was integrated and demonstrated at TIA 2012 in Dallas. In our current work, we propose the Clustered vHGWs Data center architecture to yield optimal energy conservation through virtual machine's migration among physical nodes based on the current subscriber's service access state, while ensuring SLA respective subscribers. Thus, optimized energy utilization of the data center is assured without compromising the availability of service connectivity and QoS preferences of respective subscribers.

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

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

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

6.1.8. Reservation based Usage for Energy Efficient Clouds: the Climate Architecture

The FSN XLcloud project (cf Section 7.1) strives to establish the demonstration of a High Performance Cloud Computing (HPCC) platform based on OpenStask, that is designed to run a representative set of compute intensive workloads, including more specifically interactive games, interactive simulations and 3D graphics. XLcloud is based on OpenStack, and Avalon is contributing to the energy efficiency part of this project. We have proposed and brought our contribution to Climate, a new resource reservation framework for OpenStack, developed in collaboration with Bull, Mirantis and other OpenStack contributors. Climate allows the reservation of both physical and virtual resources, in order to provide a mono-tenancy environment suitable for HPC applications. Climate chooses the most efficient hosts (flop/W). This metric is computed from the CPU / GPU informations, mixed with real power consumption measurements provided by the Kwapi framework. The user requirements may be loose, allowing Climate to choose the best time slot to place the reservation. Climate will be improved with standby mode features, to shut down automatically the unused hosts. The first release of Climate is planned at the end of January 2014, and we expect an incubation in the next version of OpenStack.

6.2. Simulation of Large Scale Distributed Systems

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

6.2.1. Toward Better Simulation of MPI Applications on Ethernet/TCP Networks

Simulation and modeling for performance prediction and profiling is essential for developing and maintaining HPC code that is expected to scale for next-generation exascale systems, and correctly modeling network behavior is essential for creating realistic simulations. In [11], we proposed an implementation of a flow-based hybrid network model that accounts for factors such as network topology and contention, which are commonly ignored by other approaches. We focused on large-scale, Ethernet-connected systems, as these currently compose 37.8% of the TOP500 index, and this share is expected to increase as higher-speed 10 and 100GbE become more available. The European Mont-Blanc project that studies exascale computing by developing prototype systems with low-power embedded devices will also use Ethernet-based interconnect. Our model is implemented within SMPI, an open-source MPI implementation that connects real applications to the SIMGRID simulation framework (cf Section 5.5). SMPI provides implementations of collective communications based on current versions of both OpenMPI and MPICH. SMPI and SIMGRID also provide methods for easing the simulation of large-scale systems, including shadow execution, memory folding, and support for both online and offline simulation. We validated our proposed model by comparing traces produced by SMPI with those from real world experiments, as well as with those obtained using other established network models. Our study shows that SMPI has a consistently better predictive power than classical LogP-based models for a wide range of scenarios including both established HPC benchmarks and real applications.

6.2.2. SimGrid: a Sustained Effort for the Versatile Simulation of Large Scale Distributed Systems

SIMGRID (cf Section 5.5) is a toolkit for the versatile simulation of large scale distributed systems, whose development effort has been sustained for the last fifteen years. Over this time period SIMGRID has evolved from a one-laboratory project in the U.S. into a scientific instrument developed by an international collaboration. The keys to making this evolution possible have been securing of funding, improving the quality of the software, and increasing the user base. We detailed in [55]how we have been able to make advances on all three fronts, on which we plan to intensify our efforts over the upcoming years.

6.2.3. Simulating Multiple Clouds from a Client Point of View: SGCB an AWS Simulator

Validating a new application over a Cloud is not an easy task and it can be costly over public Clouds. Simulation is a good solution if the simulator is accurate enough and if it provides all the features of the target Cloud. In [49], we have proposed an extension of the SIMGRID simulation toolkit to simulate the Amazon IaaS Cloud. Based on an extensive study of the Amazon platform and previous evaluations, we have integrated models into the SIMGRID Cloud Broker and exposed the same API as Amazon to the users. Our experimental results have shown that our simulator is able to simulate different parts of Amazon for different applications.

6.3. Active Data: A Data-Centric Approach to Data Life-Cycle Management

Participants: Gilles Fedak, Anthony Simonet.

Data-intensive science offers new opportunities for innovation and discoveries, provided that large datasets can be handled efficiently. Data management for data-intensive science applications is challenging; requiring support for complex data life cycles, coordination across multiple sites, fault tolerance, and scalability to support tens of sites and petabytes of data. In [28], we argue that data management for data-intensive science applications requires a fundamentally different management approach than the current ad-hoc task centric approach. We propose Active Data, a fundamentally novel paradigm for data life cycle management. Active Data follows two principles: data-centric and event-driven. We report on the Active Data programming model and its preliminary implementation, and discuss the benefits and limitations of the approach on recognized challenging data-intensive science use-cases.

6.4. HPC Component Model

Participants: Zhengxiong Hou, Vincent Lanore, Christian Perez.

6.4.1. Auto-tuning of Stencil Based Applications

We have finished designing a tuning approach for stencil applications on multi-core clusters [25]. We focused in particular on a 2D Jacobi benchmark application as well as memory bandwidth performance. The tuning approach includes 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 the Curie supercomputer.

6.4.2. Static 2D FFT Adaptation through a Component Model based on Charm++

Adaptation algorithms for HPC applications can improve performance but their implementation is often costly in terms of development and maintenance. Component models such as Gluon++, which is built on top of Charm++, propose to separate the business code, encapsulated in components, and the application structure, expressed through a component assembly. Adaptation of component-based HPC applications can be achieved through the optimization of the assembly. We have studied such an approach with the adaptation to network topology and data size of a Gluon++ 2D FFT application. Preliminary experimental results obtained on the GRID'5000 platform show the suitability of the proposed approach.

6.4.3. Towards Scalable Reconfiguration in Component Models

Some HPC applications require reconfiguration of their architecture at runtime; examples include adapting to (cloud) resource elasticity, efficient distributed deployment, Adaptive Mesh Refinement (AMR), and load balancing. This class of applications raises challenges such as handling of concurrent reconfigurations and distributed architecture representation at runtime. To our knowledge, no existing programming model addresses those challenges in the general case with both high programmability and scalability. We have identified a list of specific subproblems and use-cases and we have devised a preliminary component model to address some of them.

6.5. Resource Management and Scheduling

Participants: Eddy Caron, Frédéric Desprez, Gilles Fedak, Jose Luis Lucas, Christian Perez, Jonathan Rouzaud-Cornabas, Frédéric Suter.

6.5.1. Resource Management Architecture for Fair Scheduling of Optional Computations

Most High-Performance Computing platforms require users to submit a pre-determined number of computation requests (also called jobs). Unfortunately, this is cumbersome when some of the computations are optional, i.e., they are not critical, but their completion would improve results. For example, given a deadline, the number of requests to submit for a Monte Carlo experiment is difficult to choose. The more requests are completed, the better the results are, however, submitting too many might overload the platform. Conversely, submitting too few requests may leave resources unused and misses an opportunity to improve the results.

In cooperation with IRIT (Toulouse), we have proposed a generic client-server architecture and an implementation in DIET, a production GridRPC middleware, which auto-tunes the number of requests [12]. Real-life experiments show significant improvement of several metrics, such as user satisfaction, fairness and the number of completed requests. Moreover, the solution is shown to be scalable.

6.5.2. Advanced Promethee-based Scheduler Enriched with User-Oriented Methods

Efficiently scheduling tasks in hybrid Distributed Computing Infrastructures (DCI) is a challenging pursue because the scheduler must deal with a set of parameters that simultaneously characterize the tasks and the hosts originating from different types of infrastructure. In [27], we propose a scheduling method for hybrid DCIs, based on advanced multi-criteria decision methods. The scheduling decisions are made using pairwise comparisons of the tasks for a set of criteria like expected completion time and price charged for computation. The results are obtained with an XtremWeb-like pull-based scheduler simulator using real failure traces for a combination of three types of infrastructure. We also show how such a scheduler should be configured to enhance user satisfaction regardless their profiles, while maintaining good values for makespan and cost. We validate our approach with a statistical analysis on empirical data and show that our proposed scheduling method improves performance by 12-17% compared to other scheduling methods. Experimenting on large time-series and using realistic scheduling scenarios lead us to conclude about time consistency results of the method.

6.5.3. Fair Resource Sharing for Dynamic Scheduling of Workflows on Heterogeneous Systems

Scheduling independent workflows on shared resources in a way that satisfy users Quality of Service is a significant challenge. In [37], we described methodologies for off-line scheduling, where a schedule is generated for a set of known workflows, and on-line scheduling, where users can submit workflows at any moment in time. We consider the on-line scheduling problem in more detail and present performance comparisons of state-of-the-art algorithms for a realistic model of a heterogeneous system.

6.5.4. Image Transfer and Storage Cost Aware Brokering Strategies for Multiple Clouds

Nowadays, Clouds are used for hosting a large range of services. But between different Cloud Service Providers, the pricing model and the price of individual resources are very different. Furthermore hosting a service in one Cloud is the major cause of service outage. To increase resiliency and minimize the monetary cost of running a service, it becomes mandatory to span it between different Clouds. Moreover, due to dynamicity of both the service and Clouds, it could be required to migrate a service at run time. Accordingly, this ability must be integrated into the multi-Cloud resource manager, *i.e.* the Cloud broker. But, when migrating a VM to a new Cloud Service Provider, the VM disk image must be migrated too. Accordingly, data storage and transfer must be taken into account when choosing if and where an application will be migrated.

In [47], we have extended a cost-optimization algorithm to take into account storage costs to approximate the optimal placement of a service. The data storage management consists in taking two decisions: where to upload an image, and keep it on-line during the experiment lifetime or delete it when unused. Based on our experimentations, we have shown that the storage cost of VM disk image must not be neglected as done in previous work. Moreover, we have shown that using the accurate combinations of storage policies can dramatically reduce the storage cost (from 90% to 14% of the total bill).

6.6. Security for Virtualization and Clouds

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

6.6.1. Improving Users' Isolation in IaaS: Virtual Machine Placement with Security Constraints

Nowadays virtualization is used as the sole mechanism to isolate different users on Cloud platforms. Due to improper virtualization of micro-architectural components, data leak and modification can occur on public Clouds. Moreover, using the same attack vector (improper virtualization of micro-architectural components), it is possible to induce performance interferences, *i.e.* noisy neighbors. Using this approach, a VM can slow down and steal resources from concurrent VMs. In [43], we have proposed placement heuristics that take into account isolation requirements. We have modified three classical heuristics to take into account these requirements. Furthermore, we have proposed four new heuristics that take into account the hierarchy of the Cloud platforms and the isolation requirements. Finally, we have evaluated these heuristics and compare them with the modified classical ones. We have shown that our heuristics are performing at least as good as classical ones but are scaling better and are faster by a few order of magnitude than the classical ones.

6.6.2. Security for Cloud Environment through Information Flow Properties Formalization with a First-Order Temporal Logic

The main slowdown of Cloud activity comes from the lack of reliable security. The on-demand security concept aims at delivering and enforcing the client's security requirements. In [50], we have presented an approach, Information Flow Past Linear Time Logic (IF-PLTL), to specify how a system can support a large range of security properties. We have presented how to control those information flows from lower system events. We have given complete details over IF-PLTL syntax and semantics. Furthermore, that logic enables to formalize a large set of security policies. Our approach is exemplified with the Chinese Wall commercial-related policy. Finally, we have discussed the extension of IF-PLTL with dynamic relabeling to encompass more realistic situations through the dynamic domains isolation policy.

6.6.3. Security Metrics for the Cloud Computing and Security-aware Virtual Machine Placement

In a classic Cloud Computing scenario, a client connects to a provider platform/service and submits his computation requirements, sometimes known as Service Level Agree- ments (SLAs). Then, the platform executes the computation taking into account, in its allocation algorithms, criteria like data location, CPU usage or duration of a job. As security in Cloud Computing is a main concern, we propose to consider security as another criteria for jobs scheduling. Thus, two questions need to be answered. The first one is how a client

can describe his needs in terms of security level and the second one is how the scheduler could leverage the security to satisfy the client requirements? To provide an answer, a system of security metrics is essential. Indeed, with appropriate metrics, we can quantify and compare the security level of our resources. Moreover, a client can easily describe his security requirements and the scheduler can allocate the fitted resources using these metrics. Unfortunately, such system of metrics is not yet available. Consequently, we developed a system of security metrics specific to the Cloud Computing and scheduling algorithms using these metrics for a Security-Aware Virtual Machine (VM) placement.

6.7. Self-healing of Operational Issues for Grid Computing

Participant: Frédéric Desprez.

Many scientists now formulate their computational problems as scientific workflows. Workflows allow researchers to easily express multi-step computational task. However, their large scale and the number of middleware systems involved in these gateways lead to many errors and faults. Fair quality of service (QoS) can be delivered, yet with important human intervention. Automating such operations is challenging for two reasons. First, the problem is online by nature because no reliable user activity prediction can be assumed, and new workloads may arrive at any time. Therefore, the considered metrics, decisions and actions have to remain simple and to yield results while the application is still executing. Second, it is non-clairvoyant due to the lack of information about applications and resources in production conditions. Computing resources are usually dynamically provisioned from heterogeneous clusters, clouds or desktop grids without any reliable estimate of their availability and characteristics. Models of application execution times are hardly available either, in particular on heterogeneous computing resources.

In collaboration with Rafaël Silva and Tristan Glatard, we proposed a general self-healing process for autonomous detection and handling of operational incidents in scientific workflow executions on grids. Instances are modeled as Fuzzy Finite State Machines (FuSM) where state degrees of membership are determined by an external healing process. Degrees of membership are computed from metrics assuming that incidents have outlier performance, e.g. a site or a particular invocation behaves differently than the others. These metrics make little assumptions on the application or resource characteristics. Based on incident degrees, the healing process identifies incident levels using thresholds determined from the platform history. A specific set of actions is then selected from association rules among incident levels. The healing process is parametrized on real application traces acquired in production on the European Grid Infrastructure (EGI).

To optimize task granularity in distributed scientific workflows, we presented a method that groups tasks when the fineness degree of the application becomes higher than a threshold determined from execution traces. Controlling the granularity of workflow activities executed on grids is required to reduce the impact of task queuing and data transfer time. Our method groups tasks when the fineness degree of the application, which takes into account the ratio of shared data and the queuing/round-trip time ratio, becomes higher than a threshold determined from execution traces. The algorithm also de-groups task groups when new resources arrive. Results showed that under stationary load, our fineness control process significantly reduces the makespan of all applications. Under non-stationary load, task grouping is penalized by its lack of adaptation, but our de-grouping algorithm corrects it in case variations in the number of available resources are not too fast [21].

To address unfairness among workflow executions, we proposed an algorithm to fairly allocate distributed computing resources among workflow executions to multi-user platforms. We consider a non-clairvoyant, online fairness problem where the platform workload, task costs, and resource characteristics are unknown and not stationary. We define a novel metric that quantifies unfairness based on the fraction of pending work in a workflow. It compares workflow activities based on their ratio of queuing tasks, their relative durations, and the performance of resources where tasks are running, as information becomes available during the execution. Our method is implemented and evaluated on 4 different applications executed in production conditions on EGI. Results show that our method can very significantly reduce the standard deviation of the slowdown, and the average value of our metric [22].

CEPAGE Project-Team

6. New Results

6.1. Resource allocation and Scheduling

6.1.1. Broadcasting on Large Scale Heterogeneous Platforms under the Bounded Multi-Port Model

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

In [17], we consider the problem of broadcasting a large message in a large scale distributed network under the multi-port communication model. We are interested in building an overlay network, with the aim of maximizing the throughput and minimizing the degree of the participating nodes. We consider a classification of participating nodes into two parts: open nodes that stay in the open-Internet and "guarded" nodes that lie behind firewalls or NATs, with the constraint that two guarded nodes cannot communicate directly. Without guarded nodes, we prove that it is possible to reach the optimal throughput with a quasi-optimal (up to a small additive increase) degree of the participating nodes. In presence of guarded nodes, we provide a closed form formula for the optimal cyclic throughput and we observe that the optimal solution may require arbitrarily large degrees. In the acyclic case, we propose an algorithm that reaches the optimal acyclic throughput with low degree. Then, we prove a worst case 5/7 ratio between the optimal acyclic and cyclic throughput and show through simulations that this ratio is on average very close to 1, what makes acyclic solutions efficient both in terms of throughput maximization and degree minimization.

6.1.2. 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 [32], 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)

6.1.3. Reliable Service Allocation in Clouds

Participants: Olivier Beaumont, Lionel Eyraud-Dubois, Hubert Larchevêque, Paul Renaud-Goud, Philippe Duchon.

In [30], 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.

In [29], we extend this work to an energy minimisation framework, by considering two energy consumption models based on DVFS techniques, where the clock frequency of physical resources can be changed with a Dynamic Voltage and Frequency Scaling (DVFS) method. For each allocation problem and each energy model, we prove deterministic approximation ratios on the consumed energy for algorithms that provide guaranteed probability failures, as well as an efficient heuristic, whose energy ratio is not guaranteed.

In [37], we study the robustness of an allocation of Virtual Machines (VM) on a set of Physical Machines (PM) when the resource demand of the VMs can change over time. This may imply 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.

6.1.4. Splittable Single Source-Sink Routing on CMP Grids: A Sublinear Number of Paths Suffice

Participants: Adrian Kosowski, Przemyslaw Uznanski.

In [44], we study single chip multiprocessors (CMP) with grid topologies, where a significant part of power consumption is attributed to communications between the cores of the grid. We investigate the problem of routing communications between CMP cores using shortest paths, in a model in which the power cost associated with activating a communication link at a transmission speed of f bytes/second is proportional to f^α, for some constant exponent $\alpha > 2$. Our main result is a trade-off showing how the power required for communication in CMP grids depends on the ability to split communication requests between a given pair of node, routing each such request along multiple paths. For a pair of cores in a $m \times n$ grid, the number of available communication paths between them grows exponentially with n, m. By contrast, we show that optimal power consumption (up to constant factors) can be achieved by splitting each communication request into k paths, starting from a threshold value of $k = \Theta(n^{1/(\alpha-1)})$. This threshold is much smaller than n for typical values of $\alpha \approx 3$, and may be considered practically feasible for use in routing schemes on the grid. More generally, we provide efficient algorithms for routing multiple k-splittable communication requests between two cores in the grid, providing solutions within a constant approximation of the optimum cost. We support our results with algorithm simulations, showing that for practical instances, our approach using k-splittable requests leads to a power cost close to that of the optimal solution with arbitrarily splittable requests, starting from the stated threshold value of k.

6.1.5. Maximum matching in multi-interface networks

Participants: Adrian Kosowski, Dominik Pajak.

In [26], we consider the standard matching problem in the context of multi-interface wireless networks. In heterogeneous networks, devices can communicate by means of multiple wireless interfaces. By choosing which interfaces to switch on at each device, several connections might be established. That is, the devices at the endpoints of each connection share at least one active interface. In the studied problem, the aim is to maximize the number of parallel connections without incurring in interferences. Given a network $G = (V, E)$, nodes V represent the devices, edges E represent the connections that can be established. If node x participates in the communication with one of its neighbors by means of interface i, then another neighboring node of x can establish a connection (but not with x) only if it makes use of interface $j \neq i$. The size of a solution for an instance of the outcoming matching problem, that we call *Maximum Matching in Multi-Interface* networks (MMMI for short), is always in between the sizes of the solutions for the same instance with respect to the standard matching and its induced version problems. However, we prove that MMMI is NP-hard even for proper interval graphs and for bipartite graphs of maximum degree $\Delta \geq 3$. We also show polynomially solvable cases of MMMI with respect to different assumptions.

6.1.6. Parallel scheduling of task trees with limited memory

Participant: Lionel Eyraud-Dubois.

In a paper submitted to ACM TOPC, we have investigated the execution of tree-shaped task graphs using multiple processors. Each edge of such a tree represents some large data. A task can only be executed if all input and output data fit into memory, and a data can only be removed from memory after the completion of the task that uses it as an input data. Such trees arise, for instance, in the multifrontal method of sparse matrix factorization. The peak memory needed for the processing of the entire tree depends on the execution order of the tasks. With one processor the objective of the tree traversal is to minimize the required memory. This problem was well studied and optimal polynomial algorithms were proposed. We have extended the problem by considering multiple processors, which is of obvious interest in the application area of matrix factorization. With multiple processors comes the additional objective to minimize the time needed to traverse the tree, i.e., to minimize the makespan. Not surprisingly, this problem proves to be much harder than the sequential one. We study the computational complexity of this problem and provide inapproximability results even for unit weight trees. We design a series of practical heuristics achieving different trade-offs between the minimization of peak memory usage and makespan. Some of these heuristics are able to process a tree while keeping the memory usage under a given memory limit. The different heuristics are evaluated in an extensive experimental evaluation using realistic trees.

6.1.7. Point-to-point and congestion bandwidth estimation: experimental evaluation on PlanetLab

Participants: Lionel Eyraud-Dubois, Przemyslaw Uznanski.

In large scale Internet platforms, measuring the available bandwidth between nodes of the platform is difficult and costly. However, having access to this information allows to design clever algorithms to optimize resource usage for some collective communications, like broadcasting a message or organizing master/slave computations. In [54], we analyze the feasibility to provide estimations, based on a limited number of measurements, for the point-to-point available bandwidth values, and for the congestion which happens when several communications take place at the same time. We present a dataset obtained with both types of measurements performed on a set of nodes from the PlanetLab platform. We show that matrix factorization techniques are quite efficient at predicting point-to-point available bandwidth, but are not adapted for congestion analysis. However, a LastMile modeling of the platform allows to perform congestion predictions with a reasonable level of accuracy, even with a small amount of information, despite the variability of the measured platform.

6.1.8. Parallel Mining of Functional Dependencies

Participants: Sofian Maabout, Nicolas Hanusse.

The problem of extracting functional dependencies (FDs) from databases has a long history dating back to the 90's. Still, efficient solutions taking into account both material evolution, namely the advent of multicore machines, and the amount of data that are to be mined, are still needed. In [46] we propose a parallel algorithm which, upon small modifications, extracts (i) the minimal keys, (ii) the minimal exact FDs, (iii) the minimal approximate FDs and (iv) the Conditional functional dependencies (CFDs) holding in a table. Under some natural conditions, we prove a theoretical speed up of our solution with respect to a baseline algorithm which follows a depth first search strategy. Since mining most of these dependencies require a procedure for computing the number of distinct values (NDV) which is a space consuming operation, we show how sketching techniques for estimating the exact value of NDV can be used for reducing both memory consumption as well as communications overhead when considering distributed data while guaranteeing a certain quality of the result. Our solution is implemented in both shared, using C++ and OpenMP, and distributed memory, using Hadoop implementation of Map-Reduce. The experimental results showthe efficiency and scalability of our proposal. Most notably, the theoretical speed ups are confirmed by the experiments.

6.1.9. *Fast Skyline Query Evaluation with Skycuboids Materialization based on Functional Dependencies*

Participants: Sofian Maabout, Nicolas Hanusse.

Ranking multidimensional data via different Skyline queries gives rise to the so called skycube structure. Most of previous work on optimizing the subspaces skyline queries have concentrated on full materialization of the skycube. Due to the exponential number of skylines one must pre-compute, the full materialization is unfeasible in practice. However, due to the non monotonic nature of skylines, there is no immediate inclusion relationship between the skycuboids when we have an inclusion of the dimensions. This makes the partial materialization harder. In this paper, we identify sufficient conditions for establishing inclusions between skycuboids thanks to the functional dependencies that hold in the underlying data. This leads to the characterization of a *minimal* set of skycuboids to be materialized in order to answer all the possible skyline queries without resorting to the underlying data. We conduct an extensive set of experiments showing that with the help of a small fraction of the skycube, we can efficiently answer all the possible skyline queries. In addition, our proposal turns to be helpful even in the full materialization setting. Indeed, thanks to the inclusions we identify, we devise a full materialization algorithm which outperforms state of the art skycube computation algorithms especially when data and dimensions get large. The results are reported in the technical report submitted to SIGMOD'14.

6.2. Compact Routing

6.2.1. *On the Communication Complexity of Distributed Name-Independent Routing Schemes*

Participants: Cyril Gavoille, Nicolas Hanusse, David Ilcinkas.

In [38], we present a distributed asynchronous algorithm that, for every undirected weighted n-node graph G, constructs name-independent routing tables for G. The size of each table is $\widetilde{O}(\sqrt{n})$, whereas the length of any route is stretched by a factor of at most 7 w.r.t. the shortest path. At any step, the memory space of each node is $\widetilde{O}(\sqrt{n})$. The algorithm terminates in time $O(D)$, where D is the hop-diameter of G. In synchronous scenarios and with uniform weights, it consumes $\widetilde{O}(m\sqrt{n} + n^{3/2} \min D, \sqrt{n})$ messages, where m is the number of edges of G.

In the realistic case of sparse networks of poly-logarithmic diameter, the communication complexity of our scheme, that is $\widetilde{O}(n^{3/2})$, improves by a factor of $\sqrt{n}$ the communication complexity of *any* shortest-path routing scheme on the same family of networks. This factor is provable thanks to a new lower bound of independent interest.

6.2.2. *There are Plane Spanners of Maximum Degree 4*

Participant: Nicolas Bonichon.

Let $\mathcal{E}$ be the complete Euclidean graph on a set of points embedded in the plane. Given a fixed constant $t \geq 1$, a spanning subgraph G of $\mathcal{E}$ is said to be a *t-spanner* of $\mathcal{E}$ if for any pair of vertices u, v in $\mathcal{E}$ the distance between u and v in G is at most t times their distance in $\mathcal{E}$. A spanner is *plane* if its edges do not cross.

We consider the question: "What is the smallest *maximum degree* that can be achieved for a *plane* spanner of $\mathcal{E}$?" Without the planarity constraint, it is known that the answer is 3 which is thus the best known lower bound on the degree of any plane spanner. With the planarity requirement, the best known upper bound on the maximum degree is 6, the last in a long sequence of results improving the upper bound. In this work we show that there is a constant $t \geq 1$ such that the complete Euclidean graph always contains a plane t-spanner of maximum degree 4 and make a big step toward closing the question. Our construction leads to an efficient algorithm for obtaining the spanner from Chew's L_1-Delaunay triangulation.

6.3. Mobile Agents

6.3.1. Collision-Free Network Exploration

Participants: Ralf Klasing, Adrian Kosowski, Dominik Pajak.

A set of mobile agents is placed at different nodes of a n-node network. The agents synchronously move along the network edges in a *collision-free* way, i.e., in no round may two agents occupy the same node. In each round, an agent may choose to stay at its currently occupied node or to move to one of its neighbors. An agent has no knowledge of the number and initial positions of other agents. We are looking for the shortest possible time required to complete the collision-free *network exploration*, i.e., to reach a configuration in which each agent is guaranteed to have visited all network nodes and has returned to its starting location. In [34], we first consider the scenario when each mobile agent knows the map of the network, as well as its own initial position. We establish a connection between the number of rounds required for collision-free exploration and the degree of the minimum-degree spanning tree of the graph. We provide tight (up to a constant factor) lower and upper bounds on the collision-free exploration time in general graphs, and the exact value of this parameter for trees. For our second scenario, in which the network is unknown to the agents, we propose collision-free exploration strategies running in $O(n^2)$ rounds for tree networks and in $O(n^5 \log n)$ rounds for general networks.

6.3.2. Deterministic Rendezvous of Asynchronous Bounded-Memory Agents in Polygonal Terrains

Participant: Adrian Kosowski.

In [22], we deal with a more geometric variant of the rendezvous problem. Two mobile agents, modeled as points starting at different locations of an unknown terrain, have to meet. The terrain is a polygon with polygonal holes. We consider two versions of this rendezvous problem: exact RV, when the points representing the agents have to coincide at some time, and ϵ-RV, when these points have to get at distance less than ϵ in the terrain. In any terrain, each agent chooses its trajectory, but the movements of the agent on this trajectory are controlled by an adversary that may, e.g., speed up or slow down the agent. Agents have bounded memory: their computational power is that of finite state machines. Our aim is to compare the feasibility of exact and of ϵ-RV when agents are anonymous vs. when they are labeled. We show classes of polygonal terrains which distinguish all the studied scenarios from the point of view of feasibility of rendezvous. The features which influence the feasibility of rendezvous include symmetries present in the terrains, boundedness of their diameter, and the number of vertices of polygons in the terrains.

6.3.3. Optimal Patrolling of Fragmented Boundaries

Participant: Adrian Kosowski.

Mobile agents in geometric scenarios are also studied in [33], where a set of mobile robots is deployed on a simple curve of finite length, composed of a finite set of vital segments separated by neutral segments. The robots have to patrol the vital segments by perpetually moving on the curve, without exceeding their maximum speed. The quality of patrolling is measured by the idleness, i.e., the longest time period during which any vital point on the curve is not visited by any robot. Given a configuration of vital segments, our goal is to provide algorithms describing the movement of the robots along the curve so as to minimize the idleness. Our main contribution is a proof that the optimal solution to the patrolling problem is attained either by the cyclic strategy, in which all the robots move in one direction around the curve, or by the partition strategy, in which the curve is partitioned into sections which are patrolled separately by individual robots. These two fundamental types of strategies were studied in the past in the robotics community in different theoretical and experimental settings. However, to our knowledge, this is the first theoretical analysis proving optimality in such a general scenario.

6.3.4. *Fast Collaborative Graph Exploration*

Participants: Adrian Kosowski, Dominik Pajak, Przemyslaw Uznanski.

In [35], we study the following scenario of online graph exploration. A team of k agents is initially located at a distinguished vertex r of an undirected graph. At every time step, each agent can traverse an edge of the graph. All vertices have unique identifiers, and upon entering a vertex, an agent obtains the list of identifiers of all its neighbors. We ask how many time steps are required to complete exploration, i.e., to make sure that every vertex has been visited by some agent. We consider two communication models: one in which all agents have global knowledge of the state of the exploration, and one in which agents may only exchange information when simultaneously located at the same vertex. As our main result, we provide the first strategy which performs exploration of a graph with n vertices at a distance of at most D from r in time $O(D)$, using a team of agents of polynomial size $k = Dn^{1+\epsilon} < n^{2+\epsilon}$, for any $\epsilon > 0$. Our strategy works in the local communication model, without knowledge of global parameters such as n or D. We also obtain almost-tight bounds on the asymptotic relation between exploration time and team size, for large k. For any constant $c > 1$, we show that in the global communication model, a team of $k = Dn^c$ agents can always complete exploration in $D(1 + \frac{1}{c-1} + o(1))$ time steps, whereas at least $D(1 + \frac{1}{c} - o(1))$ steps are sometimes required. In the local communication model, $D(1 + \frac{2}{c-1} + o(1))$ steps always suffice to complete exploration, and at least $D(1 + \frac{2}{c} - o(1))$ steps are sometimes required. This shows a clear separation between the global and local communication models.

6.3.5. *A $\widetilde{O}(n2)$ Time-Space Trade-off for Undirected s-t Connectivity*

Participant: Adrian Kosowski.

The work [43] makes use of the Metropolis-type walks due to Nonaka et al. (2010) to provide a faster solution to the S-T-connectivity problem in undirected graphs (USTCON). As the main result of this research, we propose a family of randomized algorithms for USTCON which achieves a time-space product of $S \cdot T = \widetilde{O}(n^2)$ in graphs with n nodes and m edges (where the $\widetilde{O}$-notation disregards poly-logarithmic terms). This improves the previously best trade-off of $\widetilde{O}(nm)$, due to Feige (1995). Our algorithm consists in deploying several short Metropolis-type walks, starting from landmark nodes distributed using the scheme of Broder et al. (1994) on a modified input graph. In particular, we obtain an algorithm running in time $\widetilde{O}(n + m)$ which is, in general, more space-efficient than both BFS and DFS. Finally, we show how to fine-tune the Metropolis-type walk so as to match the performance parameters (e.g., average hitting time) of the unbiased random walk for any graph, while preserving a worst-case bound of $\widetilde{O}(n^2)$ on cover time.

6.3.6. *The multi-agent rotor-router on the ring: a deterministic alternative to parallel random walks*

Participants: Ralf Klasing, Adrian Kosowski, Dominik Pajak.

The *rotor-router mechanism* was introduced as a deterministic alternative to the random walk in undirected graphs. In this model, an agent is initially placed at one of the nodes of the graph. Each node maintains a cyclic ordering of its outgoing arcs, and during successive visits of the agent, propagates it along arcs chosen according to this ordering in round-robin fashion. In [42], we consider the setting in which multiple, indistinguishable agents are deployed in parallel in the nodes of the graph, and move around the graph in synchronous rounds, interacting with a single rotor-router system. We propose new techniques which allow us to perform a theoretical analysis of the multi-agent rotor-router model, and to compare it to the scenario of parallel independent random walks in a graph. Our main results concern the n-node ring, and suggest a strong similarity between the performance characteristics of this deterministic model and random walks.

We show that on the ring the rotor-router with k agents admits a cover time of between $\Theta(n^2/k^2)$ in the best case and $\Theta(n^2/\log k)$ in the worst case, depending on the initial locations of the agents, and that both these bounds are tight. The corresponding expected value of cover time for k random walks, depending on the initial locations of the walkers, is proven to belong to a similar range, namely between $\Theta(n^2/(k^2/\log^2 k))$ and $\Theta(n^2/\log k)$.

Finally, we study the limit behavior of the rotor-router system. We show that, once the rotor-router system has stabilized, all the nodes of the ring are always visited by some agent every $\Theta(n/k)$ steps, regardless of how the system was initialized. This asymptotic bound corresponds to the expected time between successive visits to a node in the case of k random walks. All our results hold up to a polynomially large number of agents ($1 \leq k < n^{1/11}$).

6.3.7. *Efficient Exploration of Anonymous Undirected Graphs*

Participant: Ralf Klasing.

In [41], we consider the problem of exploring an anonymous undirected graph using an oblivious robot. The studied exploration strategies are designed so that the next edge in the robot's walk is chosen using only local information. We present some current developments in the area. In particular, we focus on recent work on *equitable strategies* and on the *multi-agent rotor-router*.

6.3.8. *Gathering radio messages in the path*

Participant: Ralf Klasing.

In [19], we address the problem of gathering information in one node (sink) of a radio network where interference constraints are present: when a node transmits, it produces interference in an area bigger than the area in which its message can actually be received. The network is modeled by a graph; a node is able to transmit one unit of information to the set of vertices at distance at most dt in the graph, but when doing so it generates interferences that do not allow nodes at distance up to di ($di \geq dt$) to listen to other transmissions. We are interested in finding a gathering protocol, that is an ordered sequence of rounds (each round consists of non-interfering simultaneous transmissions) such that $w(u)$ messages are transmitted from any node u to a fixed node called the sink. Our aim is to find a gathering protocol with the minimum number of rounds (called *gathering time*). In [19], we focus on the specific case where the network is a path with the sink at an end vertex of the path and where the traffic is unitary ($w(u) = 1$ for all u); indeed this simple case appears to be already very difficult. We first give a new lower bound and a protocol with a gathering time that differ only by a constant independent of the length of the path. Then we present a method to construct incremental protocols. An incremental protocol for the path on $n + 1$ vertices is obtained from a protocol for n vertices by adding new rounds and new calls to some rounds but without changing the calls of the original rounds. We show that some of these incremental protocols are optimal for many values of dt and di (in particular when dt is prime). We conjecture that this incremental construction always gives optimal protocols. Finally, we derive an approximation algorithm when the sink is placed in an arbitrary vertex in the path.

6.3.9. *Computing Without Communicating: Ring Exploration by Asynchronous Oblivious Robots*

Participant: David Ilcinkas.

In [24], we consider the problem of exploring an anonymous unoriented ring by a team of k identical, oblivious, asynchronous mobile robots that can view the environment but cannot communicate. This weak scenario is standard when the spatial universe in which the robots operate is the two-dimensional plane, but (with one exception) has not been investigated before for networks. Our results imply that, although these weak capabilities of robots render the problem considerably more difficult, ring exploration by a small team of robots is still possible. We first show that, when k and n are not co-prime, the problem is not solvable in general, e.g., if k divides n there are initial placements of the robots for which gathering is impossible. We then prove that the problem is always solvable provided that n and k are co-prime, for $k \geq 17$, by giving an exploration algorithm that always terminates, starting from arbitrary initial configurations. Finally, we consider the minimum number $\rho(n)$ of robots that can explore a ring of size n. As a consequence of our positive result we show that $\rho(n)$ is $O(\log n)$. We additionally prove that $\Omega(\log n)$ robots are necessary for infinitely many n.

6.3.10. Worst-case optimal exploration of terrains with obstacles

Participant: David Ilcinkas.

A mobile robot represented by a point moving in the plane has to explore an unknown flat terrain with impassable obstacles. Both the terrain and the obstacles are modeled as arbitrary polygons. We consider two scenarios: the *unlimited vision*, when the robot situated at a point p of the terrain explores (sees) all points q of the terrain for which the segment pq belongs to the terrain, and the *limited vision*, when we require additionally that the distance between p and q is at most 1. All points of the terrain (except obstacles) have to be explored and the performance of an exploration algorithm, called its complexity, is measured by the length of the trajectory of the robot.

For unlimited vision we show in [21] an exploration algorithm with complexity $O(P + D\sqrt{k})$, where P is the total perimeter of the terrain (including perimeters of obstacles), D is the diameter of the convex hull of the terrain, and k is the number of obstacles. We do not assume knowledge of these parameters. We also prove a matching lower bound showing that the above complexity is optimal, even if the terrain is known to the robot. For limited vision we show exploration algorithms with complexity $O(P + A + \sqrt{Ak})$, where A is the area of the terrain (excluding obstacles). Our algorithms work either for arbitrary terrains (if one of the parameters A or k is known) or for c-fat terrains, where c is any constant (unknown to the robot) and no additional knowledge is assumed. (A terrain $\mathcal{T}$ with obstacles is c-fat if $R/r \leq c$, where R is the radius of the smallest disc containing $\mathcal{T}$ and r is the radius of the largest disc contained in $\mathcal{T}$.) We also prove a matching lower bound $\Omega(P + A + \sqrt{Ak})$ on the complexity of exploration for limited vision, even if the terrain is known to the robot.

6.3.11. Exploration of the T-Interval-Connected Dynamic Graphs: the Case of the Ring

Participants: David Ilcinkas, Ahmed Wade.

In [40], we study the T-interval-connected dynamic graphs from the point of view of the time necessary and sufficient for their exploration by a mobile entity (agent). A dynamic graph (more precisely, an evolving graph) is T-interval-connected ($T \geq 1$) if, for every window of T consecutive time steps, there exists a connected spanning subgraph that is stable (always present) during this period. This property of connection stability over time was introduced by Kuhn, Lynch and Oshman (STOC 2010). We focus on the case when the underlying graph is a ring of size n, and we show that the worst-case time complexity for the exploration problem is $2n - T - \Theta(1)$ time units if the agent knows the dynamics of the graph, and $n + \frac{n}{\max\{1,T-1\}}(\delta - 1) \pm \Theta(\delta)$ time units otherwise, where δ is the maximum time between two successive appearances of an edge.

6.3.12. Time vs. space trade-offs for rendezvous in trees

Participant: Adrian Kosowski.

In [23], we consider the rendezvous problem, in which two identical (anonymous) mobile agents start from arbitrary nodes of an unknown tree and have to meet at some node. Agents move in synchronous rounds: in each round an agent can either stay at the current node or move to one of its neighbors. We consider deterministic algorithms for this rendezvous task. We obtain a tight trade-off between the optimal time of completing rendezvous and the size of memory of the agents. For agents with k memory bits, we show that optimal rendezvous time is $\Theta(n + n^2/k)$ in n-node trees. More precisely, if $k \geq c \log n$, for some constant c, we design agents accomplishing rendezvous in arbitrary trees of size n (unknown to the agents) in time $O(n + n^2/k)$, starting with arbitrary delay. We also show that no pair of agents can accomplish rendezvous in time $o(n + n^2/k)$, even in the class of lines of known length and even with simultaneous start. Finally, we prove that at least logarithmic memory is necessary for rendezvous, even for agents starting simultaneously in a n-node line.

GRAND-LARGE Project-Team

5. New Results

5.1. Automated Code Generation for Lattice Quantum Chromodynamics

Participants: Denis Barthou, Konstantin Petrov, Olivier Brand-Foissac, Olivier Pène, Gilbert Grosdidier, Michael Kruse, Romain Dolbeau, Christine Eisenbeis, Claude Tadonki.

This ongoing work is about a Domain Specific Language which aims to simplify Monte-Carlo simulations and measurements in the domain of Lattice Quantum Chromodynamics. The tool-chain, called Qiral, is used to produce high-performance OpenMP C code from LaTeX sources. We discuss conceptual issues and details of implementation and optimization. The comparison of the performance of the generated code to the well-established simulation software is also made.[33][20][37]

5.2. A Fine-grained Approach for Power Consumption Analysis and Prediction

Participants: Alessandro Ferreira Leite, Claude Tadonki, Christine Eisenbeis, Alba Cristina de Melo.

Power consumption has became a critical concern in modern computing systems for various reasons including financial savings and environmental protection. With battery powered devices, we need to care about the available amount of energy since it is limited. For the case of supercomputers, as they imply a large aggregation of heavy CPU activities, we are exposed to a risk of overheating. As the design of current and future hardware is becoming more and more complex, energy prediction or estimation is as elusive as that of time performance. However, having a good prediction of power consumption is still an important request to the computer science community. Indeed, power consumption might become a common performance and cost metric in the near future. A good methodology for energy prediction could have a great impact on power-aware programming, compilation, or runtime monitoring. In this paper, we try to understand from measurements where and how power is consumed at the level of a computing node. We focus on a set of basic programming instructions, more precisely those related to CPU and memory. We propose an analytical prediction model based on the hypothesis that each basic instruction has an average energy cost that can be estimated on a given architecture through a series of micro-benchmarks. The considered energy cost per operation includes all of the overhead due to context of the loop where it is executed. Using these precalculated values, we derive an linear extrapolation model to predict the energy of a given algorithm expressed by means of atomic instructions. We then use three selected applications to check the accuracy of our prediction method by comparing our estimations with the corresponding measurements obtained using a multimeter. We show a 9.48% energy prediction on sorting.[35]

5.3. Switcheable scheduling

Participants: Lénaïc Bagnères, Cédric Bastoul, Taj Khan.

Parallel applications used to be executed alone until their termination on partitions of supercomputers. The recent shift to multicore architectures for desktop and embedded systems is raising the problem of the coexistence of several parallel programs. Operating systems already take into account the *affinity* mechanism to ensure a thread will run only onto a subset of available processors (e.g., to reuse data remaining in the cache since its previous execution). But this is not enough, as demonstrated by the large performance gaps between executions of a given parallel program on desktop computers running several processes. To support many parallel applications, advances must be made on the system side (scheduling policies, runtimes, memory management...). However, automatic optimization and parallelization can play a significant role by generating programs with dynamic-auto-tuning capabilities to adapt themselves to the complete execution context, including the system load.

Our approach is to design at compile-time programs that can adapt at run-time to the execution context. The originality of our solution is to rely on *switcheable scheduling*, a selected set of program restructuring which allows to swap between program versions at some meeting points without backtracking. A first step selects pertinent versions according to their performance behavior on some execution contexts. The second step builds the auto-adaptive program with the various versions. Then at runtime the program selects the best version by a low overhead sampling and profiling of the versions, ensuring every computation is useful.

This work has been started at Paris-Sud University by Cédric Bastoul before he joined the Inria CAMUS project team during this year. The first results have been presented in 2013 at the HiPEAC System Week and at the Rencontres Françaises de Compilation.

5.4. Solving Navier-Stokes equations on heterogeneous parallel architectures

Participants: Marc Baboulin, Jack Dongarra, Joël Falcou, Yann Fraigneau, Olivier Lemaître, Yushan Wang.

The Navier-Stokes equations describe a large class of fluid flows but are difficult to solve analytically because of their nonlinearity. We implemented a parallel solver for the 3-D Navier-Stokes equations of incompressible unsteady flows with constant coefficients, discretized by the finite difference method. We applied the prediction-projection method which transforms the Navier-Stokes equations into three Helmholtz equations and one Poisson equation. For each Helmholtz system, we applied the Alternating Direction Implicit (ADI) method resulting in three tridiagonal systems. The Poisson equation is solved using partial diagonalization which transforms the Laplacian operator into a tridiagonal one. Our implementation is based on MPI where the computations are performed on each subdomain and information is exchanged on the interfaces, and where the tridiagonal system solutions are accelerated using vectorization techniques. We provided performance results on a current multicore system.[31]

5.5. Optimizing NUMA effects in dense linear algebra software

Participants: Marc Baboulin, Adrien Rémy, Brigitte Rozoy, Masha Sosonkina.

We studied the impact of non-uniform memory accesses (NUMA) on the solution of dense general linear systems using an LU factorization algorithm. In particular we illustrated how an appropriate placement of the threads and memory on a NUMA architecture can improve the performance of the panel factorization and consequently accelerate the global LU factorization. We applied these placement strategies and presented performance results for a hybrid multicore/GPU LU algorithm as it is implemented in the public domain library MAGMA.

HIEPACS Project-Team

6. New Results

6.1. High-performance computing on next generation architectures

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

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

This paper presents an extension of StarPU, a runtime system specifically designed for heterogeneous architectures, that allows multiple parallel codes to run concurrently with minimal interference. Such parallel codes run within *scheduling contexts* that provide confined execution environments which can be used to partition computing resources. Scheduling contexts can be dynamically resized to optimize the allocation of computing resources among concurrently running libraries. We introduce a *hypervisor* that automatically expands or shrinks contexts using feedback from the runtime system (e.g. resource utilization). We demonstrate the relevance of our approach using benchmarks invoking multiple high performance linear algebra kernels simultaneously on top of heterogeneous multicore machines. We show that our mechanism can dramatically improve the overall application run time (-34%), most notably by reducing the average cache miss ratio (-50%).
This work is developed in the framework of Andra Hugo's PhD. These contributions have been presented at the international workshop on Accelerators and Hybrid Exascale Systems [19] in Boston.

6.1.2. A task-based H-matrix solver for acoustic and electromagnetic problems on multicore architectures

$\mathcal{H}$-Matrix is a hierarchical, data-sparse approximate representation of matrices that allows the fast approximate computation of matrix products, LU and LDL^T decompositions, inversion and more. This representation is suitable for the direct solution of large dense linear systems arising from the Boundary Element Method in $O(N \log_2^{\alpha}(N))$ operations. This kind of formulation is widely used in the industry for the numerical simulation of acoustics and electromagnetism scattering by large objetcs. Applications of this approach include aircraft noise reduction and antenna sitting at Airbus Group. The recursive and irregular nature of these $\mathcal{H}$-Matrix algorithms makes an efficient parallel implementation very challenging, especially when relying on a "Bulk Synchronous Parallel" paradigm. We have considered an alternative parallelization for multicore architectures using a task-based approach on top of a runtime system, namely StarPU. We have showed that our method leads to a highly efficient, fully pipelined computation on large real-world industrial test cases provided by Airbus Group.
This research activity has been conduced in the framework of the EADS-ASTRIUM, Inria, Conseil Régional initiative in collaboration with the RUNTIME Inria project, and is part of Benoit Lize's PhD.

6.1.3. A task-based 3D geophysics application

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

This work was presented during a PRACE workshop [28] as well as during a TOTAL scientific event [29].

6.1.4. Resiliency in numerical simulations

For the solution of systems of linear equations, various recovery-restart strategies have been investigated in the framework of Krylov subspace methods to address the situations of core failures. The basic underlying idea is to recover fault entries of the iterate via interpolation from existing values available on neighbor cores. The resulting results are reported in the research report [41] currently submitted to an international journal. In that resilience framework, we have extended the recovey-restart ideas to the solution of linear eigenvalue problems. Contrary to the linear system case, not only the current iterate can be interpolated but also part of the subspace where candidate eigenpairs are searched.
This work is developed in the framework of Mawussi Zounon's PhD funded by the ANR RESCUE. These contributions have been presented at the international workshop Sparse Days [27] in Toulouse. More details and results can be found in report RR-8324 [41]. Notice that theses activities are also part of our contribution to the G8 ESC (Enabling Climate Simulation at extreme scale).

6.2. High performance solvers for large linear algebra problems

6.2.1. Parallel sparse direct solver on runtime systems

The ongoing hardware evolution exhibits an escalation in the number, as well as in the heterogeneity, of the computing resources. The pressure to maintain reasonable levels of performance and portability, forces the application developers to leave the traditional programming paradigms and explore alternative solutions. Algorithms, especially those in critical domains such as linear algebra, need to undergo invasive structural changes and be adapted to new programming paradigms to be in agreement with the latest hardware advances. PaStiX is a parallel sparse direct solver, based on a dynamic scheduler for modern hierarchical architectures. In this paper, we study the replacement of the highly specialized internal scheduler in PaStiX by two generic runtime frameworks: PaRSEC and StarPU. The tasks graph of the factorization step is made available to the two runtimes, providing them with the opportunity to optimize it in order to maximize the algorithm efficiency for a predefined execution environment. A comparative study of the performance of the PaStiX solver with the three schedulers on different execution contexts is performed. The analysis highlights the similarities from a performance point of view between the different execution supports. These results demonstrate that these generic DAG-based runtimes provide a uniform and portable programming interface across heterogeneous environments, and are, therefore, a sustainable solution for hybrid environments.
This work is developed in the framework of Xavier Lacoste's PhD funder by the ANR ANEMOS. These contributions have been presented at the international workshop Sparse Days [37] in Toulouse. More details and results can be found in report RR-8446 [46].

6.2.2. Hybrid parallel implementation of hybrid solvers

In the framework of the hybrid direct/iterative MaPHyS solver, we have designed and implemented an hybrid MPI-thread variant. More precisely, the implementation rely on the multi-threaded MKL library for all the dense linear algebra calculations and the multi-threaded version of PaStiX. Among the technical difficulties, one was to make sure that the two multi-threaded libraries do not interfere with each other. The resulting software prototype is currently experimented to study its new capability to get flexibility and trade-off between the parallel and numerical efficiency. Parallel experiments have been conducted on the Plafrim plateform as well as on a large scale machine located at the USA DOE NERSC, which has a large number of CPU cores per socket.
This work is developed in the framework of the PhD thesis of Stojce Nakov funded by TOTAL. These contributions have been presented at the NVIDIA GPU Technology Conference [25] in San Jose.

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

New hybrid LU-QR algorithms for solving dense linear systems of the form $Ax = b$ have been introduced. Throughout a matrix factorization, these algorithms dynamically alternate LU with local pivoting and QR elimination steps, based upon some robustness criterion. LU elimination steps can be very efficiently parallelized, and are twice as cheap in terms of flops, as QR steps. However, LU steps are not necessarily stable, while QR steps are always stable. The hybrid algorithms execute a QR step when a robustness criterion detects some risk for instability, and they execute an LU step otherwise. Ideally, the choice between LU and QR steps must have a small computational overhead and must provide a satisfactory level of stability with as few QR steps as possible. In this paper, we introduce several robustness criteria and we establish upper bounds on the growth factor of the norm of the updated matrix incurred by each of these criteria. In addition, we describe the implementation of the hybrid algorithms through an extension of the PaRSEC software to allow for dynamic choices during execution. Finally, we analyze both stability and performance results compared to state-of-the-art linear solvers on parallel distributed multicore platforms. These contributions have been presented at the international conference IPDPS [18] in Phoenix.

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

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

1. Implementation of the FMM of multicore machines using StarPU. A new parallel scheduler was developed for this purpose. We implemented a state-of-the-art OpenMP version of the code for benchmarking purposes. It was found that StarPU significantly outperforms OpenMP. Figures show the traces of an execution of the FMM algorithm with our priority scheduler for the cube (volume) and ellipsoid (surface) with 20 million particles on a 4 deca-core Intel Xeon E7-4870 machine.
2. Implementation of the FMM of heterogeneous machines (CPU+GPU) using StarPU. The FMM was also used to demonstrate the flexibility of StarPU for handling different types of processors. In particular we demonstrated in that application that StarPU can automatically select the appropriate version of a computational kernel (CPU or GPU version) and run it on the appropriate processor in order to minimize the overall runtime. Significant speed-up were obtained on heterogeneous platforms compared to multicore only processors.

These contributions have been presented in minnisymposia at the SIAM conference on Comutational Sciences and Engineering [23], [33] in Boston and at NVIDIA GPU Technology Conference [24]. More details and results can be found in report RR-8277 [40], our paper is accepted for publication in the SIAM Journal on Scientific Computing [11].

Concerning dynamics dislocations (DD) kernels, an efficient formulation of the isotropic elastic far-field interactions between dislocations has been developed. This formulation is suitable for any polynomial interpolation based Fast Multipole Method (FMM) and is currently being implemented in OptiDis.

Meanwhile a much lighter and faster interpolation scheme based on a uniform grid (i.e. Lagrange interpolation) and the Fast Fourier Transform (FFT) was implemented into ScalFMM. This last feature was introduced in order to overcome the expensive cost of the Chebyshev FMM in the range of low interpolation orders (up to approx. 10). This should significantly improve the performances of the far-field computation in DD simulations where tensorial kernels are involved but only relatively low interpolation orders are required. This work is developed in the framework of Pierre Blanchard's PhD funded by ENS.

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

6.4.1. Dynamic load balancing for massively parallel coupled codes

As a preliminary step related to the dynamic load balancing of coupled codes, we focus on the problem of dynamic load balancing of a single parallel code, with variable number of processors. Indeed, if the workload

varies drastically during the simulation, the load must be redistributed regularly among the processors. Dynamic load balancing is a well studied subject but most studies are limited to an initially fixed number of processors. Adjusting the number of processors at runtime allows to preserve the parallel code efficiency or to keep running the simulation when the current memory resources are exceeded. We call this problem, *MxN graph repartitioning*. We propose some methods based on graph repartitioning in order to rebalance the load while changing the number of processors. These methods are split in two main steps. Firstly, we study the migration phase and we build a "good" migration matrix minimizing several metrics like the migration volume or the number of exchanged messages. Secondly, we use graph partitioning heuristics to compute a new distribution optimizing the migration according to the previous step results. Besides, we propose a direct k-way partitioning algorithm that allows us to improve our biased partitioning. Finally, an experimental study validates our algorithms against state-of-the-art partitioning tools. Our algorithms are implemented in the LBC2 library and have been integrated in the partitioning tools `Scotch` as a prototype.
This work is developed in the framework of Clément Vuchener's PhD, that will be defended on February 2014. These contributions have been presented at the international conference ParCo [22] in Munchen.

Regarding the problem of dynamic balancing of parallel coupled codes, we start to reuse results on *MxN graph repartitioning*. Given two coupled codes A and B, the key idea is to develop an algorithm of *two-graph co-partitioning*, that partitions two *coupled* graphs G_A and G_B in respectively N_A and N_B with classic objectives (*i.e.*, balancing computational load and minimizing communication cost for each code) and that minimizes the number of messages exchanged between codes in the coupling phase.
This work is developed in the framework of Maria Predari's PhD, that just started in october 2013.

6.4.2. Graph partitioning for hybrid solvers

Nested Dissection has been introduced by A. George and is a very popular heuristic for sparse matrix ordering before numerical factorization. It allows to maximize the number of parallel tasks, while reducing the fill-in and the operation count. The basic standard idea is to build a "small separator" S of the graph associated with the matrix in order to split the remaining vertices in two parts P_0 and P_1 of "almost equal size". The vertices of the separator S are ordered with the largest indices, and then the same method is applied recursively on the two sub-graphs induced by P_0 and P_1. At the end, if k levels of recursion are done, we get 2^k sets of independents vertices separated from each other by $2^k - 1$ separators.
However, if we examine precisely the complexity analysis for the estimation of asymptotic bounds for fill-in or operation count when using Nested Dissection ordering, we can notice that the size of the halo of the separated sub-graphs (set of external vertices belonging to an old separator and previously ordered) plays a crucial role in the asymptotic behavior achieved. In the perfect case, we need halo vertices to be balanced among parts.
Considering now hybrid methods mixing both direct and iterative solvers such as HIPS, MaPHyS, obtaining a domain decomposition leading to a good balancing of both the size of domain interiors and the Scalable numerical schemes for scientific applications size of interfaces is a key point for load balancing and efficiency in a parallel context. This leads to the same issue: balancing the halo vertices to get balanced interfaces.
For this purpose, we revisit the algorithm introduced by Lipton, Rose and Tarjan which performed the recursion of nested dissection in a different manner: at each level, we apply recursively the method to the sub-graphs But, for each sub-graph, we keep track of halo vertices. We have implemented that in the Scotch framework, and have studied its main algorithm to build a separator, called greedy graph growing.
This work is developed in the framework of Astrid Casadei's PhD. These contributions have been presented at the international workshop on Nested Dissection [32] in Waterloo.

6.5. Application Domains

6.5.1. Dislocation dynamics simulations in material physics

This year we have focused on the hybrid parallelization of the OptiDis code. As dislocations move in their grain, they expand, shrink, collide and annihilate, which means that we are facing a extremely dynamic n-body problem. Also, we have introduced an adaptive cache conscious data structure to manage the dislocation mesh. Moreover, two main kernels, plugged in our ScalFMM library, was built to handle the pairwise force interactions and the collisions between dislocations. Finally the code is written using hybrid parallelism based on OpenMP tasks inside on node and MPI to exchange data between nodes. The code can run on both shared and distributed memories. Future works will mainly focus on tuning the code and manage dynamically this tuning to adapt to different kind of simulations and architectures. On the physical side, we have introduced more *split node* cases to simulate irradiated materials. Now we are able to run simulations with tens of thousand of defaults in materials. Typically, our simulation box can hold lot of tiny dislocation loops such as those induced by radiation on materials, so we can observe how Frank-Read sources interact while they cross the field of loop defects.
This work is developed in the framework of Arnaud Etcheverry's PhD funded by the ANR OPTIDIS.

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

The study of the **thermo-acoustic stability of large combustion chambers** requires the solution of a nonlinear eigenvalue problem. The nonlinear problem is linearized using a fixed point iteration procedure. This leads to a sequence of linear eigenproblems which must be solved iteratively in order to obtain one nonlinear eigenpair. Therefore, efficient and robust parallel eigensolvers for the solution of linear problems have been investigated, and strategies to accelerate the solution of the sequence of linear eigenproblems have also been proposed. Among the numerical techniques that have been considered (Krylov-Schur, Implicitly Restarted Arnoldi, Subspace iteration with Chebyshev acceleration) the Jacobi-Davidson method was the best suited to be combined with techniques to recycle spectral information between the nonlinear iterations. The robustness of the parallel numerical techniques were illustrated on large problems with a few millions unknowns solved on a few tens of cores.
These results are part of the outcome of Pablo Salas PhD thesis that has been defended on November 15th.

The **Time-domain Boundary Element Method** (TD-BEM) has not been widely study but represent an interesting alternative to its frequency counterpart. Usually based on inefficient Sparse Matrix Vector-product (SpMV), we investigate other approaches in order to increase the sequential flop-rate. We have implement extremely efficient operator using intrinsic SIMD or even ASM64 instructions. We are using this novel approaches to parallelize both in shared and distributed memory and target execution on hundreds of clusters. All the implementations should be in high quality in the Software Engineering sense since the resulting library is going to be used by industrial applications.
This work is developed in the framework of Bérenger Bramas's PhD and contributes to the EADS-ASTRIUM, Inria, Conseil Régional initiative.

In a preliminary work, a **3D Cartesian SN solver**DOMINO has been designed and implemented using two nested levels of parallelism (multicore+SIMD) on shared memory computation nodes. DOMINO is written in C++, a multi-paradigm programming language that enables the use of powerful and generic parallel programming tools such as Intel TBB and Eigen. These two libraries allow us to combine multi-thread parallelism with vector operations in an efficient and yet portable way. As a result, DOMINO can exploit the full power of modern multi-core processors and is able to tackle very large simulations, that usually require large HPC clusters, using a single computing node. The very high Flops/Watt ratio of DOMINO makes it a very interesting building block for a future many-nodes nuclear simulation tool.
This work is developed in the framework of Salli Moustafa's PhD in collaboration with EDF. These contributions have been presented at the international conference on Supercomputing on Nuclear Applications [21] in Paris.

Concerning the numerical simulation of **the turbulence of plasma particules inside a tokamak**, two software tools, providing a post-mortem analysis, have been designed to manage the memory optimization of GYSELA [20]. The first one is a visualization tool. It plots the memory consumption of the code along an execution. This tool helps the developer to localize where happens the memory peak and to wonder how he can modify the code to decrease it. On the same graphic, the names of the allocated structures are labelled, which gives a significant hint on the modifications to apply. The second tool concerns the prediction of the peak memory. Given an input set of parameters, we can replay the allocations of the code in an offline mode. With this tool, we can deduce accurately the value of the memory peak and where it happens. Thank to this prediction we know which size of mesh is possible under a given architecture.
This work is carried on in the framework of Fabien Rozar's PhD in collaboration with CEA Cadarache.

In the first part of our research work concerning the parallel **aerodynamic code**FLUSEPA, an intermediate version based on the previous one has been developped. By using an hybrid OpenMP/MPI parallelism based on a domain decomposition, we achieved a faster version of the code and the temporal adaptive method used without bodies in relative motion has been tested successfully for real complex 3D-cases using up to 400 cores. Moreover, an asynchronous strategy for computing bodies in relative motion and mesh intersections has been developed and the test of this feature is currently in progress. The next step will be to design a new fully asynchronous code based on a task graph description to be executed on a modern runtine system like StarPU.
This work is carried on in the framework of Jean-Marie Couteyen's PhD in collaboration with Astrium Les Mureaux.

KERDATA Project-Team

6. New Results

6.1. A-Brain and TomusBlobs

6.1.1. Experiments with TomusBlobs at large scale

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Joint genetic and neuro-imaging data analysis may help identifying risk factors in target populations. Performing such studies on a large number of subjects is challenging as genotyping DNA chips can record several hundred thousands values per subject, while the fMRI images may contain 100k–1M volumetric picture elements. Determining statistically significant links between the two sets of data entails a massive amount of computation as one needs not only to compare all pair-wise relations but also to correct for family-wise multiple comparisons. These false positives are controlled by generating permutations of the input data set. The A-Brain initiative is such a data analysis application involving large cohorts of subjects and used to study and understand the variability that exists between individuals. Supposing that such an application could be executed on a single machine, the computation would take years. Cloud infrastructures have the potential to decrease this computation time to days, by parallellizing and scaling out the application. In order to execute this computation in parallel at a large scale, we noticed that the A-Brain application can be easily described as a MapReduce process. The problem was further divided into 28,000 computation tasks, which were executed as map jobs.

The experiment timespan was 14 days and was performed across 4 cloud deployments in 2 different US Azure data centers — North and West. The processing time for a map job is approximatively 2 hours and there are no notable time differences between the map execution time with respect to the geographical location. This is achieved due to the load balancing of the workload, the data locality within the deployments and to the geographical partition. The global result was aggregated using a MapIterativeReduce technique which was composed of 563 reduce jobs. This reduction process eliminates the implicit barrier between mappers and reducers, the reduction process happens in parallel with the map computation. During the period of the experiment the Azure services became temporary inaccessible, due to a failure of a secured certificate. Despite this problem, the framework was capable to handle the failure due to a safety mechanism that we implemented which suspended the computation until all Azure services became available again. Regarding the lost map/reduce enqueued jobs, the monitor mechanism, which supervises the computation progress, was able to restore them. The cost of the experiment was approximatively 210,000 compute hours, where 1 compute hour is equivalent to 1 CPU running for one hour. The monetary cost of the experiment adds up to almost 20,000 $. The total amount combines the cost of the compute resources, for which a value of 0.09 $/h was considered, the persistent Azure storage cost and the outbound traffic from the data centers. As a result of this experiment, we have confirmed that brain activation signals are a heritable feature.

6.1.2. Using dedicated compute nodes for data management on public clouds

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

A large spectrum of scientific applications, some generating data volumes exceeding petabytes, are currently being ported on clouds to build on their inherent elasticity and scalability. One of the critical needs in order to deal with this "data deluge" is an efficient, scalable and reliable storage. However, the storage services proposed by cloud providers suffer from high latencies, trading performance for availability. One alternative is to federate the local virtual disks on the compute nodes into a globally shared storage used for large intermediate or checkpoint data. This collocated storage supports a high throughput but it can be very intrusive and subject to failures that can stop the host node and degrade the application performance.

To deal with these limitations we proposed DataSteward [25], a data management system that provides a higher degree of reliability while remaining non-intrusive through the use of dedicated compute nodes. DataSteward harnesses the storage space of a set of dedicated VMs, selected using a topology-aware clustering algorithm, and has a lifetime dependent on the deployment lifetime. To capitalize on this separation, we introduced a set of scientific data processing services on top of the storage layer, that can overlap with the executing applications. We performed extensive experimentations on hundreds of cores in the Azure cloud: compared to state-of-the-art node selection algorithms, we show up to a 20 % higher throughput, which improves the overall performance of a real life scientific application by up to 45 %.

6.1.3. File transfers for workflows

Participants: Radu Tudoran, Alexandru Costan, Gabriel Antoniu.

Scientific workflows typically communicate data between tasks using files. Currently, on public clouds, this is achieved by using the cloud storage services, which are unable to exploit the workflow semantics and are subject to low throughput and high latencies. To overcome these limitations, we propose in [26] an alternative leveraging data locality through direct file transfers between the compute nodes. We rely on the observation that workflows generate a set of common data access patterns that our solution exploits in conjunction with context information to self-adapt, choose the most adequate transfer protocol and expose the data layout within the virtual machines to the workflow engines. This file management system was integrated within the Microsoft Generic Worker workflow engine and was validated using synthetic benchmarks and a real-life application on the Azure cloud. The results show it can bring significant performance gains: up to 5x file transfer speedup compared to solutions based on standard cloud storage and over 25 % application timespan reduction compared to Hadoop on Azure. This work was done in colaboration with Goetz Brasche and Ramin Rezai Rad from *Microsoft Advance Technology Lab Europe*.

6.2. Optimizing MapReduce Processing

6.2.1. Optimizing MapReduce in virtualized environments

Participant: Shadi Ibrahim.

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

In [17], we developed a two-level scheduling framework that decouples throughput and latency allocation to provide QoS guarantees to VMs while maintaining high disk utilization. The high-level throughput control regulates the pending requests from the VMs with an adaptive credit-rate controller, in order to meet the throughput requirements of different VMs and ensure performance isolation. Meanwhile, the low-level latency control, by the virtue of the batch and delay earliest deadline first mechanism (BD-EDF), re-orders all pending requests from VMs based on their deadlines, and batches them to disk devices taking into account the locality of accesses across VMs.

In [24], we developed a novel disk I/O scheduling framework, named *Pregather*, to improve disk I/O efficiency through exposure and exploitation of the special spatial locality in the virtualized environment (regional and sub-regional spatial locality corresponds to the virtual disk space and applications' access patterns, respectively), thereby improving the performance of disk-intensive applications (e.g., MapReduce applications) without harming the transparency feature of virtualization (without a priori knowledge of the applications' access patterns). The key idea behind Pregather is to implement an intelligent model to predict the access regularity of sub-regional spatial locality for each VM.

We evaluated *Pregather* through extensive experiments that involve multiple simultaneous applications of both synthetic benchmarks and a MapReduce application (i.e., distributed sort) on Xen-based platforms. Our experiments indicate that *Pregather* results in high disk spatial locality, yields a significant improvement in disk throughput, and ends up with improved Hadoop performance. This work was done in collaboration with Hai Jin, Song Wu and Xiao Ling from Huazhong University of Science and Technology (HUST).

6.2.2. Investigating energy efficiency in MapReduce

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

A MapReduce system spans over a multitude of computing nodes that are frequency- and voltage-scalable. Furthermore, many MapReduce applications show significant variation in CPU load during their running time. Thus, there is a significant potential for energy saving by scaling down the CPU frequency. Some power-aware data-layout techniques have been proposed to save power, at the cost of a weaker performance. MapReduce applications range from CPU-Intensive to I/O-Intensive. More importantly, a typical MapReduce application comprises many subtasks, each subtask's workload being either a computation, a disk request or a bandwidth request. As a result, there is a high potential for power reduction by scaling down the CPU when the peak CPU performance is not used.

In this ongoing work, a series of experiments are conducted to explore the implications of *Dynamic Voltage Frequency scaling* (DVFS) settings on power consumption in Hadoop-clusters, by benefitting from the current maturity in DVFS research and of the introduction of governors (e.g., *performance*, *powersave*, *ondemand*, *conservative* and *userspace*). By adapting existing DVFS governors in Hadoop clusters, we observe significant variation in performance and power consumption of the cluster with different applications when applying these governors: the different DVFS settings are only sub-optimal for different MapReduce applications. Furthermore, our results reveal that current CPU governors do not exactly reflect their design goal and may even become ineffective to manage the power consumption. Based on this analysis, we are investigating a new approach to reduce the energy consumption in Hadoop through adaptively tuning the governors and/or the CPU frequencies during the execution of MapReduce jobs.

6.2.3. Hybrid infrastructures

Participants: Alexandru Costan, Ana-Ruxandra Ion, 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 lies 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.

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.). 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 Ana-Ruxandra Ion was dedicated to this topic and showed that for reliable hybrid Map-Reduce processing, one needs to first rely on public/private cloud resources, and then to scale them up using the local, yet volatile, desktop grid resources.

6.2.4. Key partitioning techniques

Participants: Shadi Ibrahim, Gabriel Antoniu.

Data locality is a key feature in MapReduce that is extensively leveraged in data-intensive cloud systems: it avoids network saturation when processing large amounts of data by co-allocating computation and data storage, particularly for the map phase. However, our studies with Hadoop, a widely used MapReduce implementation, demonstrate that the presence of partitioning skew (partitioning skew refers to the case when a variation in either the intermediate keys' frequencies or their distributions or both among different data nodes) causes a huge amount of data transfer during the shuffle phase and leads to significant unfairness on the reduce input among different data nodes. As a result, the applications suffer from severe performance degradation due to the long data transfer during the shuffle phase along with the computation skew, particularly in reduce phase. We addressed these problems by developing a new key/value partitioning called *LEEN*.

In [16], we develop a novel algorithm named *LEEN* for locality-aware and fairness-aware key partitioning in MapReduce. *LEEN* aims at saving the network bandwidth dissipation during the shuffle phase of the MapReduce job along with balancing the reducers' inputs. *LEEN* is conducive to improve the data locality of the MapReduce execution efficiency by the virtue of the asynchronous map and reduce scheme, thereby having more control on the keys distribution in each data node. *LEEN* keeps track of the frequencies of buffered keys hosted by each data node. In doing so, *LEEN* efficiently moves buffered intermediate keys to the destination considering the location of the high frequencies along with fair distribution of reducers' inputs.

To quantify the locality, data distribution and performance of *LEEN*, we conducted a comprehensive performance evaluation study using *LEEN* in Hadoop. Our experimental results demonstrate that *LEEN* interestingly can efficiently achieve higher locality, and balance data distribution after the shuffle phase. This work was done in collaboration with Hai Jin, Song Wu and Lu Lu from Huazhong University of Science and Technology (HUST) and Bingsheng He from Nanyang Technological University (NTU).

6.3. Cloud Storage Trade-Offs: Consistency and Self-Adaptiveness

6.3.1. Cost-aware consistency management in the cloud

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

With the emergence of cloud computing, many organizations have moved their data to the cloud in order to provide scalable, reliable and highly available services. To meet ever growing user needs, these services mainly rely on geographically-distributed data replication to guarantee good performance and high availability. However, with replication, consistency comes into question. Service providers in the cloud have the freedom to select the level of consistency according to the access patterns exhibited by the applications. Most optimizations efforts then concentrate on how to provide adequate trade-offs between consistency guarantees and performance. However, as the monetary cost completely relies on the service providers, in [20] 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.

6.3.2. Analysis of the impact of consistency mangement on energy consumption

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

Energy consumption within datacenters has grown exponentially in recent years. In the era of Big Data, storage and data-intensive applications are one of the main causes of the high power usage. However, few studies have been dedicated to the analysis of the energy consumption of storage systems. Moreover, the impact of consistency management has never been investigated in spite of its high importance. In this work, we address this particular issue. We investigate the energy consumption of application workloads with different consistency models. Thereafter, we leverage the observations about power and the resource usage with every consistency level in order to provide insight into energy-saving practices. In this context, we introduce adaptive configurations of the storage cluster according to the used consistency level. Our experimental evaluations on Cassandra deployed on Grid'5000 demonstrate the existence of the inevitable tradeoff between consistency and energy saving. Moreover, they show how reconfiguring the storage cluster can lead to energy saving, enhanced performance, and better consistency.

6.3.3. Chameleon: customized consistency by means of application behavior modeling

Participants: Houssem-Eddine Chihoub, Gabriel Antoniu.

Multiple Big Data applications are being deployed worldwide to serve a very large number of clients nowadays. These applications vary in their performance and consistency requirements. Understanding such requirements at the storage system level is not possible. The high level semantics of an application is not exposed at the system level. In this context, the consequences of a stale read are not the same for all types of applications.

In [28] , we focus on managing consistency at the application level rather than at the system level. In order to achieve this goal, we propose an offline modeling approach of the application access behavior that considers its high–level consistency semantics. Furthermore, every application state is automatically associated with a consistency policy. At runtime, we introduce the *Chameleon* approach that leverages the application model to provide a customized consistency specific to that application. Experimental evaluations show the high accuracy of our modeling approach exceeding 96% of correct classification of the application states. Moreover, our experiments conducted on Grid'5000 show that *Chameleon* adapts, for every time period, according to the application behavior and requirements while providing best-effort performance.

6.4. Scalable I/O and Virtualization for Exascale Systems

6.4.1. Damaris/Viz

Participants: Matthieu Dorier, Gabriel Antoniu, Lokman Rahmani.

In the context of the Joint Inria/UIUC/ANL Laboratory for Petascale computing (JLCP), we are developing Damaris, which enables efficient I/O, data analysis and visualization at very 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 need for storage or drastically reducing the amount of data generated will be of outmost importance for exascale. In-situ visualization has therefore been proposed to run analysis and visualization tasks closer to the simulation, as it runs.

We investigated the limitations of existing in-situ visualization software and proposed Damaris/Viz, a new version of Damaris that fills the gaps of these software by providing in-situ visualization support to Damaris. The use of Damaris/Viz 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 were conducted on Blue Waters (Cray XK6 at NCSA), Intrepid (BlueGene/P at ANL) and Grid'5000 with representative visualization scenarios for the CM1 [33] atmospheric simulation and the Nek5000 [35] CFD solver. Part of these experiments were carried by NCSA researcher Roberto Sisneros, who gave us important (and very positive) feedbacks on the usability of Damaris at scale (up to 6400 cores on Blue Waters) with real applications. The results of this work were presented as a poster in the PhD forum of IEEE IPDPS 2013 [22], published in a research report [29] and at the IEEE LDAV 2013 conference [23], and a demo of Damaris/Viz was presented at Inria's exhibition booth at the Supercomputing (SC 2013) conference.

This work enlightened the fact that interactive in-situ visualization, although greatly improved by Damaris/Viz, still lakes interactivity. Several meetings were organized with Tom Peterka (ANL) and Roberto Sisneros (NCSA) during the SC conference and during the 10th workshop of the JLPC. We started working on an approach that leverages information theory metrics to automatically find important features of the simulations' data and to reduce the visualization load accordingly.

6.4.2. CALCioM

Participants: Matthieu Dorier, Gabriel Antoniu.

Unmatched computation and storage performance in new HPC systems have led to a plethora of I/O optimizations ranging from application-side collective I/O to network and disk-level request scheduling on the file system side. As we deal with ever larger machines, the interference produced by multiple applications accessing a shared parallel file system in a concurrent manner becomes a major problem. Interference often breaks single-application I/O optimizations, dramatically degrading application I/O performance and, as a result, lowering machine wide efficiency.

Following discussions initiated in 2012 with ANL's Rob Ross and Dries Kimpe, a three month internship of Matthieu Dorier at Argonne National Lab during the summer 2013 led to the design and evaluation of CALCioM (Cross-Application Layer for Coordinated I/O Management), a framework that aims to mitigate I/O interference through the dynamic selection of appropriate scheduling policies. CALCioM allows several applications running on a supercomputer to communicate and coordinate their I/O strategy in order to avoid interfering with one another. Several I/O strategies were evaluated using this framework. Experiments on Argonne's BG/P Surveyor machine and on several clusters of Grid'5000 showed how CALCioM can be used to efficiently and transparently improve the scheduling strategy between several otherwise interfering applications, given specified metrics of machine wide efficiency.

Future work will explore approaches to automatically detect the temporal I/O patterns of simulations in order to further improve the scheduling decisions made by CALCioM.

6.4.3. Scalable metadata management for WAN

Participants: Rohit Saxena, Alexandru Costan, Gabriel Antoniu.

BlobSeer-WAN is a data management service specifically optimized for geographically distributed environments. It is an extension of BlobSeer, a large scale data management service. The metadata is replicated asynchronously for low latency. There is a version manager on each site and vector clocks are used to enable collision detection and resolution under highly concurrent access. It was developed within the framework of Viet-Trung Tran' s PhD thesis, in relation to the FP3C project.

BlobSeer-WAN is used as a storage backend for HGMDS, a multi master metadata server designed for a global distributed file system, developed at University of Tsukuba. Several experiments have been conducted with this setup on the Grid'5000 testbed which have shown scalable metadata performance under geographically distributed environments.

MESCAL Project-Team

6. New Results

6.1. Simulation

6.1.1. Simulation of Parallel Computing Systems

Researchers in the area of distributed computing conduct many of their experiments in simulation. While packet-level simulation is often used to study network protocols, it can be too costly to simulate network communications for large-scale systems and applications. The alternative chosen in SimGrid and a few other simulation frameworks is to simulate the network based on less costly flow-level models. Surprisingly, in the literature, validation of these flow-level models is at best a mere verification for a few simple cases. Consequently, although distributed computing simulators are widely used, their ability to produce scientifically meaningful results is in doubt. In [13] we focus on the validation of state-of-the-art flow-level network models of TCP communication on Wide Area Networks, via comparison to packet-level simulation. While it is straightforward to show cases in which previously proposed models lead to good results, instead we systematically seek cases that lead to invalid results. Careful analysis of these cases reveal fundamental flaws and also suggest improvements. One contribution of this work is that these improvements lead to a new model that, while far from being perfect, improves upon all previously proposed models. A more important contribution, perhaps, is provided by the pitfalls and unexpected behaviors encountered in this work, leading to a number of enlightening lessons. In particular, this work shows that model validation cannot be achieved solely by exhibiting (possibly many) ”good cases.” Confidence in the quality of a model can only be strengthened through an invalidation approach that attempts to prove the model wrong.

The previous results assume steady-state and provide thus a reasonable model when message size is very large. Although, such assumptions may be reasonable when studying grid applications, when simulating HPC applications message sizes are often much smaller and phenomenon like slow-start or how communications and computations overlap have to be accurately modeled. Simulation and modeling for performance prediction and profiling is yet essential for developing and maintaining HPC code that is expected to scale for next-generation exascale systems. In [15], [34] we describe an implementation of a flow-based hybrid network model that accounts for factors such as network topology and contention, which are commonly ignored by the LogP models. Although, this may seem like a strange choice, we focus on large-scale, Ethernet-connected systems, as these currently compose 37.8% of the TOP500 index, and this share is expected to increase as higher-speed 10 and 100GbE become more available. Furthermore, the European Mont-Blanc project to study exascale computing by developing prototype systems with low-power embedded devices will also use Ethernet-based interconnect [28]. Our model is implemented within SMPI, an open-source MPI implementation that connects real applications to the SimGrid simulation framework. SMPI provides implementations of collective communications based on current versions of both OpenMPI and MPICH. SMPI and SimGrid also provide methods for easing the simulation of large-scale systems, including shadow execution, memory folding, and support for both online and offline (i.e., post-mortem) simulation. We validate our proposed model by comparing traces produced by SMPI with those from real world experiments, as well as with those obtained using other established network models. Our study shows that SMPI has a consistently better predictive power than classical LogP-based models for a wide range of scenarios including both established HPC benchmarks and real applications.

6.1.2. Perfect Simulation

Perfect simulation 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. In [7], we consider Jackson queueing networks (JQN) with finite buffer constraints and analyze the efficiency of sampling from their stationary distribution. In the context of exact sampling, the monotonicity structure of JQNs ensures

that such efficiency is of the order of the coupling time (or meeting time) of two extremal sample paths. In the context of approximate sampling, it is given by the mixing time. Under a condition on the drift of the stochastic process underlying a JQN, which we call *hyper-stability*, in our main result we show that the coupling time is polynomial in both the number of queues and buffer sizes. Then, we use this result to show that the mixing time of JQNs behaves similarly up to a given precision threshold. Our proof relies on a recursive formula relating the coupling times of trajectories that start from network states having 'distance one', and it can be used to analyze the coupling and mixing times of other Markovian networks, provided that they are monotone. An illustrative example is shown in the context of JQNs with blocking mechanisms.

In [35], we extend the technique to handle situations with infinite space state. We consider open JQN with losses with mixed finite and infinite queues and analyze the efficiency of sampling from their exact stationary distribution. Although the underlying Markov chain may have an infinite state space, we show that perfect sampling is possible. The main idea is to use a JQN with infinite buffers (that has a product form stationary distribution) to bound the number of initial conditions to be considered in the coupling from the past scheme. We also provide bounds on the sampling time of this new perfect sampling algorithm for acyclic or hyperstable networks. These bounds show that the new algorithm is considerably more efficient than existing perfect samplers even in the case where all queues are finite. We illustrate this efficiency through numerical experiments. We also extend our approach to non-monotone networks such as queueing networks with negative customers.

6.2. Interactive Analysis and Visualization of Large Distributed Systems

6.2.1. Interactive Visualization

High performance applications are composed of many processes that are executed in large-scale systems with possibly millions of computing units. A possible way to conduct a performance analysis of such applications is to register in trace files the behavior of all processes belonging to the same application. The large number of processes and the very detailed behavior that we can record about them lead to a trace size explosion both in space and time dimensions. The performance visualization of such data is very challenging because of the quantities involved and the limited screen space available to draw them all. 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.

In [33], we detail data aggregation techniques that are fully configurable by the user to control the level of details in both space and time dimensions. We also present two visualization techniques that take advantage of the aggregated data to scale. These features are part of the Viva and Triva open-source tools and framework.

The performance of parallel and distributed applications is also highly dependent on the characteristics of the execution environment. In such environments, the network topology and characteristics directly impact data locality and movements as well as contention, which are key phenomena to understand the behavior of such applications and possibly improve it. Unfortunately few visualizations available to the analyst are capable of accounting for such phenomena. In [26], we propose 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. We claim that such kind of visualization enables to explore and understand non trivial behavior that are impossible to grasp with classical visualization techniques. We also claim that the combination of multi-scale aggregation and dynamic graph layout allows our visualization technique to scale seamlessly to large distributed systems. We support these claims through a detailed analysis of a high performance computing scenario and of a grid computing scenario.

6.2.2. Entropy Based Analysis

Although the previous approaches already improve upon state of the art and are useful on current scenarios, it is clear that at very large scale they would probably not be as effective, which led us to change perspective and to investigate how entropy can help building tractable macroscopic descriptions. Indeed, data aggregation can provide such abstractions by partitioning the systems dimensions into aggregated pieces of information. This process leads to information losses, so the partitions should be chosen with the greatest caution, but

in an acceptable computational time. While the number of possible partitions grows exponentially with the size of the system, we propose in [25] an algorithm that exploits exogenous constraints regarding the system semantics to find best partitions in a linear or polynomial time. We detail two constrained sets of partitions that are respectively applied to temporal and spatial aggregation of an agent-based model of international relations. The algorithm succeeds in providing meaningful high-level abstractions for the system analysis.

Our approach is able to evaluate geographical abstractions used by the domain experts in order to provide efficient and meaningful macroscopic descriptions of the world global state [23]. We also successfully applied this technique to identify international media events by spatially and temporally aggregating RSS Flows of Newspapers [22], in particular with the case of the Syrian civil war between May 2011 and December 2012 [31], [21].

We also applied this technique to the analysis of large distributed systems and combined it with the treemap visualization technique [40], [14]. These features have been integrated in the Viva and Triva open-source tools and framework.

6.3. Trace Management and Analysis

6.3.1. Embedded Systems

The growing complexity of embedded system hardware and software makes their behavior analysis a challenging task. In this context, tracing provides relevant information about the system execution and appears to be a promising solution. However, trace management and analysis are hindered by several issues like the diversity of trace formats, the incompatibility of trace analysis methods, the problem of trace size and its storage as well as by the lack of visualization scalability. In [42], [27], [41], we present FrameSoC, a new trace management infrastructure that solves all the above issues together. It provides generic solutions for trace storage and defines interfaces and plugin mechanisms for integrating diverse analysis tools. We illustrate the benefit of FrameSoC with a case study of a visualization module that enables representation scalability of large traces by using an aggregation algorithm. Temporal aggregation techniques based on entropy are also currently integrated to the FrameSoC framework.

6.3.2. Jobs Resource Utilization

In HPC community the System Utilization metric enables to determine if the resources of the cluster are efficiently used by the batch scheduler. This metric considers that all the allocated resources (memory, disk, processors, etc) are full-time utilized. To optimize the system performance, we have to consider the effective physical consumption by jobs regarding the resource allocations. This information gives an insight into whether the cluster resources are efficiently used by the jobs. In [20], [30], we propose an analysis of production clusters based on the jobs resource utilization. The principle is to collect simultaneously traces from the job scheduler (provided by logs) and jobs resource consumption. The latter has been realized by developing a job monitoring tool, whose impact on the system has been measured as lightweight (0.35% speed-down). The key point is to statistically analyze both traces to detect and explain underutilization of the resources. This could enable to detect abnormal behavior, bottlenecks in the cluster leading to a poor scalability, and justifying optimizations such as gang scheduling or best effort scheduling. This method has been applied to two medium sized production clusters on a period of eight months.

6.4. Reconstructing the Software Environment of an Experiment

In the scientific experimentation process, an experiment result needs to be analyzed and compared with several others, potentially obtained in different conditions. Thus, the experimenter needs to be able to redo the experiment. Several tools are dedicated to the control of the experiment input parameters and the experiment replay. In parallel concurrent and distributed systems, experiment conditions are not only restricted to the input parameters, but also to the software environment in which the experiment was carried out. It is therefore essential to be able to reconstruct this type of environment. The task can quickly become complex for experimenters, particularly on research platforms dedicated to scientific experimentation, where both hardware and software are in constant rapid evolution. In [19] we discuss the concept of the reconstructability of software environments and propose a tool, Kameleon, for dealing with this problem.

6.5. Performance Evaluation

6.5.1. Computing the Throughput of Probabilistic and Replicated Streaming Applications

In [8], we investigate how to compute the throughput of probabilistic and replicated streaming applications. We are given (i) a streaming application whose dependence graph is a linear chain; (ii) a one-to-many mapping of the application onto a fully heterogeneous target platform, where a processor is assigned at most one application stage, but where a stage can be replicated onto a set of processors; and (iii) a set of random variables modeling the computation and communication times in the mapping. We show how to compute the throughput of the application, i.e., the rate at which data sets can be processed, under two execution models, the Strict model where the actions of each processor are sequentialized, and the Overlap model where a processor can compute and communicate in parallel. The problem is easy when application stages are not replicated, i.e., assigned to a single processor: in that case the throughput is dictated by the critical hardware resource. However, when stages are replicated, i.e., assigned to several processors, the problem becomes surprisingly complicated: even in the deterministic case, the optimal throughput may be lower than the smallest internal resource throughput. The first contribution of the paper is to provide a general method to compute the throughput when mapping parameters are constant or follow I.I.D. exponential laws. The second contribution is to provide bounds for the throughput when stage parameters (computation and communication times) form associated random sequences, and are N.B.U.E. (New Better than Used in Expectation) variables: the throughput is bounded from below by the exponential case and bounded from above by the deterministic case. An extensive set of simulation allows us to assess the quality of the model, and to observe the actual behavior of several distributions.

6.5.2. Optimization of Cloud Task Processing with Checkpoint-Restart Mechanism

In [17], we explain how to optimize fault-tolerance techniques based on a checkpointing/restart mechanism, in the context of cloud computing. Our contribution is three-fold. (1) We derive a fresh formula to compute the optimal number of checkpoints for cloud jobs with varied distributions of failure events. Our analysis is not only generic with no assumption on failure probability distribution, but also attractively simple to apply in practice. (2) We design an adaptive algorithm to optimize the impact of checkpointing regarding various costs like checkpointing/restart overhead. (3) We evaluate our optimized solution in a real cluster environment with hundreds of virtual machines and Berkeley Lab Checkpoint/Restart tool. Task failure events are emulated via a production trace produced on a large-scale Google data center. Experiments confirm that our solution is fairly suitable for Google systems. Our optimized formula outperforms Young's formula by 3-10 percent, reducing wallclock lengths by 50-100 seconds per job on average.

6.6. Game Theory and Applications

6.6.1. Fair Scheduling in Large Distributed Computing Sytems

Fairly sharing resources of a distributed computing system between users is a critical issue that we have investigated in two ways.

Our first proposal specifically addresses the question of designing a distributed sharing mechanism. A possible answer resorts to Lagrangian optimization and distributed gradient descent. Under certain conditions, the resource sharing problem can be formulated as a global optimization problem, which can be solved by a distributed self-stabilizing demand and response algorithm. In the last decade, this technique has been applied to design network protocols (variants of TCP, multi-path network protocols, wireless network protocols) and even distributed algorithms for smart grids. In [9], we explain how to use this technique for scheduling Bag-of-Tasks (BoT) applications on a Grid since until now, only simple mechanisms have been used to ensure a fair sharing of resources amongst these applications. Although the resulting algorithm is in essence very similar to previously proposed algorithms in the context of flow control in multi-path networks, we show using carefully designed experiments and a thorough statistical analysis that the grid context is surprisingly more difficult than the multi-path network context. Interestingly, we can show that, in practice, the convergence of the algorithm is hindered by the heterogeneity of application characteristics, which is completely overlooked

in related theoretical work. Our careful investigation provides enough insights to understand the true difficulty of this approach and to propose a set of non-trivial adaptations that enable convergence in the grid context. The effectiveness of our proposal is proven through an extensive set of complex and realistic simulations.

Our second proposal is centralized but more fine grain as it does drop the steady-state hypothesis and considers sequences of campaigns. Campaign Scheduling is characterized by multiple job submissions issued from multiple users over time. The work in [18] presents a new fair scheduling algorithm called OStrich whose principle is to maintain a virtual time-sharing schedule in which the same amount of processors is assigned to each user. The completion times in the virtual schedule determine the execution order on the physical processors. Then, campaigns are interleaved in a fair way by OStrich. For independent sequential jobs, we show that OStrich guarantees the stretch of a campaign to be proportional to campaign's size and the total number of users. The theoretical performance of our solution is assessed by simulating OStrich compared to the classical FCFS algorithm, issued from synthetic workload traces generated by two different user profiles. This is done to demonstrate how OStrich benefits both types of users, in contrast to FCFS.

6.6.2. Fundamentals of Continuous Games

We have made the following contributions:

1. Continuous-time game dynamics are typically first order systems where payoffs determine the growth rate of the players' strategy shares. In [12], we investigate what happens beyond first order by viewing payoffs as higher order forces of change, specifying e.g., the acceleration of the players' evolution instead of its velocity (a viewpoint which emerges naturally when it comes to aggregating empirical data of past instances of play). To that end, we derive a wide class of higher order game dynamics, generalizing first order imitative dynamics, and, in particular, the replicator dynamics. We show that strictly dominated strategies become extinct in n-th order payoff-monotonic dynamics n orders as fast as in the corresponding first order dynamics; furthermore, in stark contrast to first order, weakly dominated strategies also become extinct for $n \geq 2$. All in all, higher order payoff-monotonic dynamics lead to the elimination of weakly dominated strategies, followed by the iterated deletion of strictly dominated strategies, thus providing a dynamic justification of the well-known epistemic rationalizability process of Dekel and Fudenberg. Finally, we also establish a higher order analogue of the folk theorem of evolutionary game theory, and we show that convergence to strict equilibria in n-th order dynamics is n orders as fast as in first order.
2. In [37] we introduce a new class of game dynamics made of a pay-off replicator-like term modulated by an entropy barrier which keeps players away from the boundary of the strategy space. We show that these *entropy-driven* dynamics are equivalent to players computing a score as their on-going exponentially discounted cumulative payoff and then using a quantal choice model on the scores to pick an action. This dual perspective on *entropy-driven* dynamics helps us to extend the folk theorem on convergence to quantal response equilibria to this case, for potential games. It also provides the main ingredients to design a discrete time effective learning algorithm that is fully distributed and only requires partial information to converge to QRE. This convergence is resilient to stochastic perturbations and observation errors and does not require any synchronization between the players.

6.6.3. Application to Wireless Networks

We have made the following contributions:

1. Starting from an entropy-driven reinforcement learning scheme for multi-agent environments, we develop in [36] a distributed algorithm for robust spectrum management in Gaussian multiple-input, multiple-output (MIMO) uplink channels. In continuous time, our approach to optimizing the transmitters' signal distribution relies on the method of matrix exponential learning, adjusted by an entropy-driven barrier term which generates a distributed, convergent algorithm in discrete time. As opposed to traditional water-filling methods, the algorithm's convergence speed can be controlled by tuning the users' learning rate; accordingly, entropy-driven learning algorithms in MIMO systems converge arbitrarily close to the optimum signal covariance profile within a few iterations (even for large numbers of users and/or antennas per user), and this convergence remains robust even in the

presence of imperfect (or delayed) measurements and asynchronous user updates.

2. Consider a wireless network of transmitter-receiver pairs where the transmitters adjust their powers to maintain a target SINR level in the presence of interference. In [46], we analyze the optimal power vector that achieves this target in large, random networks obtained by "erasing" a finite fraction of nodes from a regular lattice of transmitter-receiver pairs. We show that this problem is equivalent to the so-called Anderson model of electron motion in dirty metals which has been used extensively in the analysis of diffusion in random environments. A standard approximation to this model is the so-called coherent potential approximation (CPA) method which we apply to evaluate the first and second order intra-sample statistics of the optimal power vector in one- and two-dimensional systems. This approach is equivalent to traditional techniques from random matrix theory and free probability, but while generally accurate (and in agreement with numerical simulations), it fails to fully describe the system: in particular, results obtained in this way fail to predict when power control becomes infeasible. In this regard, we find that the infinite system is always unstable beyond a certain value of the target SINR, but any finite system only has a small probability of becoming unstable. This instability probability is proportional to the tails of the eigenvalue distribution of the system which are calculated to exponential accuracy using methodologies developed within the Anderson model and its ties with random walks in random media. Finally, using these techniques, we also calculate the tails of the system's power distribution under power control and the rate of convergence of the Foschini-Miljanic power control algorithm in the presence of random erasures.

MOAIS Project-Team

6. New Results

6.1. Distributed Art Performance

Moais collaborated with partners fromI2cat, Barcelona, Psnc, Poznan and Grenoble-INP to setup a live distributed art performance for the ICT 2013 conference at Vilnius. This distributed performance gathered musicians located a Poznan, Barcelona and Vilnius, as well as a dancer modeled in 3D on the Grimage platform at Inria Grenoble. Though physically present in different cities these artists performed together for a numerical dance and music performance numerically assembled and transmitted in real-time at Vilnius. This joint effort relies on the FlowVR framework from Moais and the UltraGrid software from CESNET. This event received a significant attention from the medias (In France: FR3 and Tele-grenoble, France inter, etc.). A video is available at http://cyan1.grenet.fr/podcastmedia/Visionair/ICT2013_promo.m4v.

6.2. VTK Parallelization Framework

Moais developed a framework for the parallelization of scientific visualization algorithms based on on-demand task extraction and work stealing techniques. This work is developed for the VTK software and supports the OpenMP, Intel TBB and Kaapi runtime environments. Mathias Ettinger visited the Kitware company, NY, for two months to prepare the integration of this work in the next release of VTK. This work is performed in collaboration with the EDF company.

6.3. Parallel Sorting Algorithm

We developped a novel adaptive sorting algorithm, called PaVO, relying on a Packed Mermory Array data structure. Maintaining gaps in the array of elements enable to reduce the span of modifications needed when reordering elements. This is particularly relevant in a parallel context to reduce the data dependencies. Performance results on a NUMA architecute show that PaVo outperfoms standard parallel sorting algorithms even for a large amount of disorder.

6.4. High bandwidth IPSec gateways and ICMP

Internet Control Message Protocol (ICMP)i is essential for performance aspects in particular for Path Maximum Transmission Unit discovery but is also known to be a cause of attacks. In collaboration with Planet, we demonstrate, through a real exploit on a testbed, that an external attacker having eavesdropping and traffic injection capabilities in the black untrusted network, without any access to clear-text (thesis of Ludovic Jacquin). This impacts out current research on trusted outsourced computations.

6.5. Efficient Parallel multi-GPUs execution

We developped a novel scheduling algorithm in Kaapi to perform multi-GPUs execution of task' based program [19]. Performance results on Cholesky factorization on up to 8-GPUs shows that Kaapi outperfoms similar runtime systems and even hand code parallel version.

6.6. Porting Kaapi for Native Mode on Intel Xeon Phi

Kaapi was ported natively on Intel Xeon Phi co-processor. Specific memory hierarchy was managed transparently to the applicaiton by the development of specific hierarchical work stealing scheduler. Experimentations on dense linear algebra kernels (Cholesky, LU and QR factorization) shows a very promising gain compared to the standard parallel implementation available in the Intel MKL [16]. Extension of theses resultats are under publication process.

6.7. Adaptive loop scheduling in GCC OpenMP runtime library

We port an adaptive loop scheduler from Kaapi into the OpenMP runtime library of GCC called libGOMP [12]. The loop scheduler is consencious of the bloc data mapping to improve locality of computation.

6.8. Kaapi in EPX standard distribution

Kaapi software developped by the MOAIS team was included in the standard EPX distribution. EPX has won the 2013 Grand Prix SFEN (http://www-epx.cea.fr).

ROMA Team

6. New Results

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

In this work [37], we investigate 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. Here, we extend 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 study the computational complexity of this problem and provide an inapproximability result even for unit weight trees. Several heuristics are proposed, each with a different optimization focus, and they are analyzed in an extensive experimental evaluation using realistic trees.

6.2. Model and complexity results for tree traversals on hybrid platforms

In this work [35], we study the complexity of traversing tree-shaped workflows whose tasks require large I/O files. We target a heterogeneous architecture with two resources of different types, where each resource has its own memory, such as a multicore node equipped with a dedicated accelerator (FPGA or GPU). Tasks in the workflow are tagged with the type of resource needed for their processing. Besides, a task can be processed on a given resource only if all its input files and output files can be stored in the corresponding memory. At a given execution step, the amount of data stored in each memory strongly depends upon the ordering in which the tasks are executed, and upon when communications between both memories are scheduled. The objective is to determine an efficient traversal that minimizes the maximum amount of memory of each type needed to traverse the whole tree. In this work, we establish the complexity of this two-memory scheduling problem, provide inapproximability results, and show how to determine the optimal depth-first traversal. Altogether, these results lay the foundations for memory-aware scheduling algorithms on heterogeneous platforms.

6.3. On the combination of silent error detection and checkpointing

In this work [19], we revisit traditional checkpointing and rollback recovery strategies, with a focus on silent data corruption errors. Contrarily to fail-stop failures, such latent errors cannot be detected immediately, and a mechanism to detect them must be provided. We consider two models: (i) errors are detected after some delays following a probability distribution (typically, an Exponential distribution); (ii) errors are detected through some verification mechanism. In both cases, we compute the optimal period in order to minimize the waste, i.e., the fraction of time where nodes do not perform useful computations. In practice, only a fixed number of checkpoints can be kept in memory, and the first model may lead to an irrecoverable failure. In this case, we compute the minimum period required for an acceptable risk. For the second model, there is no risk of irrecoverable failure, owing to the verification mechanism, but the corresponding overhead is included in the waste. Finally, both models are instantiated using realistic scenarios and application/architecture parameters.

6.4. Checkpointing algorithms and fault prediction

In this series of work [22], [49], we deal with the impact of fault prediction techniques on checkpointing strategies, when the fault-prediction system provides either prediction windows or exact predictions. We extend the classical first-order analysis of Young and Daly in the presence of a fault prediction system, characterized by its recall and its precision. In this framework, we provide optimal algorithms to decide whether and when to take predictions into account, and we derive the optimal value of the checkpointing period. These results allow us to analytically assess the key parameters that impact the performance of fault predictors at very large scale.

6.5. Mapping applications on volatile resources

In this series of work [12], [27], [28], we study the execution of iterative applications on volatile processors such as those found on desktop grids. We envision two models, one where all tasks are assumed to be independent, and another where all tasks are tightly coupled and keep exchanging information throughout the iteration. These two models cover the two extreme points of the parallelization spectrum. We develop master-worker scheduling schemes that attempt to achieve good trade-offs between worker speed and worker availability. Any iteration entails the execution of a fixed number of independent tasks or of tightly-coupled tasks. A key feature of our approach is that we consider a communication model where the bandwidth capacity of the master for sending application data to workers is limited. This limitation makes the scheduling problem more difficult both in a theoretical sense and in a practical sense. Furthermore, we consider that a processor can be in one of three states: available, down, or temporarily preempted by its owner. This preempted state also complicates the scheduling problem. In practical settings, e.g., desktop grids, master bandwidth is limited and processors are temporarily reclaimed. Consequently, addressing the aforementioned difficulties is necessary for successfully deploying master-worker applications on volatile platforms. Our first contribution is to determine the complexity of the scheduling problems in their offline versions, i.e., when processor availability behaviors are known in advance. Even with this knowledge, the problems are NP-hard. Our second contribution is an evaluation of the expectation of the time needed by a worker to complete a set of tasks. We obtain a close formula for independent tasks and an analytical approximation for tightly-coupled tasks. Those evaluations rely on a Markovian assumption for the temporal availability of processors, and are at the heart of some heuristics that aim at favoring “reliable” processors in a sensible manner. Our third contribution is a set of heuristics for both models, which we evaluate in simulation. Our results provide guidance to selecting the best strategy as a function of processor state availability versus average task duration.

6.6. Using group replication for resilience on exascale systems

High performance computing applications must be resilient to faults. The traditional fault-tolerance solution is checkpoint-recovery, by which application state is saved to and recovered from secondary storage throughout execution. It has been shown that, even when using an optimal checkpointing strategy, the checkpointing overhead precludes high parallel efficiency at large scale. Additional fault-tolerance mechanisms must thus be used. Such a mechanism is replication, i.e., multiple processors performing the same computation so that a processor failure does not necessarily imply an application failure. In spite of resource waste, replication can lead to higher parallel efficiency when compared to using only checkpoint-recovery at large scale. In this work [11], we propose to execute and checkpoint multiple application instances concurrently, an approach we term group replication. For Exponential failures we give an upper bound on the expected application execution time. This bound corresponds to a particular checkpointing period that we derive. For general failures, we propose a dynamic programming algorithm to determine non-periodic checkpoint dates as well as an empirical periodic checkpointing solution whose period is found via a numerical search. Using simulation we evaluate our proposed approaches, including comparison to the non-replication case, for both Exponential and Weibull failure distributions. Our broad finding is that group replication is useful in a range of realistic application and checkpointing overhead scenarios for future exascale platforms.

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

In this work [10], we present 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 identify a set of crucial parameters, instantiate them and compare the expected efficiency of the fault tolerant protocols, for a given application/platform pair. We then propose 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 outline comparative behaviors of checkpoint strategies at very large scale, thereby providing insight that is hardly accessible to direct experimentation.

6.8. Revisiting the double checkpointing algorithm

In this work [33], we study fast checkpointing algorithms which require distributed access to stable storage. This work revisits the approach base upon double checkpointing, and compares the blocking algorithm of Zheng, Shi, and Kalé, with the non-blocking algorithm of Ni, Meneses, and Kalé in terms of both performance and risk. We also extend the model that they have proposed to assess the impact of the overhead associated to non-blocking communications. We then provide a new peer-to-peer checkpointing algorithm, called the triple checkpointing algorithm, that can work at constant memory, and achieves both higher efficiency and better risk handling than the double checkpointing algorithm. We provide performance and risk models for all the evaluated protocols, and compare them through comprehensive simulations.

6.9. Multi-criteria checkpointing strategies: Optimizing response-time versus resource utilization

Failures are increasingly threatening the efficiency of HPC systems, and current projections of Exascale platforms indicate that rollback recovery, the most convenient method for providing fault tolerance to general-purpose applications, reaches its own limits at such scales. One of the reasons explaining this unnerving situation comes from the focus that has been given to per-application completion time, rather than to platform efficiency. In this work [26], we discuss the case of uncoordinated rollback recovery where the idle time spent waiting recovering processors is used to progress a different, independent application from the system batch queue. We then propose an extended model of uncoordinated checkpointing that can discriminate between idle time and wasted computation. We instantiate this model in a simulator to demonstrate that, with this strategy, uncoordinated checkpointing per application completion time is unchanged, while it delivers near-perfect platform efficiency.

6.10. Optimal checkpointing period: Time vs. energy

In this work [18], we deal with parallel scientific applications using non-blocking and periodic coordinated checkpointing to enforce resilience. We provide a model and detailed formulas for total execution time and consumed energy. We characterize the optimal period for both objectives, and we assess the range of time/energy trade-offs to be made by instantiating the model with a set of realistic scenarios for Exascale systems. We give a particular emphasis to I/O transfers, because the relative cost of communication is expected to dramatically increase, both in terms of latency and consumed energy, for future Exascale platforms.

6.11. Energy-aware checkpointing of divisible tasks with soft or hard deadlines

In this work [20], we aim at minimizing the energy consumption when executing a divisible workload under a bound on the total execution time, while resilience is provided through checkpointing. We discuss several variants of this multi-criteria problem. Given the workload, we need to decide how many chunks to use, what are the sizes of these chunks, and at which speed each chunk is executed. Furthermore, since a failure may occur during the execution of a chunk, we also need to decide at which speed a chunk should be re-executed in

the event of a failure. The goal is to minimize the expectation of the total energy consumption, while enforcing a deadline on the execution time, that should be met either in expectation (soft deadline), or in the worst case (hard deadline). For each problem instance, we propose either an exact solution, or a function that can be optimized numerically. The different models are then compared through an extensive set of experiments.

6.12. Assessing the performance of energy-aware mappings

In this work [8], we aim at mapping streaming applications that can be modeled by a series-parallel graph onto a 2-dimensional tiled chip multiprocessor (CMP) architecture. The objective of the mapping is to minimize the energy consumption, using dynamic voltage and frequency scaling (DVFS) techniques, while maintaining a given level of performance, reflected by the rate of processing the data streams. This mapping problem turns out to be NP-hard, and several heuristics are proposed. We assess their performance through comprehensive simulations using the StreamIt workflow suite and randomly generated series-parallel graphs, and various CMP grid sizes.

6.13. Computing the throughput of probabilistic and replicated streaming applications

In this work [7], we investigate how to compute the throughput of probabilistic and replicated streaming applications. We are given (i) a streaming application whose dependence graph is a linear chain; (ii) a one-to-many mapping of the application onto a fully heterogeneous target platform, where a processor is assigned at most one application stage, but where a stage can be replicated onto a set of processors; and (iii) a set of random variables modeling the computation and communication times in the mapping. We show how to compute the throughput of the application, i.e., the rate at which data sets can be processed, under two execution models, the Strict model where the actions of each processor are sequentialized, and the Overlap model where a processor can compute and communicate in parallel. The problem is easy when application stages are not replicated, i.e., assigned to a single processor: in that case the throughput is dictated by the critical hardware resource. However, when stages are replicated, i.e., assigned to several processors, the problem becomes surprisingly complicated: even in the deterministic case, the optimal throughput may be lower than the smallest internal resource throughput. The first contribution of this work is to provide a general method to compute the throughput when mapping parameters are constant or follow I.I.D. exponential laws. The second contribution is to provide bounds for the throughput when stage parameters (computation and communication times) form associated random sequences, and are N.B.U.E. (New Better than Used in Expectation) variables: the throughput is bounded from below by the exponential case and bounded from above by the deterministic case. An extensive set of simulation allows us to assess the quality of the model, and to observe the actual behavior of several distributions.

6.14. Reliability and performance optimization of pipelined real-time systems

In this work [6], we consider pipelined real-time systems that consist of a chain of tasks executing on a distributed platform. The processing of the tasks is pipelined: each processor executes only one interval of consecutive tasks. We are interested in minimizing both the input-output latency and the period of application mapping. For dependability reasons, we are also interested in maximizing the reliability of the system. We therefore assign several processors to each interval of tasks, so as to increase the reliability of the system. Both processors and communication links are unreliable and subject to transient failures. We assume that the arrival of the failures follows a constant parameter Poisson law, and that the failures are statistically independent events. We study several variants of this multiprocessor mapping problem, with several hypotheses on the target platform (homogeneous/heterogeneous speeds and/or failure rates). We provide NP-hardness complexity results, and optimal mapping algorithms for polynomial problem instances. Efficient heuristics are presented to solve the general case, and experimental results are provided.

6.15. Scheduling linear chain streaming applications on heterogeneous systems with failures

In this work [5], we study the problem of optimizing the throughput of streaming applications for heterogeneous platforms subject to failures. Applications are linear graphs of tasks (pipelines), with a type associated to each task. The challenge is to map each task onto one machine of a target platform, each machine having to be specialized to process only one task type, given that every machine is able to process all the types before being specialized in order to avoid costly setups. The objective is to maximize the throughput, i.e., the rate at which jobs can be processed when accounting for failures. Each instance can thus be performed by any machine specialized in its type and the workload of the system can be shared among a set of specialized machines. For identical machines, we prove that an optimal solution can be computed in polynomial time. However, the problem becomes NP-hard when two machines may compute the same task type at different speeds. Several polynomial time heuristics are designed for the most realistic specialized settings. Simulation results assess their efficiency, showing that the best heuristics obtain a good throughput, much better than the throughput obtained with a random mapping. Moreover, the throughput is close to the optimal solution in the particular cases where the optimal throughput can be computed.

6.16. A survey of pipelined workflow scheduling: Models and algorithms

In this survey [4], we consider a large class of applications that need to execute the same workflow on different data sets of identical size. Efficient execution of such applications necessitates intelligent distribution of the application components and tasks on a parallel machine, and the execution can be orchestrated by utilizing task-, data-, pipelined-, and/or replicated-parallelism. The scheduling problem that encompasses all of these techniques is called pipelined workflow scheduling, and it has been widely studied in the last decade. Multiple models and algorithms have flourished to tackle various programming paradigms, constraints, machine behaviors or optimization goals. This work surveys the field by summing up and structuring known results and approaches.

6.17. Reclaiming the energy of a schedule: Models and algorithms

In this work [1], we consider a task graph to be executed on a set of processors. We assume that the mapping is given, say by an ordered list of tasks to execute on each processor, and we aim at optimizing the energy consumption while enforcing a prescribed bound on the execution time. Although it is not possible to change the allocation of a task, it is possible to change its execution speed. Rather than using a local approach such as backfilling, we consider the problem as a whole and study the impact of several speed variation models on its complexity. For continuous speeds, we give a closed-form formula for trees and series-parallel graphs, and we cast the problem into a geometric programming problem for general directed acyclic graphs. We show that the classical dynamic voltage and frequency scaling (DVFS) model with discrete modes leads to an NP-complete problem, even if the modes are regularly distributed (an important particular case in practice, which we analyze as the incremental model). On the contrary, the Vdd-hopping model that allows to switch between different supply voltages (VDD) while executing a task leads to a polynomial solution. Finally, we provide an approximation algorithm for the incremental model, which we extend for the general DVFS model.

6.18. Non-clairvoyant reduction algorithms for heterogeneous platforms

In this work [24], we revisit the classical problem of the reduction collective operation in a heterogeneous environment. We discuss and evaluate four algorithms that are non-clairvoyant, i.e., they do not know in advance the computation and communication costs. On the one hand, Binomial-stat and Fibonacci-stat are static algorithms that decide in advance which operations will be reduced, without adapting to the environment; they were originally defined for homogeneous settings. On the other hand, Tree-dyn and Non-Commut-Tree-dyn are fully dynamic algorithms, for commutative or non-commutative reductions. With identical computation costs, we show that these algorithms are approximation algorithms with constant or asymptotic ratios. When costs are exponentially distributed, we perform an analysis of Tree-dyn based on Markov chains.

Finally, we assess the relative performance of all four non-clairvoyant algorithms with heterogeneous costs through a set of simulations.

6.19. Non-linear divisible loads: There is no free lunch

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 this work [23], 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.

6.20. Direct solvers for sparse linear systems

This work is closely related to the MUMPS solver (see Section 5.1) and was performed in close collaboration with INPT (Toulouse). First, we have pursued the study of low-rank representations to speed-up sparse direct solvers using the so called BLR (Block Low Rank) format [44]. This work was done in collaboration with LSTC (Livermore Software Technology Corp., USA) and in the context of a contract with EDF which funded the PhD thesis of Clément Weisbecker at INPT. We also worked on shared-memory parallelism [61] in the context of the PhD thesis of Wissam M. Sid-Lakhdar. Concerning low-rank approximations, they were experimented on geophysics applications [38] (Helmholtz equations) in the context of a collaboration with members of the ISTerre and Geoazur laboratories. The impact of both low-rank compression and shared-memory parallelism was also studied on electromagnetism problems [17], in collaboration with University of Padova (Italy) and CEDRAT.

We have started the design and implementation of a distributed-memory low-rank multifrontal solver. When computations are faster (thanks to low-rank compression or multithreading within each node), we observed that communications become critical; we are therefore currently studying the limits of the communication schemes from the MUMPS approach and their possible improvements.

On numerical and industrial aspects, we worked on rank detection and null space basis computations (in collaboration with CERFACS and Total/Hutchinson) as well as on improved parallel pivoting strategies for symmetric indefinite systems, in collaboration with ESI-Group (see Section 7.1).

6.21. Push-relabel based algorithms for the maximum transversal problem

In this work [14], 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 algorithms based on augmenting paths and pseudoflows. 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.

6.22. 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 [82], [83]. The construction algorithm has a worst-case time complexity of $\Theta(mn)$ for an $n \times n$ unsymmetric matrix having m off-diagonal nonzeros. In this work [15], we propose another algorithm that has a worst-case time complexity of $\mathcal{O}(m \log n)$. We compare the two algorithms experimentally and show that both algorithms are efficient in general. The algorithm of Eisenstat and Liu 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.

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

In this work [25], we study 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. The problem is known to be NP-complete for bipartite graphs with general vertex (task) weights, and solvable in polynomial time for unweighted graphs with unit weights (i.e., unit-weight 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 for these low-complexity algorithms, they return solutions close to the optimal (or a known lower bound) in average.

6.24. Maximum cardinality bipartite matching algorithms on GPUs

In two studies [30], [31], we propose, develop, and evaluate maximum cardinality matching algorithms from two different families (called push-relabel and augmenting-path based) on GPUs. The problem of finding a maximum cardinality matching in bipartite graphs has applications in computer science, scientific computing, bioinformatics, and other areas. To the best of our knowledge, the proposed algorithms are the first investigation of the push-relabel and augmenting-path based on GPUs/ We compare the proposed algorithms with serial and multicore implementations from the literature on a large set of real-life problems where in majority of the cases one of our GPU-accelerated algorithms is demonstrated to be faster than both the sequential and multicore implementations.

6.25. Analysis of partitioning models and metrics in parallel sparse matrix-vector multiplication

Graph/hypergraph partitioning models and methods have been successfully used to minimize the communication among processors in several parallel computing applications. Parallel sparse matrix-vector multiplication (SpMxV) is one of the representative applications that renders these models and methods indispensable in many scientific computing contexts. In this work [36], [55], we investigate the interplay of the partitioning metrics and execution times of SpMxV implementations in three libraries: Trilinos, PETSc, and an in-house one. We carry out experiments with up to 512 processors and investigate the results with regression analysis. Our experiments show that the partitioning metrics influence the performance greatly in a distributed memory setting. The regression analyses demonstrate which metric is the most influential for the execution time of the libraries.

6.26. On partitioning and reordering problems in a hierarchically parallel hybrid linear solver

PDSLin is a general-purpose algebraic parallel hybrid (direct/iterative) linear solver based on the Schur complement method. The most challenging step of the solver is the computation of a preconditioner based

on the global Schur complement. Efficient parallel computation of the preconditioner gives rise to partitioning problems with sophisticated constraints and objectives. In this work [39], we identify two such problems and propose hypergraph partitioning methods to address them. The first problem is to balance the workloads associated with different subdomains to compute the preconditioner. We first formulate an objective function and a set of constraints to model the preconditioner computation time. Then, to address these complex constraints, we propose a recursive hypergraph bisection method. The second problem is to improve the data locality during the parallel solution of a sparse triangular system with multiple sparse right-hand sides. We carefully analyze the objective function and show that it can be well approximated by a standard hypergraph partitioning method. Moreover, an ordering compatible with a post ordering of the subdomain elimination tree is shown to be very effective in preserving locality. To evaluate the two proposed methods in practice, we present experimental results using linear systems arising from some applications of our interest. First, we show that in comparison to a commonly-used nested graph dissection method, the proposed recursive hypergraph partitioning method reduces the preconditioner construction time, especially when the number of subdomains is moderate. This is the desired result since PDSLin is based on a two-level parallelization to keep the number of subdomains small by assigning multiple processors to each subdomain. We also show that our second proposed hypergraph method improves the data locality during the sparse triangular solution and reduces the solution time. Moreover, we show that partitioning time can be greatly reduced while maintaining its quality by removing quasi-dense rows from the solution vectors.

6.27. UMPA: A Multi-objective, multi-level partitioner for communication minimization

In this work [42], we propose a directed hypergraph model and a refinement heuristic to distribute communicating tasks among the processing units in a distributed memory setting. The aim is to achieve load balance and minimize the maximum data sent by a processing unit. We also take two other communication metrics into account with a tie-breaking scheme. With this approach, task distributions causing an excessive use of network or a bottleneck processor which participates to almost all of the communication are avoided. We show on a large number of problem instances that our model improves the maximum data sent by a processor up to 34% for parallel environments with 4, 16, 64, and 256 processing units compared to the state of the art which only minimizes the total communication volume.

6.28. A Partitioning-based divisive clustering technique for maximizing the modularity

In this work [43], we present a new graph clustering algorithm aimed at obtaining clusterings of high modularity. The algorithm pursues a divisive clustering approach and uses established graph partitioning algorithms and techniques to compute recursive bipartitions of the input as well as to refine clusters. Experimental evaluation shows that the modularity scores obtained compare favorably to many previous approaches. In the majority of test cases, the algorithm outperformed the best known alternatives. In particular, among 13 problem instances common in the literature, the proposed algorithm improves the best known modularity in 9 cases.

6.29. Randomized matching heuristics with quality guarantees on shared memory parallel computers

In this work [56], we propose two heuristics for the bipartite matching problem that are amenable to shared-memory parallelization. The first heuristic is very intriguing from parallelization perspective. It has no significant algorithmic synchronization overhead and no conflict resolution is needed across threads. We show that this heuristic has an approximation ratio of around 0.632. The second heuristic is designed to obtain a larger matching by employing the well-known Karp-Sipser heuristic on a judiciously chosen subgraph of the original graph. We show that the Karp-Sipser heuristic always finds a maximum cardinality matching in the chosen subgraph. Although the Karp-Sipser heuristic is hard to parallelize for general graphs, we exploit the

structure of the selected subgraphs to propose a specialized implementation which demonstrates a very good scalability. Based on our experiments and theoretical evidence, we conjecture that this second heuristic obtains matchings with cardinality of at least 0.866 of the maximum cardinality. We discuss parallel implementations of the proposed heuristics on shared memory systems. Experimental results, for demonstrating speed-ups and verifying the theoretical results in practice, are provided.

6.30. On the minimum edge cover and vertex partition by quasi-cliques problems

A γ-quasi-clique in a simple undirected graph is a set of vertices which induces a subgraph with the edge density of at least γ for $0 < \gamma < 1$. A cover of a graph by γ-quasi-cliques is a set of γ-quasi-cliques where each edge of the graph is contained in at least one quasi-clique. The minimum cover by γ-quasi-cliques problem asks for a γ-quasi-clique cover with the minimum number of quasi-cliques. A partition of a graph by γ-quasi-cliques is a set of γ-quasi-cliques where each vertex of the graph belongs to exactly one quasi-clique. The minimum partition by γ-quasi-cliques problem asks for a vertex partition by γ-quasi-cliques with the minimum number of quasi-cliques. In this work [60], we show that the decision versions of the minimum cover and partition by γ-quasi-cliques problems are NP-complete for any fixed γ satisfying $0 < \gamma < 1$.

RUNTIME Project-Team

6. New Results

6.1. SIMD Analysis Support in MAQAO

Either on ARM and x86 architectures, compilers and tools are needed for automatic and efficient vectorization. Although commercial compilers (e.g. IBM xlc, Intel icc, PGI pgcc) have made significant advances in auto-vectorization, a lot of source codes still remain too complicated for a compiler to vectorize, particularly when complex data structures are involved, or because of the lack of information at compile time. However, when vectorization fails, compilers leave the user with little clues about the cause of the failure, even though in certain cases moderate modifications could be applied on the source code to enable the compiler to vectorize.

Thus, the main objective for this work was to analyse SIMD vectorization potentials through loop detection. Parallelism detection is done through the instrumentation of the binary codes, capturing all memory streams in target loops and computing memory dependences using MAQAO. When combined to a static analysis for register dependences, this technique ensures that parallel slices of computation will be detected.

From a practical point of view, this work consists in the capture of the trace and its processing to extract memory reference patterns. To do so, we made use of the current state of the art MAQAO for instrumentation and trace capture on Intel architectures. We then implemented the dependence analysis on memory traces for performing loop pattern recognition. Finally, using this mechanism for loop pattern recognition, we can conclude about the vectorization potential of computation intensive loop nests. The dependence analysis does not depend on the target architecture, hence results computed for x86 architectures are valuable for ARM target as well.

6.2. NUMA-aware fine grain parallelization for multi-core architecture

Today, popular frameworks like Intel TBB or OpenMP offer a task based programming interface that allows to easily parallelize algorithms in shared memory. We have proposed some improvements to these task-based parallelization frameworks in order to cope with the problem of expressing an algorithm with a suitable task grain size and with the problem of Non Uniform Memory Accesses that degrades performance. In its current prototype state, our framework does not fully automate the selection of an optimal grain size. However, it significantly helps the programmer by proposing a simple interface to deal with DAG coarsening.

We have shown the benefits of this work on the parallelization of a sparse ILU preconditioner which is a challenging application with respect to task grain tuning and NUMA effect to an Intel TBB implementation. To improve even more the NUMA aspects, we are working on improving the task scheduler with cache-aware hierarchical scheduling support using a similar approach as the one implemented in the Bubblesched thread scheduler.

6.3. Task scheduling over heterogeneous architectures

We continued our work on extending StarPU to master exploitation of Heterogeneous Platforms through dynamic task scheduling, leading to the release of StarPU 1.1. We have extended our lightweight DSM to support out-of-core scheduling over disks. We have finished integrating StarPU with SimGrid and obtained very accurate simulated times, which allows to experiment scheduling heuristics without having to actually execute the application on the target platform, thus tremendously reducing experimentation time and resource consumption.

We have modularized the scheduling part of StarPU, which permits to create complex schedulers by assembling simple scheduling components. This will allow theoreticians to work on writing the simple scheduling components without having to deal with the technical parts of the scheduling, performed in other scheduling components.

We have also collaborated with various research project to leverage the potential of StarPU: for instance, the PaStiX sparse matrix solver was ported over StarPU, so that we improved the dynamic task and management for applications with such fine-grain task size. This resulted with fair-enough performance on CPUs, compared to the hand-optimized static scheduler of PaStiX, and very promising performance on CPUs + GPUs. EADS ported its sparse hmatrix solver over StarPU, and we collaborated to work on adding StarPU support for communicating sparse data over MPI.

6.4. Task Size Control with XcalableMP/StarPU

On the work sharing among GPUs and CPU cores on GPU equipped clusters, it is a critical issue to select the task computational weights suited to these heterogeneous computing resources. We have been developing a solution for this problem, based on the cooperation of a PGAS language named XcalableMP (developed at the University of Tsukuba) together with a runtime sytem named XMP-dev/StarPU building on the work of the University of Tsukuba and on the StarPU platform developed by the Inria Runtime Team. Through the development, we found the necessity of adaptive task weight control for the GPU/CPU work sharing to achieve the best performance for various application codes. In particular, the language was extended to add a new feature allowing to alter the task size to be assigned to these heterogeneous resources dynamically during application execution. As a result of performance evaluation on several benchmarks, we confirmed the proposed feature correctly works and perform well even for relatively small size of problems.

6.5. Scheduling contexts for StarPU

Scheduling context is an extension of StarPU that allows multiple parallel codes to run concurrently with minimal interference. A scheduling context encapsulates an instance of the runtime system, and runs on top of a subset of the available processing units (i.e. regular cores or GPU accelerators). In order to maximize the overall efficiency of applications, contexts can be dynamically shrunk or expanded by a *hypervisor* that periodically gathers performance statistics inside each context (e.g. resource utilization, computation progress) and tries to determine how resources should be assigned to contexts so as to minimize the overall execution time. We have demonstrated the relevance of this approach using benchmarks invoking multiple high performance linear algebra kernels simultaneously on top of heterogeneous multicore machines. We have shown that our mechanism can dramatically improve the overall application run time (-34%), most notably by reducing the average cache miss ratio (-50%).

6.6. Load-balancing with TreeMatch

In the context of the Joint Laboratory for Petascale Computing (JLPC) included Inria and the University of Illinois at Urbana, we developed two load balancers for Charm++.

The two load-balancers we wrote take into account both the computing power and the hierarchical topology depending on the fact that the application is compute-bound or communication-bound. This work is based on our TreeMatch library that computes process placement in order to reduce an application communication costs based on the hardware topology. Compared to some other solutions based on weighted topologies (latency, bandwidth, ...), ours is fully dynamic because we use only a qualitative approach for our representation of the hardware architecture.

The first load balancer is designed for compute-bound applications as it favours the leveling of CPU loads. The second load balancer focuses on communication-bound applications as it first reduces the congestion on the upper links in the topology tree.

These two load balancers gave us improvements for some applications up to 10% of the execution time.

6.7. List scheduling in embedded systems taking into account memroy constraints

Video decoding and image processing in embedded systems are subject to strong resource constraints, particularly in terms of memory. List-scheduling heuristics with static priorities (HEFT, SDC, etc.) being the often-cited solution due to both their good performance and their low complexity, we propose a method aimed at introducing the notion of memory into them. Moreover, we show that through appropriate adjustment of task priorities and judicious resort to insertion-based policy, speedups up to 20% can be achieved. Lastly, we show that our technique allows to prevent deadlock and to substantially reduce the required memory footprint compared to classic list-scheduling heuristics.

6.8. NewMadeleine generic multi-threading

The PIOMan progression engine utilized in NewMadeleine used to rely on the Marcel specific multi-threading library, with dedicated hooks and close co-operation between libraries. It restricted the target platforms and applications, and was considered as a constraint by users. We have designed mechanisms to make communication progress without hooks in the thread scheduler, able to run on any system with a `pthread` library. We have re-written PIOMan from the ground up to implement these mechanisms, and based on lock-free structures, with scalability in mind. A proof-of-concept port to the Intel Xeon Phi has been implemented in cooperation with the University of Tokyo, using the DCFA (Direct Communication Facility for manycore-based Accelerators) library to access InfiniBand boards from the Xeon Phi.

ASCOLA Project-Team

6. New Results

6.1. Software composition

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

6.1.1. Service-oriented computing

Services are frequently implemented using object-oriented frameworks. In this context, two properties are particularly important: (i) a loose coupling between the service layer and the object layer, allowing evolution of the service layer with a minimal impact on the object layer, (ii) interoperability induced by the substitution principle associated to subtyping in the object layer, thus allowing to freely convert a value of a subtype into a supertype. However, through experimentation with Apache's popular service framework CXF, we observed some undesirable coupling and interoperability issues due to the failure of the substitution principle [23]. Therefore we have proposed a new specification method for the data binding used to translate data between the object and service layers [24]. We have shown that if the CXF framework follows the specification, the substitution principle is satisfied, with all its advantages.

6.1.2. Modularity and program transformations

Refactoring tools are commonly used for remodularization tasks. Basic refactoring operations are 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 [31], we have recast two calculus for static composition of refactorings in a type system framework and we have discussed their use for inferring useful properties. We have illustrated the value of support for static composition in refactoring tools with a complex remodularization use case: a round-trip transformation between programs conforming to the Composite and Visitor patterns. Composite and Visitor design patterns have dual properties with respect to modularity, thus they are good candidates to explore their transformations. In [22] we have extended our initial refactoring-based round-trip transformation between these two structures and we have studied how that transformation is impacted by four variations in the implementation of these patterns. We have validated that study by computing the smallest preconditions for the resulting transformations. We have also automated the transformation and applied it to JHotDraw, where the studied variations occur. Finally, [11] presents more exhaustively modular transformations and design patterns. We have also proposed a reversible transformation in the Singleton pattern to benefit from optimization by introducing this pattern and flexibility by its suppression according to the requirements of the software user.

6.1.3. Domain specific languages

In the context of Charles Prud'hommes' PhD Thesis, we have developed a domain specific language in order to specify strategies of filtering propagation in constraint solvers. Indeed, constraint programming replaces brute force generate-and-test by the exploration of the solution space based on incremental instantiation and constraint propagation. Strategies of incremental instantiation (also known as heuristics) have been heavily studied. However, most solvers propagate constraints with a simple fix point computation based on a queue of constraints to propagate (or several queues in order to deal with the grain/cost of filtering algorithms). This technique has a good behavior in general but for a given problem a dedicated strategy can be more efficient. Our declarative DSL and its support in the new version of the constraint solver Choco [19], [52] enables us to easily experiment with different propagation strategies. Moreover, our DSL supports properties such as completeness, intended incompleteness or non ambiguity.

6.1.4. Constructive security

In the field of techniques for the development of secure software systems we have presented results on the enforcement of security properties in service-oriented systems and Javascript programs.

Concerning the security of service-based systems, we have first presented a software framework that harnesses a type based policy language and aspect-based support for protocol adaptation in service-oriented systems by means of flexible reference monitors [29], [28]. We have shown how this framework improves the security, interoperability and evolution issues of service systems using the OAuth 2.0 standard for the authorization of resource accesses. The OAuth 2 protocol is a recent IETF standard devoted to providing authorization to clients requiring access to specific resources over HTTP. It was recently adopted by major internet companies and software editors, such as Google, Facebook, Microsoft, and SAP. We have shown how to improve the security of software systems that use OAuth 2 in the presence of different kinds of clients.

Furthermore, we have developed a new notion of transformation operators, so-called workflow adaptation schemas (WASs) for service compositions that facilitates the integration and modification of security functionalities of service-oriented systems [30]. These schemas may be generic and specialized through parameter instantiation. A set of schemas therefore effectively provides a domain-specific language for the transformation of service-oriented applications. We have developed a set of specific schemas and applied them to the OAuth 2 standard in order to implement state-based security hardening strategies. We have also implemented tool support for WASs and implemented some of the security scenarios involving OAuth 2 (see Sec. 5.4).

Finally, we have shown that a wide range of strategies to make secure JavaScript-based applications can be described pertinently using aspects [42]. To this end, we have reviewed major categories of approaches to make client-side applications secure and have discussed uses of aspects that exist for some of them. We also propose aspect-based techniques for the categories that have not been studied previously. We have given examples of applications where aspects are useful as a general means to flexibly express and implement security policies for JavaScript.

6.2. Aspect-Oriented Programming

Participants: Rémi Douence, Ismael Figueroa, Jacques Noyé, Mario Südholt, Nicolas Tabareau, Jurgen Van Ham.

6.2.1. Aspects in a concurrent and distributed setting

Aspect oriented programming modularizes crosscutting concerns by gathering several join points. In the context of distributed applications these point cuts can be on different machines. In this case, a sequence of join points must be defined as a sequence of logical joint points (à la Lamport). We propose an aspect oriented languages to define distributed aspects in JavaScript in a distributed context. Our proposal [18] is based on vector clocks in order to logically relate join points and can ignore "illogical" (that is late or early) join points. In can also enforce causal communications when no join point must be discarded. We have exemplified the advantages of our technique with different applications such as a discussion forum, a retweet scenario and a web browser.

Multiparty session types allow the definition of distributed processes with strong communication safety properties. A global type is a choreographic specification of the interactions between peers, which is then projected locally in each peer. Well-typed processes behave accordingly to the global protocol specification. Multiparty session types are however monolithic entities that are not amenable to modular extensions. Also, session types impose conservative requirements to prevent any race condition, which prohibit the uniform application of extensions at different points in a protocol. We have proposed a means to support modular extensions with *aspectual session types* [47], a static pointcut/advice mechanism at the session type level. To support the modular definition of crosscutting concerns, we augment the expressivity of session types to allow harmless race conditions. We formally prove that well-formed aspectual session types entail communication safety. As a result, aspectual session types make multiparty session types more flexible, modular, and extensible.

We have added dedicated concurrency support to EScala, our extension of Scala that introduces composable *declarative events* as a way to integrate Aspect-Oriented Programming and Event-Based Programming in the context of Object-Oriented Programming. In JEScala, Events, which were synchronous in EScala, can be declared as *asynchronous* so that they are handled concurrently to their emitter. Moreover, two new operators, a join and a choice operator, inherited from the join calculus - hence the name of the new prototype, can now be used to compose events and control concurrency. In [48], we present JEScala, show that it captures coordination schemas in a more expressive and modular way than plain join languages and provide a first performance assessment.

6.2.2. Effective aspects

We have proposed a novel approach to embed pointcut/advice aspects in a typed functional programming language like Haskell. Aspects are first-class, can be deployed dynamically, and the pointcut language is extensible. Type soundness is guaranteed by exploiting the underlying type system, in particular phantom types and a new anti-unification type class. The use of monads brings type-based reasoning about effects for the first time in the pointcut/advice setting and enables modular extensions of the aspect language [46], [16].

To allow a type-safe embedding of aspects in Haskell, we had to develop a notion of anti-unification in Haskell type system. The anti-unification problem is that of finding the most specific pattern of two terms. While dual to the unification problem, anti-unification has rarely been considered at the level of types. We have developed an algorithm to compute the least general type of two types in Haskell, using the logic programming power of type classes [53]. That is, we have defined a type class for which the type class instances resolution performs anti-unification.

6.2.3. Reasoning about aspect interference

When a software system is developed using several aspects, special care must be taken to ensure that the resulting behavior is correct. This is known as the *aspect interference problem*, and existing approaches essentially aim to detect whether a system exhibits problematic interferences of aspects. We have described how to control aspect interference by construction by relying on the type system. More precisely, we combine a monadic embedding of the pointcut/advice model in Haskell with the notion of membranes for aspect-oriented programming [34]. Aspects must explicitly declare the side effects and the context they can act upon. Allowed patterns of control flow interference are declared at the membrane level and statically enforced. Finally, computational interference between aspects is controlled by the membrane topology. To combine independent and reusable aspects and monadic components into a program specification we use *monad views*, a recent technique for conveniently handling the monadic stack.

Oliveira and colleagues recently developed a powerful model to reason about mixin-based composition of effectful components and their interference, exploiting a wide variety of techniques such as equational reasoning, parametricity, and algebraic laws about monadic effects. Our work addresses the issue of reasoning about interference with effectful aspects in the presence of unrestricted quantification through pointcuts. While global reasoning is required, we have shown that it is possible to reason in a compositional manner, which is key for the scalability of the approach in the face of large and evolving systems. We have established a general equivalence theorem that is based on a few conditions that can be established, reused, and adapted separately as the system evolves. Interestingly, one of these conditions, local harmlessness, can be proven by a translation to the mixin setting, making it possible to directly exploit previously established results about certain kinds of harmless extensions [33].

In aspect-oriented programming (AOP) languages, advice evaluation is usually considered as part of the base program evaluation. While viewing aspects as part of base level computation clearly distinguishes AOP from reflection, it also comes at a price: because aspects observe base level computation, evaluating pointcuts and advice at the base level can trigger infinite regression. To avoid these pitfalls, we have introduced levels of execution in the programming language, thereby allowing aspects to observe and run at specific, possibly different, levels. We adopt a defensive default that avoids infinite regression, and gives advanced programmers the means to override this default using level-shifting operators [21].

6.3. Resource management in Cloud computing

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

We have contributed on several topics: multiple autonomic managers for Cloud infrastructure, SLA management for Cloud elasticity, fully distributed and autonomous virtual machine scheduling, and simulator toolkits for IaaS platforms.

6.3.1. Cloud infrastructure based on multiple autonomic managers

One of the main reasons for the wide adoption of Cloud Computing is the concept of elasticity. Implementing elasticity to tackle varying workloads while optimizing infrastructures (e.g. utilization rate) and fulfilling the application requirements on Quality of Service should be addressed by self-adaptation techniques able to manage complexity and dynamism. However, since Cloud systems are organized in different but dependent Cloud layers, self-management decisions taken in isolation in a certain layer may indirectly interfere with the decision taken by an other layer. Indeed, non-coordinated managers may lead to conflicting decisions and consequently to non-desired states.

We have proposed a framework for the coordination of multiple autonomic managers in cloud environments [25]. The PhD thesis of Frederico Alvares [12], defended in April 2013, is based on this framework. This thesis proposes a self-adaptation approach that considers both application internals (architectural elasticity) and infrastructure (resource elasticity), managed by multiple autonomic managers, to reduce the energy footprint in Cloud infrastructures.

6.3.2. SLA Management for Cloud elasticity

Elasticity is the intrinsic element that differentiates Cloud Computing from traditional computing paradigms, since it allows service providers to rapidly adjust their needs for resources to absorb the demand and hence guarantee a minimum level of Quality of Service (QoS) that respects the Service Level Agreements (SLAs) previously defined with their clients. However, due to non-negligible resource initiation time, network fluctuations or unpredictable workload, it becomes hard to guarantee QoS levels and SLA violations may occur. The main challenge of service providers is to maintain its consumer's satisfaction while minimizing the service costs due to resources fees. The PhD thesis of Yousri Kouki [13], defended in December, proposes different contributions to address this issue: CSLA, a specific language to describe SLA for Cloud services ; HybridScale, an auto-scaling framework driven by SLA [39], [17].

6.3.3. Fully Distributed and Autonomous Virtualized Environments

We have consolidated the DVMS system to obtain a fully distributed virtual machine scheduler [44]. 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 [44] 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) [56]. 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.

6.3.4. Testing the cloud

Computer science, as other sciences, needs instruments to validate theoretical research results, as well as software developments. Although simulation and emulation are generally used to get a glance of the behavior of new algorithms, they use over-simplified models in order to reduce their execution time and thus cannot be accurate enough. Leveraging a scientific instrument to perform actual experiments is an undeniable advantage. However, conducting experiments on real environments is still too often a challenge for researchers, students, and practitioners: first, because of the unavailability of dedicated resources, and second, because of the inability to create controlled experimental conditions, and to deal with the wide variability of software

requirements. During 2013, we have contributed to a new topic addressing the "testing the cloud" challenge. First, we have presented the latest mechanisms we have designed to enable the automated deployment of the major open-source IaaS cloudkits (i.e., Nimbus, OpenNebula, CloudStack, and OpenStack) on Grid'5000 [26]. Providing automatic, isolated and reproducible deployments of cloud environments lets end-users study and compare each solution or simply leverage one of them to perform higher-level cloud experiments (such as investigating Map/Reduce frameworks or applications). Moreover, we have presented EXECO, a library that provides easy and efficient control of large scale experiments through a set of tools well as tools designed for scripting distributed computing experiments on any computing platform. We have illustrated its interest by presenting two experiments dealing with virtualization technologies on the Grid'5000 testbed [37].

6.3.5. Adding virtualization abstractions into the Simgrid toolkit

In the context of the ANR SONGS project and in collaboration with Takahiro Hirofuchi, researcher at AIST (Japan), we have extended the Simgrid framework to be able to simulate virtualized distributed infrastructures [35]. In addition, we have proposed the first class support of live migration operations within such a simulator toolkit for large scale distributed infrastructures. We have developed a resource share calculation mechanism for VMs and a live migration model implementing the precopy migration algorithm of Qemu/KVM. We have confirmed that our simulation framework correctly reproduced live migration behaviors of the real world under various conditions [36].

6.3.6. Power and energy management in the cloud

Power management has become one of the main challenges for data center infrastructures. Currently, the cost of powering a server is approaching the cost of the server hardware itself, and, in a near future, the former will continue to increase, while the latter will go down. In this context, virtualization is used to decrease the number of servers, and increase the efficiency of the remaining ones.

First, in [43] we have proposed an approach and a model to estimate the total power consumption of a virtual machine, by taking into account its static (e.g. memory) and dynamic (e.g. CPU) consumption of resources. Second, we have rewritten the Entropy framework (in OptiPlace) to give it the support of external models, named views. Entropy, based on the Constraint Programming solver Choco written in Java, does not really scale well. We have studied Entropy's scalability properties [32] and have then integrated heuristics and constraints in OptiPlace [40].

The evaluation of these policies on real infrastructures has become an important and difficult issue. The corresponding techniques have become so complex that there is a need for load injection frameworks able to inject resource load in a tested datacenter instead of model-driven simulation. For this reason we have developed StressCloud [41], [51], a framework to manipulate the activities of a group of Virtual Machines and observe the resulting performance.

FOCUS Project-Team

6. New Results

6.1. Service-oriented computing

Participants: Mario Bravetti, Ivan Lanese, Fabrizio Montesi, Gianluigi Zavattaro.

6.1.1. Primitives

We have studied primitives used in the context of service-oriented computing, at different levels of abstraction and in different contexts. At the abstract level, we considered both standard web services and Internet of Things, where computational and communication capabilities are attached to real-world objects such as smartphones or alarm clocks. For web services, we defined SSCC (Stream-based Service-Centered Calculus) [17], a calculus allowing to describe both service composition (orchestration) via streams and the protocols that services follow when invoked (conversation). We assessed the expressive power of SSCC by modeling van der Aalst's workflow patterns and an automotive case study from the European project Sensoria. For analysis, we presented a simple type system ensuring compatibility of client and service protocols. We also studied the behavioral theory of the calculus, highlighting some axioms that capture the behavior of the different primitives. As a final application of the theory, we defined and proved correct some program transformations. For Internet of Things, a main contribution [37] has been the definition of a calculus and of an equivalence allowing to capture the behavior of the system as seen by the human end-user. Since this equivalence is not compositional we defined also a finer equivalence which is compositional. We showed how our equivalences can be applied to reason on simple Internet of Things examples.

At a more concrete level, we have continued to study and extend the Jolie language. In [44] we present a detailed description of the Jolie language. We put our emphasis on how Jolie can deal with heterogeneous services. On the one hand, Jolie combines computation and composition primitives in an intuitive and concise syntax. On the other hand, the behavior and deployment of a Jolie program are orthogonal: they can be independently defined and recombined as long as they have compatible typing. In [42] we extended Jolie to model process-aware web information systems. Our major contribution is to offer a unifying approach for the programming of distributed architectures based on HTTP that support typical features of the process-oriented paradigm, such as structured communication flows and multiparty sessions.

6.1.2. Choreographies

Choreographies are high-level descriptions of distributed interacting systems featuring as basic unit a communication between two participants. A main feature of choreographies is that they are deadlock-free by construction. From a choreography one can automatically derive the behavior of each participant using a notion of projection. Under suitable conditions on the structure of a choreography, the correctness of its projection can be established in terms of a trace-based semantics. In [24] we have proposed a purely-global programming model. The idea is to define abstract choreographies – called protocol specifications – and use them to type a more concrete choreography. This more concrete choreography is used to generate executable code for the different participants. The approach is based on a novel interpretation of asynchrony and parallelism. We evaluated the approach by providing a prototype implementation for a concrete programming language and by applying it to some examples from multicore and service-oriented programming [49]. In [43] we tackled one of the main limitations of choreographies, namely the fact that they model closed systems. To this end we proposed a notion of composable choreographies. The key of our approach is the introduction of partial choreographies, which can mix global descriptions with communications among external peers. We prove that if two choreographies are composable, then the endpoints independently generated from each choreography are also composable, preserving their typability and deadlock-freedom. In [39] we showed how to transform choreographies which do not satisfy the conditions for their projection into choreographies that satisfy them preserving their behavior and enabling a correct projection.

6.2. Models for reliability

Participants: Ivan Lanese, Michael Lienhardt, 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. In [38] we present croll-pi, a concurrent calculus extending roll-pi – an higher-order pi-calculus featuring a rollback operator – allowing the specification of alternatives to a computation to be used upon rollback. Alternatives in croll-pi are attached to messages. We show the robustness of this mechanism by encoding more complex idioms for specifying alternatives. We illustrate the expressiveness of our approach by encoding a calculus of communicating transactions and by modeling the 8-queens problem. We also formally prove that croll-pi is strictly more expressive than roll-pi.

6.2.2. Compensations

We have continued the study of the expressive power of primitives for specifying compensations in long running transactions. Dynamic compensation installation allows for easier specification of fault handling in complex interactive systems since it enables to update the compensation policies according to run-time information. In [40] we show that in a simple pi-like calculus with static compensations the termination of a process is decidable, but it is undecidable in one with dynamic compensations. We then consider three commonly used patterns for dynamic compensations: parallel compensations, where new compensation items can only be added in parallel, replacing compensations, where old compensations are replaced, and nested compensations, where old compensations can be used (linearly) to build new ones. We show that termination is decidable in the first two cases and undecidable in the last one.

6.3. Cloud Computing

Participants: Elena Giachino, Michael Lienhardt, Tudor Alexandru Lascu, Jacopo Mauro, Gianluigi Zavattaro.

6.3.1. Languages for cloud applications

To foster the industrial adoption of virtualized services, it is necessary to address two important problems: (1) the efficient analysis, dynamic composition of services with qualitative and quantitative service levels and (2) the dynamic control of resources such as storage and processing capacities according to the internal policies of the services. Current technologies for cloud computing, addresses these problems at deployment and run time. The ENVISAGE project and the position paper [20] proposes, on the contrary, to overcome these problems by leveraging service-level agreements into software models and resource management into early phases of service design.

6.3.2. Models for cloud application deployment

Cloud computing offers the possibility to build sophisticated software systems on virtualized infrastructures at a fraction of the cost necessary just few years ago, but deploying/maintaining/reconfiguring such software systems is a serious challenge. The AEOLUS project, aims to tackle the scientific problems that need to be solved in order to ease the problem of efficient and cost-effective deployment and administration of the complex distributed architectures which are at the heart of cloud applications [25]. In particular, it is necessary to define appropriate models for the representation of the interdependencies among the software components of a cloud application as well as declarative languages for the specification of the desired application configuration. We have proposed [31] a model for the representation of the component lifecycle and of its dependencies/conflicts with the other components. Based on such model, we have defined a sound and complete algorithm that efficiently computes a deployment plan (i.e. a sequence of low-level component deployment actions) capable of reaching a final configuration including at least some predefined basic components [48] and we have realized a prototypical implementation of such algorithm which was proved to be effective on case-studies of realistic size (i.e. hundreds of components) [41].

6.4. Resource Control

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

6.4.1. Expressive type systems for complexity analysis

Along 2013, our work on expressive methodologies for complexity analysis of higher-order languages has proceeded. In particular, we have focused our attention on extending linear dependent types to languages with control operators in the style of `callcc` [27]. This has taken the form of a generalization of bounded linear logic towards Laurent's polarized linear logic, which is then turned into a type system for the lambda-mu-calculus (in which the aforementioned control operator can indeed be implemented). In the introduced type system, all typable terms can be reduced in polynomial time. We also worked on the linear dependent type inference and on its implementation (though the work has not yet been transferred onto the *Lideal* tool implementing type inference for dependently linear type systems, see http://lideal.cs.unibo.it/); more specifically, we showed that type inference can in this context be reduced to a form of constraint amenable to be solved by SMT solvers [28]. Finally, a call-by-value version of $\mathsf{d}\ell\mathsf{PCF}$ has been defined and proved sound but also relatively complete as a tool for complexity analysis of programs [16].

6.4.2. Complexity analysis and process algebras

Complexity analysis methodologies drawn from linear logic have been adapted to higher-order process algebras, obtaining linear versions of the higher-order π-calculus in which reduction sequences are guaranteed to have a length bounded by a polynomial [14]. This is done by following the exponential discipline Lafont's Soft Linear logic suggests.

6.4.3. Characterizing probabilistic complexity classes

We have also been looking [10] (papers extracted from the thesis should appear soon) at probabilistic computation and at whether probabilistic complexity classes like **BPP**, **ZPP** and **PP** can be characterized by logics and λ-calculi. We encountered some problems in doing the above for **BPP** and **ZPP**, which are semantic classes and which, as a consequence, cannot be easily enumerated (and captured by ICC systems). On the other hand, probabilistic classes like **PP** can indeed be characterized by λ-calculi, as shown by our recent work on RSLR, a system derived from Hofmann's SLR that captures the (deterministic) polytime computable functions.

6.4.4. Ensuring differential privacy

Differential privacy offers a strong guaranteed bound on loss of private information due to release of query results, even under worst-case assumptions. One of the challenges in proving queries differentially private is to prove an upper bound on the query's sensitivity, i.e., the maximum change in the query's output that can result from changing the data of a single individual. Reed and Pierce have recently proposed a type analysis using numerical annotations in types to describe bounds on the sensitivity of the queries. A first delicate aspect of this approach is that in order to verify if a program is typable or not one needs to come up with numerical annotations and verify their consistency. Finding a "small" annotation is crucial, since the privacy depends on it. For this reason we designed a sensitivity inference tool [26] that combined with the Z3 SMT solver is able to verify and infer a minimal sensitivity bound in an automatic and efficient way. Another delicate aspect of this approach is the *expressivity* of the type analysis. Reed and Pierce's type system offers only a very limited form of numerical annotations. These numerical annotations are not enough to provide a bound for programs whose sensitivity depends on data available only at runtime. To recover this problem we introduced *Dfuzz* [32], a language combining linear types and lightweight dependent types.

6.5. Verification of extensional properties

Participants: Ornela Dardha, Elena Giachino, Michael Lienhardt, Cosimo Laneve, Fabrizio Montesi.

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. Most work carried out this year has to do with type systems for concurrent objects and components ensuring safe and reliable interactions, and on deadlock analysis for systems of concurrent objects or within process sessions.

6.5.1. Type systems for objects and components

In previous work, we had developed an integration of session types, for specifying and validating structured communication sequences (sessions) into a class-based core object language for building network applications. We have defined [12] a constraint-based type system that reconstructs the appropriate session types of session declarations instead of assuming that session types are explicitly given by the programmer, and used static analysis via types to ensure that, once a session has started, computation cannot get stuck on a communication deadlock.

In previous papers, we had proposed a component layer for object-oriented language ABS (studied in the EU project Hats), that allows one to perform updates on objects by means of communication ports and their rebinding. We have now [29] introduced a type system for this component model that statically enforces that no object will attempt illegal rebinding.

6.5.2. Deadlock analysis

Deadlock represents an insidious and recurring threat when systems also exhibit a high degree of resource and data sharing. We address deadlock analysis of two such systems: (1) concurrent object-oriented languages; (2) protocol specifications.

For (1), we have developed a framework for statically detecting deadlocks in a concurrent object-oriented language with asynchronous method calls and cooperative scheduling of method activations. Since this language features recursion and dynamic resource creation, deadlock detection is extremely complex and state-of-the-art solutions either give imprecise answers or do not scale. In order to augment precision and scalability we propose a modular framework that allows several techniques to be combined. The basic component of the framework is a front-end inference algorithm that extracts abstract behavioral descriptions of methods, called contracts, which retain resource dependency information [33]. This component is integrated with a number of possible different back-ends that analyze contracts and derive deadlock information. As a proof-of-concept, we discuss two such back-ends: (i) an evaluator that computes a fixpoint semantics [33] and (ii) an evaluator using abstract model checking [34].

For (2), in [24], we develop a typing discipline that verifies choreographies against protocol specifications, based on multiparty sessions. Exploiting the nature of global descriptions, our type system defines a new class of deadlock-free concurrent systems (deadlock-freedom-by-design), provides type inference, and supports session mobility. We give a notion of Endpoint Projection (EPP) which generates correct entity code (as pi-calculus terms) from a choreography. Finally, we evaluate our approach by providing a prototype implementation for a concrete programming language and by applying it to some examples from multicore and service-oriented programming.

Finally, en passant we remind [23], that studies deadlock analysis of concurrent object-oriented languages via encoding into Petri nets, which had already been discussed in last year's report.

6.6. Expressiveness of computational models

Participants: Roberto Amadini, Ornela Dardha, Maurizio Gabbrielli, Daniel Hirschkoff, Jean-Marie Madiot, Jacopo Mauro, Davide Sangiorgi, Gianluigi Zavattaro.

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

The fusion calculi are a simplification of the π-calculus in which input and output are symmetric and restriction is the only binder. We show [35] a major difference between these calculi and the π-calculus from the point of view of types, proving some impossibility results for subtyping in fusion calculi. We propose a modification of fusion calculi in which the name equivalences produced by fusions are replaced by name preorders, so to be able to import subtype systems, and related results, from the π-calculus. We have studied the consequences of the modification on behavioural equivalence and expressiveness.

In Focus we use notions of constraint to define in a succinct way models of computation and current constraint solving technologies to solve problems modeled using constraints. For this reason we have studied the expressive power of various computational models involving constraints and their practical impact in terms of solving/execution performances. In [18] we have investigated how the notion of constraint augments the expressive power of a concurrent language if priorities are introduced. The chosen language is Constraint Handling Rules, a committed-choice declarative language originally designed for writing constraint solvers and that is nowadays a general purpose language. The result has been otbained by first formalising the meaning of language encodings and language embedding, widely used in concurrency theory. Different ways to model and define disaster scenarios are analyzed and compared in [11], where we study a model expressive enough to define a disaster scenario that, at the same time, can be used to find plans to save the victims of a disaster using modern constraint solving technology. Similarly, different computation models are considered in [22] where we study how machine learning techniques can be used to boost the performances of constraint solvers. A technique dubbed “portfolio approach” is used to combine the different performances of constraint solvers to obtain a globally better solver using, as a starting point, a simple low-level constraint language.

In [30] we propose an integration of structural sub-typing with boolean connectives and semantic sub-typing to define a Java-like programming language that exploits the benefits of both techniques. The resulting language has a more expressive set of types that comes from the use of boolean constructs, negation types, and the integration of structural and nominal sub-typing in an object-oriented setting. By implementing traditional Java-language constructs we show that the proposed language is also expressive enough w.r.t. the Java language.

OASIS Project-Team

6. New Results

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

6.1.1. Multi-active Objects

Participants: Ludovic Henrio, Fabrice Huet, Justine Rochas.

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:

- publication of the multiactive object programming model in COORDINATION 2013 [19]
- extension of the annotations to support the specification of:
  - thread management. This aims at specifying (i) thread reservation and (ii) thread limitation in order to control more finely the allocation of threads in a multiactive object.
  - priority of requests. The progammer can now specify a priority graph to have an influence on the order of execution of requests in a multiactive object

  This extension was initially explored in a master thesis [34] and led to a publication in SAC 2014 [21].
- extensive use of multiactive objects in our CAN P2P network and implementation of usecases.

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

6.1.2. Algorithmic skeletons

Participant: Ludovic Henrio.

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

- Encapsulation of the skandium skeleton runtime in a component in order to allow distributed execution of skeletons: local parallelism is handled by skandium while distributed execution is handled by the GCM component library.
- We applied the event framework for skeletons to design a framework allowing the skeleton execution to adapt autonomically in order to achieve a required qualiti of service. We have first promising results on this aspect and a publication has just been accepted to PMAM 2014.

6.1.3. Behavioural models for Distributed Components

Participants: Eric Madelaine, Nuno Gaspar, Oleksandra Kulankhina, Ludovic 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 improved the specification of the behavioural model generation for component systems that we specified last year [36]. A journal version is under submission.
- We extended the formal model of the GCM architecture and included the specification of the non-functional aspects.
- We worked on the design of a bisimulation equivalence relation adapted to pNets; such an equivalence relation would justify some of the verifications and simplifications we do in our verification platform. Bisimulation theory gives tools to prove the equivalent behaviour of two processes, but adapting it to the structural nature and to the parameterized definitions of pNets is a challenging task. We have obtained promising preliminary results on this aspect: we have a good library of examples illustrating the expressiveness of pNets and use it to study bisimulation techniques.
- We additionally have put considerable efforts on the improvement of the Vercors platform (see Section 5.2). We have totally updated the Vercors Components Editor. We have integrated the UML state machines editor from Obeo UML Designer (http://marketplace.obeonetwork.com/module/uml) into Vercors platform. The integrated editor provides the tools for the specification of the components behavior.
- We have started implementing the behavioural semantics of [36] in the Vercors platform. This task consists in generating the behavior of GCM components in the form of pNets from the GCM architecture defined using VCE. This is an important task, involving intricate engineering issues, but also interesting research on methods for reducing the size of the generated models.

This work was done in collaboration with Rabéa Ameur-Boulifa from Télécom-Paristech and Min Zhang from ECNU Shanghai.

In parallel with core developments of the behavioural specification environment, we further collaborated with our industrial partners and enhanced our work around the use of proof assistants for our specification and verification purposes. In particular, this year:

- We made significant improvements on Mefresa, our Mechanized Framework for the Reasoning on Software Architectures. These were published in [17]. Moreover, we obtained preliminary results regarding its integration with GCM/ProActive, our java middleware for parallel and distributed programming.
- We specified, verified and implemented the HyperManager, a GCM distributed application for the management and monitoring of E-Connectware — a solution for the management of distributed RFID infrastructures. This work was published as an industrial case study in [18].

6.1.4. Autonomic Monitoring and Management of Components

Participants: Françoise Baude, Bastien 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. The framework for autonomic computing allows the designer to describe in a separate way each phase of the MAPE autonomic control loop (Monitoring, Analysis, Planning, and Execution), and to plug them or unplug them dynamically.

- This year, we worked on a journal paper presenting our implementation of GCM component model using active objects, and its use to provide autonomic components. The paper is under revision for SPE journal.

6.1.5. Optimization of data transfer in SOA and EDA models

Participants: Amjad Alshabani, Iyad Alshabani, Françoise Baude, Laurent Pellegrino, Bastien Sauvan, Quirino Zagareze.

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, and allowing the receiver to proceed without only some of them is a meaningful idea that we proved to be effective for active objects in the past [38]. 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.

- We contributed to the offloading of objects representing parameters of the web service Java Apache CXF API [46]. 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 from ASP.
- 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 [26]. 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 EventCloud middleware, see Section 5.5) and its attachements are transferred to the subscriber 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. Also it features very low performance overhead, as additional experiments conducted show it: they are reported in an extended version of the SAC 2013 paper that is under minor revision for a special issue of the Science of Computer Programming journal.

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: Olivier 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. This project is composed of four levels of specification: the two first levels are ready for public realease, but some work is still needed for the development of the validation prototypes.

6.2. Middleware for Grid and Cloud computing

6.2.1. Distributed algorithms for CAN-like P2P networks

Participants: Ludovic Henrio, Fabrice Huet, Justine Rochas.

The nature of some large-scale applications, such as content delivery systems or publish/subscribe systems, built on top of Structured Overlay Networks (SONs), demands application-level dissemination primitives which do not overwhelm the overlay, i.e. which are efficient, and which are also reliable. Building such communication primitives in a reliable manner 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 each other. Our objective is to design and prove an efficient (in terms of messages and execution time) and reliable broadcast algorithm for CAN-like P2P networks. To this aim, this year, we realized the following:

- publication in FASE 2013 of a formalisation, in Isabelle/HOL, of CAN-like P2P networks [15]. Thank to this work, we proved that there exist a broadcast algorithm that does not produce any duplicated message in those networks. A first naive algorithm was exhibited to prove it.
- design and and publication of an optimal broadcast algorithm for CAN-like P2P networks in OPODIS 2013 [20]. The solution we have proposed is proven to be correct, optimal in terms of number of messages, and also efficient, as it provides a good parallelization during the dissemination.

We are also investigating new algorithms to efficiently build a SON when the peer involved already have data. Most of the work on SONs assume that new peers joining the network will arrive without data and thus get assigned a random position. However, if they already have data, they will have to send them to other peers, depending on the key space they are responsible of. In 2013, we continued on the tracks investigated 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.
- We have conducted preliminary experiments which shows a reduction of data transfer between 20% and 90%.

6.2.2. Open Virtual Machines Placement Algorithms

Participants: Fabien Hermenier, Vincent Kherbache, Huynh Tu Dang.

Clients of IaaS providers are looking for dependable infrastructures that can cope with their SLA requirements. To stay attractive, a cloud must then rely on a Virtual Machine (VM) placement algorithm with features matching wrt the SLAs expectations. These constraints are however very specific to each of the tenants but also the infrastructure. They also cover a large range of concerns (reliability, performance, security, energy-efficiency, ...) that are continuously evolving according to new trends and new technologies. To address these issue, we advocate for a flexible VM placement algorithm that can be specialized through plugins to address new concerns.

This year, we first validate our approach with BtrPlace, a composable VM placement algorithm built over Constraint Programming [9]. The usage of Constraint Programming makes placement constraints independent of each other. New constraints can be added without changing the existing implementation. The expressivity of BtrPlace has been verified by implementing more than 20 placement constraints that reproduce, extend but also bring new meaningful restrictions on the VM placement with regards to constraints available in commercial placement algorithm. Each constraint was implemented by an average of 30 lines of Java code. An experienced developer implemented some of the them in half a day, while external developers, without any background in CP, have implemented constraints related to power efficiency [43].

Secondly, we exhibited a lack of reliability in the common approach to address placement constraints in some algorithms. Usually, a constraint controls the VM placement only at the end of the reconfiguration process and ignores the datacenter intermediary states between the beginning and the end of the reconfiguration process. In [11], we advocated that this discrete approach is not sufficient to satisfy the SLAs continuously as an uncontrolled actions schedule may indeed lead to temporary violations. We relied on the flexibility provided by BtrPlace to exhibit these violations and to propose *continuous constraints* to control the quality of service at any moment. We implemented preliminary version of continuous constraints and confirmed they improve the datacenter reliability by removing any temporary violations.

6.2.3. GPU-based High Performance Cloud Computing

Participants: Michael Benguigui, Françoise Baude, Fabrice Huet.

To address HPC, GPU devices are now considered as unavoidable cheap, energy efficient, and very efficient alternative computing units. 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 was to gain some insight on how to efficiently program a non trivial but well known algorithm. Our previous work [40] highlights the necessity to target a GPU rather than distributed CPUs to provide the same performance level. By this way we price complex American basket options through the Picazo pricing algorithm, in the same order of time than a CPU cluster implementation on a 64-core cluster. This year, we achieved the following tasks

- We proposed a multi GPU based implementation of this method, allowing pricing time to fall below 1 hour on 18 GPUs, for a 40-assets American option [14].
- We are currently designing a task dispatching model to load balance tasks in a CPU-GPU cluster. This will allow us to drastically lower the overall computation time of a portfolio estimation, and moreover, the computation time of the Monte Carlo value at risk of a portfolio of complex assets.

6.2.4. MapReduce Based Frameworks for Big Data

Participants: Fabrice Huet, Ge Song.

MapReduce is a programming model which allows the processing of vast amounts of data in parallel, on a large number of machines. It is particularly well suited to static or slow changing set of data since the execution time of a job is usually high. However, in practice data-centers collect data at fast rates which makes it very difficult to maintain up-to-date results. To address this challenge, we propose in [25] a generic mechanism for dealing with dynamic data in MapReduce frameworks. Long-standing MapReduce jobs, called *continuous Jobs*, are *automatically* re-executed to process new incoming data at a minimum cost. We present a simple and clean API which integrates nicely with the standard MapReduce model. Furthermore, we describe cHadoop, an implementation of our approach based on Hadoop which does not require modifications to the source code of the original framework. Thus, cHadoop can quickly be ported to any new version of Hadoop. We evaluate our proposal with two standard MapReduce applications (WordCount and WordCount-N-Count), and one real world application (RDF Query) on real datasets. Our evaluations on clusters ranging from 5 to 40 nodes demonstrate the benefit of our approach in terms of execution time and ease of use.

Another important point is the difficulty to predict the performance of a MapReduce job. This is particularly important when using pay-as-you-go resources such a Cloud. We have proposed a simple framework to predict the performance of Hadoop jobs. It is composed of a dynamic light-weight Hadoop job analyzer, and a prediction module using locally weighted regression methods. Our framework makes some theoretical cost models more practical, and also fits well with jobs and clusters diversity. It can also help those users who want to predict the cost when applying for an on- demand cloud service.

6.3. Application Domains

6.3.1. Publish-Subscribe in Distributed Environments

Participants: Françoise Baude, Fabrice Huet, Laurent Pellegrino, Bastien Sauvan, Iyad Alshabani, Maeva Antoine, Amjad Alshabani.

In the context of the FP7 STREP PLAY and French SocEDA ANR research projects we have developed a middleware dubbed EventCloud (Section 5.5). This last aims to store and retrieve Resource Description Framework (RDF) data but also to relay them to interested parties through a publish/subscribe layer that allows the formulation of content-based subscriptions. Content-based subscriptions are automatically deduced from more complex rules deployed onto a Complex Event Processing engine, the aim of these CEP rules being to trigger new (complex) events after detecting interesting situations [24]. The EventCloud architecture relies on a CAN structured P2P overlay network we initially designed and implemented for the former SOA4ALL FP7-IP project [44].

This year we continued to improve the performances of the EventCloud middleware and its usability as a standalone component but also as a component integrated within the previous projects' platform. Concretely, we proposed a new publish/subscribe matching algorithm for RDF events made of several related RDF triples, that was thoroughly presented in [31] and [22]. To further improve performance, we pursue some efforts to finalize the usage of the newest multi-active object library (cf. Section 6.1.1). Also, to handle more efficiently multicast messaging, we replaced our initial and naive solution with the optimal one presented in Section 6.2.1 . Finally, we proposed a solution for managing multiple EventCloud instances on various cloud platforms, especially for the integration of our middleware in the PLAY and SocEDA platforms (whose latest assessment can be found in [32]). Details about EventCloud management are provided in [29].

Since RDF resources have the property to be poorly balanced, we are also investigating new algorithms that decrease load imbalance for events and data.

6.3.2. Large-scale Simulation Platform: Techniques and methodologies

Participants: Olivier Dalle, E. Mancini, Damian Vicino.

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

Distributed Network Simulation NetStep[16], is a prototype we developped for the distributed simulation of very large scale network simulations, such as the simulation of peer-to-peer applications. We use simulation micro-steps as a means for optimizing the overlap of communications and computations, without changing the original event-driven model. As a consequence, NetStep allows for the reuse of unmodified existing sequential simulators for building large-scale distributed simulations: the overall simulation is divided both in time and space, into a large number of simulation micro-steps, each of which being executed by a legacy sequential simulator. By choosing the time-step smaller than the minimal look-ahead due to communications, we avoid the need for synchronization between logical processes (LPs) during the simulation. Instead, the simulated communications become inputs and outputs of the simulation micro-steps, and are routed in parallel between LPs by a NetStep dedicated entity. Our prototype is based on the SimGrid sequential simulator.

Discrete Time Representation The representation of time in simulations is a long standing issue, for which many solutions and formalisms have been proposed. However, once the formalism is chosen, the implementation of the time representation is still a non trivial problem: Integer values have a limited range and require the selection of a minimal fixed step that does not support well the multi-scale models; Floating Points numbers have numerous limitations and hidden effects such as rounding due to quantization; those issues result in inaccuracies or even timing errors. In collaboration with our partner in the DISSIMINET Associated Team, we have started a new research on this topic. This research will be released in the form of a new Discrete Event Simulation engine library for the DEVS formalism, designed to fully exploit the 2011 C++ standard; it is candidate for inclusion in the BoostC++ Libraries.

PHOENIX Project-Team

6. New Results

6.1. Design-Driven Development Methodology for Resilient Computing

Critical systems have to face changes, either to meet new user requirements or because of changes in the execution context. Often, these changes are made at runtime because the system is too critical to be stopped. Such systems are called *resilient systems*. They have to guarantee dependability despite runtime evolution. For example, in the domain of pervasive computing, building management systems (*e.g.,* anti-intrusion, fire protection system, access control) have to be resilient as they are in charge of people safety and have to run in a continuous way.

To mitigate faults at runtime, dependable systems are augmented with fault tolerance mechanisms such as replication techniques or degraded modes of operation. However, these mechanisms cover a large spectrum of areas ranging from hardware to distributed platforms, to software components. As a consequence, the need of fault-tolerance expertise is spread throughout the software development process, making it difficult to trace the dependability requirements. The fault tolerance mechanisms have to be systematically and rigorously applied in order to guarantee the conformance between the application runtime behavior and the dependability requirements. This integration becomes even more complex when taking into account runtime adaptation. Indeed, a change in the execution context of an application may require to adapt the fault tolerance mechanisms. For example, a decrease of the network bandwidth may require to change the replication mechanism for one requiring less network bandwidth (*e.g.,* Leader-Follower Replication instead of Primary-Backup Replication).

Without a clear separation of the functional and fault-tolerance concerns, ensuring dependability becomes a daunting task for programmers, whose outcome is unpredictable. In this context, design-driven development approaches are of paramount importance because the design drives the development of the application while ensuring the traceability of the dependability requirements. 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 development process of a resilient application in a systematic manner.

In this work, we propose a novel approach that relies on a design language which is extended with fault-tolerance declarations. To further raise the level of abstraction, our development approach revolves around the Sense-Compute-Control paradigm. The design is then compiled into a customized programming framework that drives and supports the development of the application. To face up changes in the execution context, our development methodology relies on a component-based approach, allowing fine-grained runtime adaptation. This design-driven development approach ensures the traceability of the dependability requirements and preserves the separation of concerns, despite runtime evolution.

This work was funded by the Inria collaboration program (in french, actions de recherches collaboratives). The Serus ARC includes the Phoenix project-team (Bordeaux), the ADAM project-team (Lille) and the TSF-LAAS group (Toulouse). These accomplishments were part of Quentin Enard's PhD studies [14]. This work has been published at the International Conference on Component-based Software Engineering (CBSE'13) [23].

6.2. A Case for Human-Driven Software Development

Human-Computer Interaction (HCI) defines a range of principles and methodologies to design User Interfaces (UIs), aiming (1) to improve the interaction between users and computers, (2) to address how interfaces are implemented, leveraging techniques such as program generation and component architectures, and (3) to propose methods to evaluate and compare interfaces.

Despite the many successes of HCI, when it comes to software development, this domain expertise often does not go beyond guidelines (*e.g.,* ISO/TR 22411:2008 addressing the needs of the elderly and users with disabilities). Sometimes, guidelines are mapped into UI design artifacts. However, for a lack of tools, these artifacts remain contemplative. As a consequence, there exists a gap between UI design and software development. This gap is not typical of the HCI domain. Yet, its consequences are dramatically increasing in importance as software systems intertwine with our daily activities, both professional and domestic. Nowadays, a host of systems are playing a critical role for users in terms of safety, privacy, *etc*.

To bridge the gap between UI design and software development, our approach consists in making UI design a full-fledged dimension of software design. We introduce a language dedicated to designing UIs in a high-level manner, while capturing the key requirements of user interaction. We go beyond a contemplative approach and process a UI design artifact to produce a dedicated programming framework that supports the implementation of all the dimensions expressed in a design artifact. This programming framework guides the stakeholders during the development process, while ensuring the conformance between the UI design and its implementation over time.

This work has been published at the International Conference on Software Engineering (ICSE'13, NIER track) [21].

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

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

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

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

This work is in the context of the School+ national research project funded by the French Ministry of National Education. This work is part of Charles Fage's PhD studies.

RMOD Project-Team

6. New Results

6.1. Tools for understanding applications: IDEs and Visualization

Performance Evolution Blueprint: Understanding the Impact of Software Evolution on Performance. Understanding the root of a performance drop or improvement requires analyzing different program executions at a fine grain level. Such an analysis involves dedicated profiling and representation techniques. JProfiler and YourKit, two recognized code profilers fail, on both providing adequate metrics and visual representations, conveying a false sense of the performance variation root. We propose performance evolution blueprint, a visual support to precisely compare multiple software executions. Our blueprint is offered by Rizel, a code profiler to efficiently explore performance of a set of benchmarks against multiple software revisions. [31]

Seamless Composition and Reuse of Customizable User Interfaces with Spec Implementing UIs is often a tedious task. To address this, UI Builders have been proposed to support the description of widgets, their location, and their logic. A missing aspect of UI Builders is however the ability to reuse and compose widget logic. In our experience, this leads to a significant amount of duplication in UI code. To address this issue, we built Spec: a UIBuilder for Pharo with a focus on reuse. With Spec, widget properties are defined declaratively and attached to specific classes known as composable classes. A composable class defines its own widget description as well as the model-widget bridge and widget interaction logic. Spec enables seamless reuse of widgets, its use in Pharo 2.0 has cut in half the amount of lines of code of six of its tools, mostly through reuse. This shows that Spec meets its goals of allowing reuse and composition of widget logic. [17]

Pragmatic Visualizations for Roassal: a Florilegium Software analysis and in particular reverse engineering often involves a large amount of structured data. This data should be presented in a meaningful form so that it can be used to improve software artefacts. The software analysis community has produced numerous visual tools to help understand different software elements. However, most of the visualization techniques, when applied to software elements, produce results that are difficult to interpret and comprehend. We present five graph layouts that are both expressive for polymetric views and agnostic to the visualization engine. These layouts favor spatial space reduction while emphasizing on clarity. Our layouts have been implemented in the Roassal visualization engine and are available under the MIT License. [23]

6.2. Software Quality: Bugs and Debuggers

BugMaps-Granger: A Tool for Causality Analysis between Source Code Metrics and Bugs. Despite the increasing number of bug analysis tools for exploring bugs in software systems, there are no tools supporting the investigation of causality relationships between internal quality metrics and bugs. We propose an extension of the BugMaps tool called BugMaps-Granger that allows the analysis of source code properties that caused bugs. For this purpose, we relied on Granger Causality Test to evaluate whether past changes to a given time series of source code metrics can be used to forecast changes in a time series of defects. Our tool extracts source code versions from version control platforms, generates source code metrics and defects time series, computes Granger, and provides interactive visualizations for causal analysis of bugs. We also provide a case study in order to evaluate the tool. [22]

Mining Architectural Patterns Using Association Rules

Software systems usually follow many programming rules prescribed in an architectural model. However, developers frequently violate these rules, introducing architectural drifts in the source code. We present a data mining approach for architecture conformance based on a combination of static and historical software analysis. For this purpose, the proposed approach relies on data mining techniques to extract structural and historical architectural patterns. In addition, we propose a methodology that uses the extracted patterns to detect both absences and divergences in source-code based architectures. We applied the proposed approach in an industrial strength system. As a result we detected 137 architectural violations, with an overall precision of 41.02%. [27]

Heuristics for Discovering Architectural Violations

Software architecture conformance is a key software quality control activity that aims to reveal the progressive gap normally observed between concrete and planned software architecture. We present ArchLint, a lightweight approach for architecture conformance based on a combination of static and historical source code analysis. For this purpose, ArchLint relies on four heuristics for detecting both absences and divergences in source code based architectures. We applied ArchLint in an industrial-strength system and as a result we detected 119 architectural violations, with an overall precision of 46.7% and a recall of 96.2%, for divergences. We also evaluated ArchLint with four open-source systems, used in an independent study on reflexion models. In this second study, ArchLint achieved precision results ranging from 57.1% to 89.4%. [26]

6.3. Software Quality: History and Changes

Representing Code History with Development Environment Events. Modern development environments handle information about the intent of the programmer: for example, they use abstract syntax trees for providing high-level code manipulation such as refactorings; nevertheless, they do not keep track of this information in a way that would simplify code sharing and change understanding. In most Smalltalk systems, source code modifications are immediately registered in a transaction log often called a ChangeSet. Such mechanism has proven reliability, but it has several limitations. We analyse such limitations and describe scenarios and requirements for tracking fine-grained code history with a semantic representation. We want to enrich code sharing with extra information from the IDE, which will help understanding the intention of the changes and let a new generation of tools act in consequence. [24]

Mining System Specific Rules from Change Patterns A significant percentage of warnings reported by tools to detect coding standard violations are false positives. Thus, there are some works dedicated to provide better rules by mining them from source code history, analyzing bug-fixes or changes between system releases. However, software evolves over time, and during development not only bugs are fixed, but also features are added, and code is refactored. In such cases, changes must be consistently applied in source code to avoid maintenance problems. We propose to extract system specific rules by mining systematic changes over source code history, i.e., not just from bug-fixes or system releases, to ensure that changes are consistently applied over source code. We focus on structural changes done to support API modification or evolution with the goal of providing better rules to developers. Also, rules are mined from predefined rule patterns that ensure their quality. In order to assess the precision of such specific rules to detect real violations, we compare them with generic rules provided by tools to detect coding standard violations on four real world systems covering two programming languages. The results show that specific rules are more precise in identifying real violations in source code than generic ones, and thus can complement them. [25]

6.4. Reconciling Dynamic Languages and Isolation

Virtual Smalltalk Images: Model and Applications. Reflective architectures are a powerful solution for code browsing, debugging or in-language process handling. However, these reflective architectures show some limitations in edge cases of self-modification and self-monitoring. Modifying the modifier process or monitoring the monitor process in a reflective system alters the system itself, leading to the impossibility to perform some of those tasks properly. We analyze the problems of reflective architectures in the context of image based object-oriented languages and solve them by providing a first-class representation of an image: a virtualized image. We present Oz, our virtual image solution. In Oz, a virtual image is represented by an object space. Through an object space, an image can manipulate the internal structure and control the execution of other images. An Oz object space allows one to introspect and modify execution information such as processes, contexts, existing classes and objects. We show how Oz solves the edge cases of reflective architectures by adding a third participant, and thus, removing the self modification and self-observation constraints. [30]

Bootstrapping Reflective Systems: The Case of Pharo. Bootstrapping is a technique commonly known by its usage in language definition by the introduction of a compiler written in the same language it compiles. This process is important to understand and modify the definition of a given language using the same language,

taking benefit of the abstractions and expression power it provides. A bootstrap, then, supports the evolution of a language. However, the infrastructure of reflective systems like Smalltalk includes, in addition to a compiler, an environment with several self-references. A reflective system bootstrap should consider all its infrastructural components. We propose a definition of bootstrap for object-oriented reflective systems, we describe the architecture and components it should contain and we analyze the challenges it has to overcome. Finally, we present a reference bootstrap process for a reflective system and Hazelnut, its implementation for bootstrapping the Pharo Smalltalk-inspired system. [15]

Object Graph Isolation with Proxies More and more software systems are now made of multiple collaborating third-party components. Enabling fine-grained control over the communication between components becomes a major requirement. While software isolation has been studied for a long time in operating systems (OS), most programming languages lack support for isolation. In this context we explore the notion of proxy. A proxy is a surrogate for another object that controls access to this object. We are particularly interested in generic proxy implementations based on language-level reflection. We present an analysis that shows how these reflective proxies can propagate a security policy thanks to the transitive wrapping mechanism. We present a prototype implementation that supports transitive wrapping and allows a fine-grained control over an isolated object graph. [33]

6.5. Dynamic Languages: Compilers

Towards a flexible Pharo Compiler The Pharo Smalltalk-inspired language and environment started its development with a codebase that can be traced back to the original Smalltalk-80 release from 1983. Over the last years, Pharo has been used as the basis of many research projects. Often these experiments needed changes related to the compiler infrastructure. However, they did not use the existing compiler and instead implemented their own experimental solutions. This shows that despite being an impressive achievement considering its age of over 35 years, the compiler infrastructure needs to be improved. We identify three problems: (i) The architecture is not reusable, (ii) compiler can not be parametrized and (iii) the mapping between source code and bytecode is overly complex. Solving these problems will not only help researchers to develop new language features, but also the enhanced power of the infrastructure allows many tools and frameworks to be built that are important even for day-to-day development, such as debuggers and code transformation tools. [20]

Gradual Typing for Smalltalk Being able to combine static and dynamic typing within the same language has clear benefits in order to support the evolution of prototypes or scripts into mature robust programs. While being an emblematic dynamic object-oriented language, Smalltalk is lagging behind in this regard. We report on the design, implementation and application of Gradualtalk, a gradually-typed Smalltalk meant to enable incremental typing of existing programs. The main design goal of the type system is to support the features of the Smalltalk language, like metaclasses and blocks, live programming, and to accomodate the programming idioms used in practice. We studied a number of existing projects in order to determine the features to include in the type system. As a result, Gradualtalk is a practical approach to gradual types in Smalltalk, with a novel blend of type system features that accomodate most programming idioms. [13]

TRISKELL Project-Team

6. New Results

6.1. Support for Reverse Engineering and Maintaining Feature Models

Feature Models (FMs) are a popular formalism for modelling and reasoning about commonality and variability of a system. In essence, FMs aim to define a set of valid combinations of features, also called configurations. In [35], we tackle the problem of synthesising an FM from a set of configurations. The main challenge is that numerous candidate FMs can be extracted from the same input configurations, yet only a few of them are meaningful and maintainable. We first characterise the different meanings of FMs and identify the key properties allowing to discriminate between them. We then develop a generic synthesis procedure capable of restituting the intended meanings of FMs based on inferred [72] or user-specified knowledge. Using tool support, we show how the integration of knowledge into FM synthesis can be realized in different practical application scenarios that involve reverse engineering and maintaining FMs.

6.2. Feature Model Extraction from Large Collections of Informal Product Descriptions

Feature Models (FMs) are used extensively in software product line engineering to help generate and validate individual product configurations and to provide support for domain analysis. As FM construction can be tedious and time-consuming, researchers have previously developed techniques for extracting FMs from sets of formally specified individual configurations, or from software requirements specifications for families of existing products. However, such artifacts are often not available. In [44] we present a novel, automated approach for constructing FMs from publicly available product descriptions found in online product repositories and marketing websites such as SoftPedia and CNET. While each individual product description provides only a partial view of features in the domain, a large set of descriptions can provide fairly comprehensive coverage. Our approach utilizes hundreds of partial product descriptions to construct an FM and is described and evaluated against antivirus product descriptions mined from SoftPedia.

6.3. On Product Comparison Matrices and Variability Models

Product comparison matrices (PCMs) provide a convenient way to document the discriminant features of a family of related products and now abound on the internet. Despite their apparent simplicity, the information present in existing PCMs can be very heterogeneous, partial, ambiguous, hard to exploit by users who desire to choose an appropriate product. Variability Models (VMs) can be employed to formulate in a more precise way the semantics of PCMs and enable automated reasoning such as assisted configuration. Yet, the gap between PCMs and VMs should be precisely understood and automated techniques should support the transition between the two. We propose variability patterns that describe PCMs content and conduct an empirical analysis of 300+ PCMs mined from Wikipedia [62], we also identify the limits of existing comparators, configurators and PCMs [67], [62]. Our findings are a first step toward better engineering techniques for maintaining and configuring PCMs.

6.4. Generating Counterexamples of Model-based Software Product Lines: An Exploratory Study

Model-based Software Product Line (MSPL) engineering aims at deriving customized models corresponding to individual products of a family. The design space of an MSPL is extremely complex to manage for the engineer, since the number of variants may be exponential and the derived product models have to conform to numerous well-formedness and business rules. We provide a way to generate MSPLs, called counterexamples, that can produce invalid product models despite a valid configuration in the variability model [49]. We provide a systematic and automated process, based on the Common Variability Language (CVL), to randomly search the space of MSPLs for a specific formalism. We validate the effectiveness of this process for three formalisms at different scales (up to 247 metaclasses and 684 rules).

6.5. Composing your Compositions of Variability Models

Modeling and managing variability is a key activity in a growing number of software engineering contexts. Support for composing variability models is arising in many engineering scenarios, for instance, when several subsystems or modeling artifacts, each coming with their own variability and possibly developed by different stakeholders, should be combined together. We consider in [34] the problem of composing feature models (FMs), a widely used formalism for representing and reasoning about a set of variability choices. We show that several composition operators can actually be defined, depending on both matching/merging strategies and semantic properties expected in the composed FM. We present four alternative forms and their implementations. We discuss their relative trade-offs w.r.t. reasoning, customizability, traceability, composability and quality of the resulting feature diagram. We summarize these findings in a reading grid which is validated by revisiting some relevant existing works. Our contribution should assist developers in choosing and implementing the right composition operators.

6.6. Extraction and Evolution of Architectural Variability Models in Plugin-based Systems

Variability management is a key issue when building and evolving software-intensive systems, making it possible to extend, configure, customize and adapt such systems to customers' needs and specific deployment contexts. A wide form of variability can be found in extensible software systems, typically built on top of plugin-based architectures that offer a (large) number of configuration options through plugins. In an ideal world, a software architect should be able to generate a system variant on-demand, corresponding to a particular assembly of plugins. To this end, the variation points and constraints between architectural elements should be properly modeled and maintained over time (i.e., for each version of an architecture). A crucial, yet error-prone and time-consuming, task for a software architect is to build an accurate representation of the variability of an architecture, in order to prevent unsafe architectural variants and reach the highest possible level of flexibility. In [23], we propose a reverse engineering process for producing a variability model (i.e., a feature model) of a plugin-based architecture. We develop automated techniques to extract and combine different variability descriptions, including a hierarchical software architecture model, a plugin dependency model and the software architect knowledge. By computing and reasoning about differences between versions of architectural feature models, software architect can control both the variability extraction and evolution processes. The proposed approach has been applied to a representative, large-scale plugin-based system (FraSCAti), considering different versions of its architecture. We report on our experience in this context.

6.7. FAMILIAR: A Domain-Specific Language for Large Scale Management of Feature Models

The feature model formalism has become the de facto standard for managing variability in software product lines (SPLs). In practice, developing an SPL can involve modeling a large number of features representing di erent viewpoints, sub-systems or concerns of the software system. This activity is generally tedious and error-prone. In [24], we present FAMILIAR a Domain-Specific Language (DSL) that is dedicated to the large scale management of feature models and that complements existing tool support. The language provides a powerful support for separating concerns in feature modeling, through the provision of composition and decomposition operators, reasoning facilities and scripting capabilities with modularization mechanisms. We illustrate how an SPL consisting of medical imaging services can be practically managed using reusable FAMILIAR scripts that implement reasoning mechanisms. We also report on various usages and applications of FAMILIAR and its operators, to demonstrate their applicability to di erent domains and use for di erent purposes.

6.8. Web Configurators

Nowadays, mass customization has been embraced by a large portion of the industry. As a result, the web abounds with sales configurators that help customers tailor all kinds of goods and services to their specific

needs. In many cases, configurators have become the single entry point for placing customer orders. As such, they are strategic components of companies' information systems and must meet stringent reliability, usability and evolvability requirements. However, the state of the art lacks guidelines and tools for efficiently engineering web sales configurators. To tackle this problem, empirical data on current practice is required. The paper [51] reports on a systematic study of 111 web sales configurators along three essential dimensions: rendering of configuration options, constraint handling, and configuration process support. Based on this, we highlight good and bad practices in engineering web sales configurator. The reported quantitative and qualitative results open avenues for the elaboration of methodologies to (re-)engineer web sales configurators. In [48] we focus on how to associate product configurations to visual representations in a Web configurator. We present a formal statement of the problem and a model-driven perspective.

6.9. Separating Concerns in Feature Models

Feature models (FMs) are a popular formalism to describe the commonality and variability of a set of assets in a software product line (SPL). SPLs usually involve large and complex FMs that describe thousands of features whose legal combinations are governed by many and often complex rules. The size and complexity of these models is partly explained by the large number of concerns considered by SPL practitioners when managing and configuring FMs. In the chapter [68], we first survey concerns and their separation in FMs, highlighting the need for more modular and scalable techniques. We then revisit the concept of view as a simplified representation of an FM. We finally describe a set of techniques to specify, visualize and verify the coverage of a set of views. These techniques are implemented in complementary tools providing practical support for feature-based configuration and large scale management of FMs.

6.10. Bridging the Chasm between Executable Metamodeling and Models of Computation

The complete and executable definition of a Domain Specific Language (DSL) relies on the specification of two essential facets: a model of the domain-specific concepts with actions and their semantics; a scheduling model that orchestrates the actions of a domain-specific model. Metamodels can capture the former facet, while Models of Computation (MoCs) capture the latter facet. Unfortunately, theories and tools for metamodeling and MoCs have evolved independently, creating a cultural and technical chasm between the two communities. We introduce a new framework to bridge a metamodel and a MoC in a modular fashion [43]. This bridge allows (i) the complete and executable definition of a DSL, (ii) the reuse of MoCs for different domain-specific metamodels, and (iii) the use of different MoCs for a given metamodel, to cope with variation points of a DSL.

6.11. Reifying Concurrency for Executable Metamodeling

Current metamodeling techniques can be used to specify the syntax and semantics of domain specific modeling languages (DSMLs). However, there is currently very little support for explicitly specifying concurrency semantics using metamodels. We reify concurrency as a metamodeling facility, leveraging formalization work from the concurrency theory and models of computation (MoC) community [42]. The essential contribution of this paper is a proposed language workbench for binding domain-specific concepts and models of computation through an explicit event structure at the metamodel level. We illustrate these novel metamodeling facilities for designing two variants of a concurrent and timed final state machine, and provide other experiments to validate the scope of our approach.

6.12. Using Model Types to Support Contract-Aware Model Substitutability

Model typing brings the benefit associated with well-defined type systems to model-driven development (MDD) through the assignment of specific types to models. In particular, model type systems enable reuse of model manipulation operations (e.g., model transformations), where manipulations defined for models of a supertype can be used to manipulate models of subtypes. Existing model typing approaches are limited to

structural typing defined in terms of object-oriented metamodels (e.g., MOF) in which the only structural (well-formedness) constraints are those that can be expressed directly in metamodeling notations (e.g., multiplicity and element containment constraints). We propose an extension to model typing that takes into consideration structural invariants, other than those that can be expressed directly in metamodeling notation, and specifications of behaviors associated with model types [64]. The approach supports contract-aware substitutability, where contracts are defined in terms of invariants and pre-/postconditions expressed using OCL. Support for behavioral typing paves the way for behavioral substitutability. We also describe a technique to rigorously reason about model type substitutability as supported by contracts and apply the technique in use cases from the optimizing compiler community.

6.13. Variability Support in Domain-Specific Language Development

Domain Specific Languages (DSLs) are widely adopted to capitalize on business domain experiences. Consequently, DSL development is becoming a recurring activity. Unfortunately, even though it has its benefits, language development is a complex and time-consuming task. Languages are commonly realized from scratch, even when they share some concepts and even though they could share bits of tool support. This cost can be reduced by employing modern modular programming techniques that foster code reuse. However, selecting and composing these modules is often only within the reach of a skilled DSL developer. We propose to combine modular language development and variability management, with the objective of capitalizing on existing assets [63]. This approach explicitly models the dependencies between language components, thereby allowing a domain expert to configure a desired DSL, and automatically derive its implementation. The approach is tool supported, using Neverlang to implement language components, and the Common Variability Language (CVL) for managing the variability and automating the configuration. We illustrate our approach with the help of different case studies, including the implementation of a family of DSLs to describe variants of state machines.

6.14. Automatically Searching for Metamodel Well-Formedness Rules in Examples and Counter-Examples

Current metamodeling formalisms support the definition of a metamodel with two views: classes and relations, that form the core of the metamodel, and well-formedness rules, that constraints the set of valid models. While a safe application of automatic operations on models requires a precise definition of the domain using the two views, most metamodels currently present in repositories have only the first one part. We propose in [47] to start from valid and invalid model examples in order to automatically retrieve well-formedness rules in OCL using Genetic Programming. The approach is evaluated on metamodels for state machines and features diagrams. The experiments aim at demonstrating the feasibility of the approach and at illustrating some important design decisions that must be considered when using this technique.

6.15. Building Modular and Efficient DSLs: Mashup of Meta-Languages and its Implementation in the Kermeta Language Workbench

With the growing use of domain-specific languages (DSL) in industry, DSL design and implementation goes far beyond an activity for a few experts only and becomes a challenging task for thousands of software engineers. DSL implementation indeed requires engineers to care for various concerns, from abstract syntax, static semantics, behavioral semantics, to extra-functional issues such as run-time performance. We propose an approach that uses one meta-language per language implementation concern [27] in the new version (v2) of the Kermeta workbench. We show that the usage and combination of those meta-languages is simple and intuitive enough to deserve the term "mashup". We evaluate the approach by completely implementing the non trivial fUML modeling language, a semantically sound and executable subset of the Unified Modeling Language (UML) ; Kompren, a DSL for designing and implementing model slicers ; and KCVL, the Commun Variability Language dedicated to variability management in software design models [65].

6.16. On the Globalization of Modeling Languages

In the software and systems modeling community, research on domain-specific modeling languages (DSMLs) is focused on providing technologies for developing languages and tools that allow domain experts to develop system solutions efficiently in a particular domain. Unfortunately, the current lack of support for explicitly relating concepts expressed in different DSMLs makes it very difficult for software and system engineers to reason about information spread across models describing different system aspects. Supporting coordinated use of DSMLs leads to what we call the globalization of modeling languages. We present a research initiative that broadens the current DSML research focus beyond the development of independent DSMLs to one that provides support for globalized DSMLs, that is, DSMLs that facilitate coordination of work across different domains of expertise [31]. We explore this new grand challenge in recent workshops, *e.g.,* GlobalDSL'13 at ECSA, ECMFA and ECOOP 2013 [69], and GEMOC'13 at MODELS 2013 [70].

6.17. Automating the Maintenance of Non-functional System Properties using Demonstration-based Model Transformation

Given a base model with functional components, maintaining the non-functional properties that crosscut the base model has become an essential modeling task when using DSMLs. We present a demonstration-based approach to automate the maintenance of non-functional properties in DSMLs [29]. Instead of writing model transformation rules explicitly, users demonstrate how to apply the non-functional properties by directly editing the concrete model instances and simulating a single case of the maintenance process. An inference engine generates generic model transformation patterns, which can be refined by users and then reused to automate the same evolution and maintenance task in other models. Our demonstration-based approach has been applied to several scenarios, such as auto-scaling and model layout.

6.18. Improving Reusability and Automation in Software Process Lines

Software processes orchestrate manual or automatic tasks to create new software products that meet the requirements of specific projects. While most of the tasks are about inventiveness, modern developments also require recurrent, boring and time-consuming tasks (e.g., the IDE configuration, or the continuous integration setup). Such tasks struggle to be automated due to their various execution contexts according to the requirements of specific projects. We propose a methodology that benefits from an explicit modeling of a family of processes to identify the possible reuse of automated tasks in software processes [60]. Then, we propose a tool-supported approach that integrates both reuse and automation [61]. It consists of reusing processes from an SPL according to projects' requirements. The processes are bound to components that automate their execution. When the variability of a process to execute is not fully resolved, our approach consists of resolving this variability during the execution of this process. We illustrate our approach on industrial projects in a software company, as well as on a family of processes for designing and implementing modeling languages. Our approach promoted the identification of possible automated tasks for configuring IDEs and continuous integration, their reuse in various projects of the company, and the automation of their execution, while enabling to resolve process variability during the execution.

6.19. Towards Trust-Aware and Self-Adaptive Systems

The dynamic conditions under which Future Internet (FI) applications must execute call for self-adaptive software to cope with unforeseeable changes in the application environment. Software engineering currently provides frameworks to develop reasoning engines that support the runtime adaptation of distributed, heterogeneous applications. However, these frameworks have very limited support to address security concerns of these application, hindering their usage for FI scenarios. We address this challenge by enhancing self-adaptive systems with the concepts of trust and reputation [58]. Trust improves decision-making processes under risk and uncertainty, in turn improving security of self-adaptive FI applications.

6.20. SOA Antipatterns: an Approach for their Specification and Detection

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 in their architecture, called antipatterns. We introduce a novel and innovative approach supported by a framework for specifying and detecting antipatterns in SBSs [25]. We specify ten well-known and common antipatterns, including Multi Service and Tiny Service, and automatically generate their detection algorithms. We validate the detection algorithms in terms of precision and recall on two systems developed independently. This validation demonstrates that our approach enables the specification and detection of SOA antipatterns with an average precision of 90% and recall of 97.5%.

6.21. Automated Measurement of Models of Requirements

One way to formalize system requirements is to express them using the object-oriented paradigm. In this case, the class model representing the structure of requirements is called a requirements metamodel, and requirements themselves are object-based models of natural-language requirements. We show that such object-oriented requirements are well-suited to support a large class of requirements metrics[28]. We define a requirements metamodel and use an automated measurement approach proposed in our previous work to specify requirements metrics. We show that it is possible to integrate 78 metrics from 11 different papers in the proposed framework. The software that computes the requirements metric values is fully generated from the specification of metrics.

6.22. Empirical Evidence of Large-Scale Diversity in API Usage of Object-Oriented Software

In this paper, we study how object-oriented classes are used across thousands of software packages. We concentrate on "usage diversity'", defined as the different statically observable combinations of methods called on the same object. We present empirical evidence that there is a significant usage diversity for many classes. For instance, we observe in our dataset that Java's String is used in 2460 manners. We discuss the reasons of this observed diversity and the consequences on software engineering knowledge and research [56].

6.23. Efficient high-level abstractions for web programming

Writing large Web applications is known to be difficult. One challenge comes from the fact that the application's logic is scattered into heterogeneous clients and servers, making it difficult to share code between both sides or to move code from one side to the other. Another challenge is performance: while Web applications rely on ever more code on the client-side, they may run on smart phones with limited hardware capabilities. These two challenges raise the following problem: how to benefit from high-level languages and libraries making code complexity easier to manage and abstracting over the clients and servers differences without trading this ease of engineering for performance? In [59], we present high-level abstractions defined as deep embedded DSLs in Scala that can generate efficient code leveraging the characteristics of both client and server environments. We compare performance on client-side against other candidate technologies and against hand written low-level JavaScript code. Though code written with our DSL has a high level of abstraction, our benchmark on a real world application reports that it runs as fast as hand tuned low-level JavaScript code.

6.24. Exploring Optimal Service Compositions in Highly Heterogeneous and Dynamic Service-Based Systems

Service-oriented pervasive systems, composed of a large number of devices with heterogeneous capabilities where devices' resources are abstracted as software services, challenge the creation of high-quality composite applications. Resource heterogeneity, dynamic network connectivity, and a large number of highly distributed service providers complicate the process of creating applications with specific QoS requirements. Existing approaches to service composition control the QoS of an application solely by changing the set of participating

concrete services which is not suitable for ad-hoc service-based systems characterised by high intermittent connectivity and resource heterogeneity. In [46], we propose a flexible way of formulating composition configurations suitable for such service-based systems. Our formulation proposes the combined consideration of the following factors that affect the QoS of a composed service: (a) service selection, (b) orchestration partitioning, and (c) orchestrator node selection. We show that the proposed formulation enables the definition of service composition configurations with 49% lower response time, 28% lower network latency, 36% lower energy consumption, and 13% higher success ratio compared to those defined with the traditional approach. In [45], we present the problem of efficiently exploring at runtime the serach space of possible configurations for a service orchestration with various Quality of Services.

COATI Project-Team

6. New Results

6.1. Network Design and Management

Participants: Julio Araújo, Jean-Claude Bermond, Luca Chiaraviglio, David Coudert, Frédéric Giroire, Alvinice Kodjo, Aurélien Lancin, Remigiusz Modrzejewski, Christelle Molle-Caillouet, Joanna Moulierac, Nicolas Nisse, Stéphane Pérennes, Truong Khoa Phan, Ronan Pardo Soares, Issam Tahiri.

6.1.1. Optimization in backbone networks

6.1.1.1. Shared Risk Link Group

The notion of Shared Risk Link Groups (SRLG) has been introduced to capture survivability issues where some links of a network fail simultaneously. In this context, the diverse routing problem is to find a set of pairwise SRLG-disjoint paths between a given pair of end nodes of the network. This problem has been proved NP-complete in general and some polynomial instances have been characterized.

In [33], [32], we investigate the diverse-routing problem in networks where the SRLGs are localized and satisfy the *star property*. This property states that a link may be subject to several SRLGs, but all links subject to a given SRLG are incident to a common node. We first provide counterexamples to the polynomial-time algorithm proposed in the literature for computing a pair of SRLG-disjoint paths in networks with SRLGs satisfying the star property, and then prove that this problem is in fact NP-complete. We have also characterized instances that can be solved in polynomial time or are fixed parameter tractable, in particular when the number of SRLGs is constant, the maximum degree of the vertices is at most 4, and when the network is a directed acyclic graph. Moreover, we have considered the problem of finding the maximum number of SRLG-disjoint paths in networks with SRLGs satisfying the star property. We have proved that such problem is NP-hard and hard to approximate. Then, we have provided exact and approximation algorithms for relevant subcases.

6.1.1.2. Wavelength assignment in WDM networks

Let $\mathcal{P}$ be a family of directed paths in a directed graph G. The load of an arc is the number of directed paths containing this arc. Let $\pi(G,\mathcal{P})$ be the maximum of the load of all the arcs and let $w(G,\mathcal{P})$ be the minimum number of wavelengths (colours) needed to colour $\mathcal{P}$ in such a way that two directed paths with the same wavelength are arc-disjoint. These two parameters correspond respectively to the clique number and the chromatic number of the associated conflict graph, and $\pi(G,\mathcal{P}) \leq w(G,\mathcal{P})$. It was known that there exists directed acyclic graphs (DAGs) such that the ratio between $w(G,\mathcal{P})$ and $w(G,\mathcal{P})$ is arbitrarily large. In [18], solving a conjecture of an earlier article, we show that the same is true for a very restricted class of DAGs, the UPP-DAGs, those for which there is at most one directed path from a vertex to another. We also characterized the DAGs such that $\pi(G,\mathcal{P}) = w(G,\mathcal{P})$ for all families of directed paths.

6.1.1.3. Mutli-operators microwave backhaul networks

In [35], we consider the problem of sharing the infrastructure of a backhaul network for routing. We investigate on the revenue maximization problem for the physical network operator (PNO) when subject to stochastic traffic requirements of multiple virtual network operators (VNO) and prescribed service level agreements (SLA). We use robust optimization to study the tradeoff between revenue maximization and the allowed level of uncertainty in the traffic demands. This mixed integer linear programming model takes into account end-to-end traffic delays as example of quality-of-service requirement in a SLA. To show the effectiveness of our model, we present a study on the price of robustness, i.e. the additional price to pay in order to obtain a feasible solution for the robust scheme, on realistic scenarios.

6.1.2. Energy efficiency

With one third of the world population online in 2013 and an international Internet bandwidth multiplied by more than eight since 2006, the ICT sector is a non-negligible contributor of worldwide greenhouse gases emissions and power consumption. Indeed, power consumption of telecommunication networks has become a major concern for all the actors of the domain, and efforts are made to reduce their impact on the overall figure of ICTs, and to support its foreseen growth in a sustainable way. In this context, the contributors of the European Network of Excellence TREND have developed innovative solutions to improve the energy efficiency of optical networks summarized in [45].

6.1.2.1. Energy aware routing with redundancy elimination

Many studies have shown that energy-aware routing (EAR) can significantly reduce energy consumption of a backbone network. Redundancy Elimination (RE) techniques provide a complementary approach to reduce the amount of traffic in the network. In particular, the GreenRE model combines both techniques, offering potentially significant energy savings.

In [44], we enhance the MIP formulation proposed in [75] for the GreenRE model. We derive cutting planes, extending the well-known cutset inequalities, and report on preliminary computations.

In [37], we propose a concept for respecting uncertain rates of redundant traffic within the GreenRE model, closing the gap between theoretical modeling and drawn-from-life data. To model redundancy rate uncertainty, the robust optimization approach in [73] is adapted and the problem is formally defined as mixed integer linear program. An exemplary evaluation of this concept with real-life traffic traces and estimated fluctuations of data redundancy shows that this closer-to-reality model potentially offers significant energy savings in comparison to GreenRE and EAR.

6.1.2.2. Energy Efficient Content Distribution

The basic protocols of the Internet are point-to-point in nature. However, the traffic is largely broadcasting, with projections stating that as much as 80-90% of it will be video by 2016. This discrepancy leads to an inefficiency, where multiple copies of essentially the same messages travel in parallel through the same links. We have studied approaches to mitigate this inefficiency and reduce the energy consumption of future networks, in particular in [13].

In [29], we study the problem of reducing power consumption in an Internet Service Provider (ISP) network by designing the content distribution infrastructure managed by the operator. We propose an algorithm to optimally decide where to cache the content inside the ISP network. We evaluate our solution over two case studies driven by operators feedback.

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 [57], [46], [29], we study the impact of using in-network caches and content delivery network (CDN) cooperation on an energy-efficient routing. Experimental results show that by placing a cache on each backbone router to store the most popular content, along with well choosing the best content provider server for each demand to a CDN, we can save up to 23% of power in the backbone.

6.1.3. Distributed systems

6.1.3.1. Distributed Storage systems.

In a P2P storage system using erasure codes, a data block is encoded in many redundancy fragments. These fragments are then sent to distinct peers of the network. In [24], we study the impact of different placement policies of these fragments on the performance of storage systems.

In [39], we propose a new analytical framework that takes into account the correlation between data reconstructions when estimating the repair time and the probability of data loss. The models and schemes proposed are validated by mathematical analysis, extensive set of simulations, and experimentation using the GRID5000 test-bed platform. This new model allows system designers to operate a more accurate choice of system parameters in function of their targeted data durability.

6.1.3.2. P2P Streaming systems

In [41], [68], we propose and analyze a simple localized algorithm to balance a tree. The motivation comes from live distributed streaming systems in which a source diffuses a content to peers via a tree, a node forwarding the data to its children. Such systems are subject to a high churn, peers frequently joining and leaving the system. It is thus crucial to be able to repair the diffusion tree to allow an efficient data distribution. In particular, due to bandwidth limitations, an efficient diffusion tree must ensure that node degrees are bounded. Moreover, to minimize the delay of the streaming, the depth of the diffusion tree must also be controlled. We propose here a simple distributed repair algorithm in which each node carries out local operations based on its degree and on the subtree sizes of its children.

6.1.4. 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.

In [19], we give an algorithm to construct minimum makespan transmission schedules for data gathering under the following hypotheses: the communication graph G is a tree network, and no buffering is allowed at intermediate nodes and $d_I \geq 2$. In the interesting case in which all nodes in the network have to deliver an arbitrary positive number of packets, we provide a closed formula for the makespan of the optimal gathering schedule. Additionally, we consider the problem of determining the computational complexity of data gathering in general graphs and show that the problem is NP–complete. On the positive side, we design a simple $(1 + 2/d_I)$-factor approximation algorithm for general networks.

In [59], we focus on the gathering and personalized broadcasting problem in grids. We still consider the non-buffering model. In this setting, though the problem of determining the complexity of computing the optimal makespan in a grid is still open, we present linear (in the number of messages) algorithms that compute schedules for gathering with $d_I = 0, 1, 2$. In particular, we present an algorithm that achieves the optimal makespan up to a small additive constant. Note that, the approximation algorithms that we present also provide approximation up to a ratio 2 for the gathering with buffering. All our results are proved in terms of personalized broadcasting.

In [20], we now allow transmission till a distance d_T and buffering in intermediate nodes. We focus on the specific case where the network is a path with the sink at an end vertex of the path and where the traffic is unitary ($w(u) = 1$ for all u); indeed this simple case appears to be already very difficult. We first give a new lower bound and a protocol with a gathering time that differs only by a constant independent from the length of the path. Then we present a method to construct incremental protocols which are optimal for many values of d_T and d_I (in particular when d_T is prime).

In [50], we focus on gathering uncertain traffic demands in mesh networks with multiple sources and sinks. The scheduling is relaxed into the round weighting problem in which a set of pairwise non-interfering links is called a round, and we seek to successively activate rounds in order to get enough capacity on links to route the demand from the set of sources to the set of sinks. We propose a new robust model considering traffic demand uncertainty, efficiently solved by column generation, and quantify the price of robustness, i.e., the additional cost to pay in order to obtain a feasible solution for the robust scheme.

6.1.5. Routing

6.1.5.1. Routing models evaluation

The Autonomous System (AS)-level topology of the Internet that currently comprises more than 40k ASs, is growing at a rate of about 10% per year. In these conditions, Border Gateway Protocol (BGP), the inter-domain routing protocol of the Internet starts to show its limits, among others in terms of the number of routing table entries it can dynamically process and control. To overcome this challenging situation, the design but also the evaluation of alternative dynamic routing models and their comparison with BGP will be performed by means of simulation. However, existing routing models simulators such as DRMSim, the Dynamic Routing Model

Simulator developped in COATI in collaboration with Alcatel-Lucent [72], are limited in terms of the number of routing table entries they can dynamically process and control on a single computer.

In [63], we have conducted a feasibility study of the extension of DRMSim so as to support the Distributed Parallel Discrete Event paradigm. We have studied several distribution models and their associated communication overhead. We have in particular evaluated the expected additional time required by a distributed simulation of BGP (border gate protocol) on topologies with 100k ASes compared to its sequential simulation. We show that such a distributed simulation of BGP is possible with a reasonable time overhead.

6.1.5.2. Complexity of Shortest Path Routing

In telecommunication networks packets are carried from a source s to a destination t on a path that is determined by the underlying routing protocol. Most routing protocols belong to the class of shortest-path routing protocols. For better protection and efficiency, one wishes to use multiple (shortest) paths between two nodes. Therefore the routing protocol must determine how the traffic from s to t is distributed among the shortest paths. In the protocol called OSPF-ECMP (for Open Shortest Path First-Equal Cost Multiple Path) the traffic incoming at every node is uniformly balanced on all outgoing links that are on shortest paths. In [43], [42], we show that the problem of maximizing even a single commodity flow for the OSPF-ECMP protocol cannot be approximated within any constant factor ratio. Besides this main theorem, we derive some positive results which include polynomial-time approximations and an exponential-time exact algorithm.

6.2. Graph Algorithms

Participants: Julio Araújo, Jean-Claude Bermond, David Coudert, Frédéric Havet, Frédéric Giroire, Bi Li, Fatima Zahra Moataz, Christelle Molle-Caillouet, Nicolas Nisse, Ronan Pardo Soares, Stéphane Pérennes.

COATI is also interested in the algorithmic aspects of Graph Theory. In general we try to find the most efficient algorithms to solve various problems of Graph Theory and telecommunication networks. More information on several results presented in this section may be found in R. Soares's thesis [14].

6.2.1. Complexity and Computation of Graph Parameters

We use graph theory to model various network problems. In general we study their complexity and then we investigate the structural properties of graphs that make these problems hard or easy. In particular, we try to find the most efficient algorithms to solve the problems, sometimes focusing on specific graph classes from which the problems are polynomial-time solvable.

6.2.1.1. Parameterized Complexity

Parameterized complexity is a way to deal with intractable computational problems having some parameters that can be relatively small with respect to the input size. This area has been developed extensively during the last decade. More precisely, we consider problems that consist in deciding whether a graph G satisfies some property (i.e., if G belongs to some given family of graphs). For decision problems with input size n and parameter k, the goal is to design an algorithm with running time $f(k).n$, where f depends only on k. Problems for which we can find an optimal algorithm with such time complexity are said to be fixed-parameter tractable (FPT). Equivalently, the goal is to design a polynomial-time algorithm (in k and n) that computes a pair (H, k') where H is a graph (the kernel) with size polynomial in k and $P(G) \leq k$ if and only if $P(H) \leq k'$.

We study the parameterized complexity of the edge-modification problems. Given a graph $G = (V, E)$ and a positive integer k, an edge modification problem for a graph property Π consists in deciding whether there exists a set F of pairs of V of size at most k such that the graph $H = (V, E \Delta F)$ satisfies the property Π. In [25], it is proved that parameterized cograph edge-modification problems have cubic vertex kernels whereas polynomial kernels are unlikely to exist for the P_l-free edge-deletion and the C_l-free edge-deletion problems for $l \geq 7$ and $l \geq 4$ respectively.

We also design a unified parameterized algorithm for computing various widths of graphs (such as branched tree-width, branch-width, cut-width, etc.) [60].

6.2.1.2. Convexity in Graphs

The geodesic convexity of graphs naturally extends the notion of convexity in euclidean metric spaces. A set S of vertices of a graph $G = (V, E)$ is *convex* if any vertex on a shortest path between two vertices of S also belongs to S. The *convex hull* of $S \subset V$ is the smallest convex set containing S. Finally, a *hull set* of a graph is a set of vertices whose convex hull is V. The hull number of a graph G is the minimum size of a hull set in G. In [16] , we prove that computing the hull number is NP-complete in bipartite graphs. We also provide bounds and design various polynomial-time algorithms for this problem in different graph classes such as co-bipartite graphs, P_4-sparse graphs, etc. In [30], we first 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 is the size of a vertex cover of G or, more generally, its neighborhood diversity. We also use these reductions to characterize the hull number of the lexicographic product of any two graphs.

6.2.1.3. Hyperbolicity

The Gromov hyperbolicity is an important parameter for analyzing complex networks since it expresses how the metric structure of a network looks like a tree. In other words, it provides bounds on the stretch resulting from the embedding of a network topology into a weighted tree. It is therefore used to provide bounds on the expected stretch of greedy-routing algorithms in Internet-like graphs. However, the best known algorithm for computing this parameter has time complexity in $O(n^{3.69})$, which is prohibitive for large-scale graphs. In [36], we proposed a novel algorithm for determining the hyperbolicity of a graph that is scalable for large graphs. The time complexity of this algorithm is output-sensitive and depends on the shortest-path distances distribution in the graph and on the computed value of the hyperbolicity. Although its worst case time complexity is in $O(n^4)$, it is in practice much faster than previous proposals as it uses bounds to cut the search space. This algorithm allowed us for computing the hyperbolicity of all maps of the Internet provided by CAIDA and DIMES.

6.2.2. Graph searching and applications

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. We investigated several variants of these games.

6.2.2.1. Variants of graph searching.

We study non-deterministic graph searching where the searchers have to capture an invisible fugitive but can see him a bounded number of times. This variant generalizes the notion of pathwidth and treewidth of graphs. In this setting, we provide a polynomial-time algorithm that approximates the minimum number of searchers needed in trees, up to a factor of two [56].

In [34], [61], we define another variant of graph searching, where searchers have to capture an invisible fugitive with the constraint that no two searchers can occupy the same node simultaneously. This variant seems promising for designing approximation algorithms for computing the pathwidth of graphs. The main contribution in [34], [61] is the characterization of trees where k searchers are necessary and sufficient to win. Our characterization leads to a polynomial-time algorithm to compute the minimum number of searchers needed in trees.

We also study graph searching in directed graphs. We prove that the graph processing variant is monotone which allows us to show its equivalence with a particular digraph decomposition [47].

6.2.2.2. Surveillance Game and Fractional Game.

A surprising application of some variant of pursuit-evasion games is the problem for a web-browser to download documents in advance while an internaut is surfing on the Web. In a previous work, we model this problem as a Pursuit-evasion game called Surveillance game. In [40], [67], we continue our study of the Surveillance game. We provide some bounds on the connected and online variants of this game. In particular,

we show that, in the online variant (when the searchers discover the graph during the game), the best strategy is the trivial one that consists in downloading the document in the neighborhood of the position of the internaut.

In [69], [48], [52], we define a framework generalizing and relaxing many games (including the Surveillance game) where Players use fractions of their token at each turn. We design an algorithm for solving the fractional games. In particular, our algorithm runs in polynomial-time when the length of the game is bounded by 2 (in contrast, computing the surveillance game is NP-hard even when the game is limited to two turns). For some games, we also prove that the fractional variant provides some good approximation. This direction of research seems promising for solving many open problems related to Pursuit-evasion games.

6.2.2.3. 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. We proposed a general approach [38], [66] to solve the three problems in rings even in case of symmetric initial configurations.

6.2.3. Algorithm design in biology

In COATI, we have recently started a collaboration with EPI ABS (Algorithms Biology Structure) from Sophia Antipolis on "minimal connectivity complexes in mass spectrometry based macro-molecular complex reconstruction" [28], [55]. This problem turns out to be a minimum color covering problem (minimum number of colors to cover colored edges with connectivity constraints on the subgraphs induced by the colors) of the edges of a graph, and is surprizingly similar to a capacity maximization problem in a multi-interfaces radio network we were studying.

6.3. Structural Graph Theory

Participants: Julio Araújo, Jean-Claude Bermond, Frédéric Havet, Nicolas Nisse, Ana Karolinna Maia de Oliveira, Stéphane Pérennes.

6.3.1. Graph colouring and applications

Graph colouring is a central problem in graph theory and it has a huge number of applications in various scientific domains (telecommunications, scheduling, bio-informatics, ...). We mainly study graph colouring problems that model ressource allocation problems.

6.3.1.1. Backbone colouring

A well-known channel assignment problem is the following: we are given a graph G, whose vertices correspond to transmitters, together with an edge-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 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 [22], we prove that if G is planar and T is a tree of diameter at most 4, then $BBC_2(G,T) \leq 6$ hence giving an evidence to a conjecture of Broersma et al. [74] stating that the same holds if T has an arbitrary diameter.

6.3.1.2. Weighted colouring

We also studied weighted colouring which models various problems of shared resources allocation. Given a vertex-weighted graph G and a (proper) r -colouring $c = \{C_1, ..., C_r\}$ of G , the *weight* of a colour class C_i is the maximum weight of a vertex coloured i and the *weight* of c is the sum of the weights of its colour classes. The objective of the Weighted Colouring Problem is, given a vertex-weighted graph G , to determine the minimum weight of a proper colouring of G, that is, its *weighted chromatic number*. In [17], we prove that the Weighted Colouring Problem admits a version of Hajós' Theorem and so we show a necessary and sufficient condition for the weighted chromatic number of a vertex-weighted graph G to be at least k, for any positive real k. The Weighted Colouring Problem problem remains NP-complete in some particular graph classes as bipartite graphs. In their seminal paper [77], Guan and Zhu asked whether the weighted chromatic number of bounded tree-width graphs (partial k-trees) can be computed in polynomial-time. Surprisingly, the time-complexity of computing this parameter in trees is still open. We show [58] that, assuming the Exponential Time Hypothesis (3-SAT cannot be solved in sub-exponential time), the best algorithm to compute the weighted chromatic number of n-node trees has time-complexity $n^{\Theta(\log n)}$. Our result mainly relies on proving that, when computing an optimal proper weighted colouring of a graph G, it is hard to combine colourings of its connected components, even when G is a forest.

6.3.1.3. On-line colouring

Since many applications, and in particular channel assignment problems, must be solved on-line, we studied on-line colouring algorithms. The most basic and most widespread of them is the greedy algorithm. The largest number of colours that can be given by the greedy algorithm on some graph. is called its *Grundy number* and is denoted $\Gamma(G)$. Trivially $\Gamma(G) \leq \Delta(G) + 1$, where $\Delta(G)$ is the maximum degree of the graph. In [26], we show that deciding if $\Gamma(G) \leq \Delta(G)$ is NP-complete. We then show that deciding if $\Gamma(G) \geq |V(G)| - k$ is fixed-parameter tractable with respect to the parameter k. We also gave similar complexity results on b-colourings, which is a manner of improving colourings on-line.

In [27], we study a game version of greedy colouring. Given a graph G, two players, Alice and Bob, alternate their turns in choosing uncoloured vertices to be coloured. Whenever an uncoloured vertex is chosen, it is coloured by the least positive integer not used by any of its coloured neighbors. Alice's goal is to minimize the total number of colours used in the game, and Bob's goal is to maximize it. The *game Grundy number* of G is the number of colours used in the game when both players use optimal strategies. It is proved in this paper that the maximum game Grundy number of forests is 3, and the game Grundy number of any partial 2-tree is at most 7.

6.3.1.4. Enumerating edge-colourings and total colourings

With the success of moderately exponential algorithms, there is an increasing interest for enumeration problems, because of their own interest but also because they might be crucial to solve optimization problems. In [21], we are interested in computing the number of edge colourings and total colourings of a connected graph. We prove that the maximum number of k-edge-colourings of a connected k-regular graph on n vertices is $k \cdot ((k-1)!)^{n/2}$. Our proof is constructive and leads to a branching algorithm enumerating all the k-edge-colourings of a connected k-regular graph in time $O^*(((k-1)!)^{n/2})$ and polynomial space. In particular, we obtain a algorithm to enumerate all the 3-edge-colourings of a connected cubic graph in time $O^*(2^{n/2}) = O^*(1.4143^n)$ and polynomial space. This improves the running time of $O^*(1.5423^n)$ of the algorithm of Golovach et al. [76]. We also show that the number of 4-total-colourings of a connected cubic graph is at most $3 \cdot 2^{3n/2}$. Again, our proof yields a branching algorithm to enumerate all the 4-total-colourings of a connected cubic graph.

6.3.2. Directed graphs

Graph theory can be roughly partitioned into two branches: the areas of undirected graphs and directed graphs (digraphs). Even though both areas have numerous important applications, for various reasons, undirected graphs have been studied much more extensively than directed graphs. One of the reasons is that many problems for digraphs are much more difficult than their analogues for undirected graphs.

6.3.2.1. Finding a subdivision of a digraph

One of the cornerstones of modern (undirected) graph theory is minor theory of Robertson and Seymour. Unfortunately, we cannot expect an equivalent for directed graphs. Minor theory implies in particular that, for any fixed F, detecting a subdivision of a fixed graph F in an input graph G can be performed in polynomial time by the Robertson and Seymour linkage algorithm. In contrast, the analogous subdivision problem for digraph can be either polynomial-time solvable or NP-complete, depending on the fixed digraph F. In a previous paper, we gave a number of examples of polynomial instances, several NP-completeness proofs as well as a number of conjectures and open problems. In [71], we conjecture that, for every integer k greater than 1, the directed cycles of length at least k have the Erdős-Pósa Property : for every n, there exists an integer t_n such that for every digraph D, either D contains n disjoint directed cycles of length at least k, or there is a set T of t_n vertices that meets every directed cycle of length at least k. This generalizes a celebrated result of Reed, Robertson, Seymour and Thomas which is the case $k = 2$ of this conjecture. We prove the conjecture for $k = 3$. We also show that the directed k-Linkage problem is polynomial-time solvable for digraphs with circumference at most 2. From these two results, we deduce that if F is the disjoint union of directed cycles of length at most 3, then one can decide in polynomial time if a digraph contains a subdivision of F.

6.3.2.2. Oriented trees in digraphs

Let $f(k)$ be the smallest integer such that every $f(k)$-chromatic digraph contains every oriented tree of order k. Burr proved $f(k) \leq (k-1)^2$ in general, and he conjectured $f(k) = 2k-2$. Burr also proved that every $(8k-7)$-chromatic digraph contains every antidirected tree. We improve both of Burr's bounds. We show [15] that $f(k) \leq k^2/2 - k/2 + 1$ and that every antidirected tree of order k is contained in every $(5k-9)$-chromatic digraph. We also make a conjecture that explains why antidirected trees are easier to handle. It states that if $|E(D)| > (k-2)|V(D)|$, then the digraph D contains every antidirected tree of order k. This is a common strengthening of both Burr's conjecture for antidirected trees and the celebrated Erdős-Sós Conjecture. The analogue of our conjecture for general trees is false, no matter what function $f(k)$ is used in place of $k-2$. We prove our conjecture for antidirected trees of diameter 3 and present some other evidence for it. Along the way, we show that every acyclic k-chromatic digraph contains every oriented tree of order k and suggest a number of approaches for making further progress on Burr's conjecture.

DANTE Team

6. New Results

6.1. Probabilistic resource management

Participants: Paulo Gonçalves [correspondant], Thomas Begin, Shubhabrata Roy, Thibaud Trolliet.

This contribution is part of the PhD work of S. Roy (Dec. 2010 – March 2014) on probabilistic resource management in the context of highly volatile workloads. We proposed a Markovian model that can reproduce the workload volatility occurring in real-life VoD systems, such as Video On Demand (VoD). We derived an original MCMC based identification procedure to calibrate model on real data. We assess the accuracy of the proposed procedure in terms of bias and variance through several numerical experiments, and we compared its outcome with a former ad-hoc method that we had designed. We also compared the performance of our approach to that of other existing models examining the goodness-of-fit of the steady state distribution and of the autocorrelation function of real workload traces. Results show that the combination of out model and its MCMC based calibration clearly outperforms the existing state-of-the art. (See [17], [18])

peta

6.2. Semi-supervised machine learning

Participant: Paulo Gonçalves [correspondant].

This contribution is part of the PhD work of M. Sokol (EPI MAESTRO, Oct. 2009 – May 2014), co-supervised with K. Avrachenkov and Ph. Nain, on the classification of content and users in peer-to-peer networks using graph-based semi-supervised learning methods. Semi-supervised learning methods constitute a category of machine learning methods which use labelled points together with unlabelled data to tune the classifier. The main idea of the semi-supervised methods is based on an assumption that the classification function should change smoothly over a similarity graph, which represents relations among data points. This idea can be expressed using kernels on graphs such as graph Laplacian. Different semi-supervised learning methods have different kernels which reflect how the underlying similarity graph influences the classification results. In a recent work, we analysed a general family of semi-supervised methods, provided insights about the differences among the methods and gave recommendations for the choice of the kernel parameters and labelled points. In particular, it appeared that it was preferable to choose a kernel based on the properties of the labelled points. We illustrated our general theoretical conclusions with an analytically tractable characteristic example, clustered preferential attachment model and classification of content in P2P networks. (See [8])

6.3. Analysis of heart beat rate variability

Participant: Paulo Gonçalves [correspondant].

Intrapartum fetal heart rate monitoring constitutes an important stake aiming at early acidosis detection. Measuring heart rate variability is often considered a powerful tool to assess the intrapartum health status of fetus and has has been envisaged using various techniques. In the present contribution, the power of scale invariance parameters, such as the Hurst exponent and the global regular- ity exponent, estimated from wavelet coefficients of intrapartum fetal heart rate time series, to evaluate the health status of fetuses is quantified from a case study database, constituted at a French Academic Hospital in Lyon. Notably, the ability of such parameters to discriminate subjects incorrectly classified according to FIGO rules as abnormal will be discussed. Also, the impact of the occurrence of decelerations identified as complicated by obstetricians on the values taken by Hurst parameter is investigated in detail. (See [7])

6.4. Hierarchical Modeling of IEEE 802.11 Multi-hop Wireless Networks

Participants: Thiago Wanderley Matos de Abreu, Thomas Begin, Isabelle Guérin Lassous.

IEEE 802.11 is implemented in many wireless networks, including multi-hop networks where communications between nodes are conveyed along a chain. We present a modelling framework to evaluate the performance of flows conveyed through such a chain. Our framework is based on a hierarchical modelling composed of two levels. The lower level is dedicated to the modelling of each node, while the upper level matches the actual topology of the chain. Our approach can handle different topologies, takes into account Bit Error Rate and can be applied to multi-hop flows with rates ranging from light to heavy workloads. We assess the ability of our model to evaluate loss rate, throughput, and end-to-end delay experienced by flows on a simple scenario, where the number of nodes is limited to three. Numerical results show that our model accurately approximates the performance of flows with a relative error typically less than 10%. (See [6])

6.5. Available Bandwidth Estimation in GPSR for VANETs

Participant: Isabelle Guérin Lassous.

We propose an adaptation of the collision probability used in the measure of the available bandwidth designed for Mobile Ad hoc Networks (MANETs) and which is used in ABE. Instead, we propose a new ABE+ that includes a new function to estimate the probability of losses. This new function has been specially designed for Vehicular Ad hoc Networks, to be suited to the high mobility and variable density in vehicular environments. In this new solution, we do not only consider the packet size, but also other metrics, such as, density and speed of the nodes. We include the ABE+ algorithm in the forwarding decisions of the GBSR-B protocol, which is an improvement of the well-known GPSR protocol. Finally through simulations, we compare the performance of our new ABE+ compared to the original ABE. These results show that ABE+ coupled with GBSR-B achieves a good trade-off in terms of packet losses and packet end-to-end delay. (See [19])

6.6. Reduced complexity in $M/Ph/c/N$ queues

Participant: Thomas Begin [correspondant].

This contribution stems from a long-existing collaboration with Pr. Brandwajn (UCSC), which is devoted to innovative numerical solution of classical queueing systems. Many real-life systems can be modelled using the classical $M/G/c/N$ queue. A frequently-used approach is to replace the general service time distribution by a phase-type distribution since the $M/Ph/c/N$ queue can be described by familiar balance equations. The downside of this approach is that the size of the resulting state space suffers from the "dimensionality curse", *i.e.*, exhibits combinatorial growth as the number of servers and/or phases increases. To circumvent this complexity issue, we propose to use a reduced state description in which the state of only one server is represented explicitly, while the other servers are accounted for through their rate of completions. The accuracy of the resulting approximation is generally good and, moreover, tends to improve as the number of servers in the system increases. Its computational complexity in terms of the number of states grows only linearly in the number of servers and phases. (See [9])

6.7. Throughput maximisation in multi-radio wireless networks

Participants: Isabelle Guérin Lassous, Busson Anthony.

Wireless mesh network offers a simple and costless solution to deploy wireless based infrastructure network. They are particularly suitable when the network is deployed temporarily, such as substitution networks (studied in the ANR RESCUE project). In order to ensure an important capacity, the mesh nodes may be equipped with several 802.11 network interfaces. The classical approach to assign 802.11 channels to these interfaces aim to minimise global interference, *i.e.* minimise the conflict graph. Our proposition is two folds. We define a new benefit function that describes the network capacity rather than interference/conflicts. Also, we derive an efficient algorithm that maximises this function. Simulation results show that the proposed function is very close to the measured end-to-end throughputs, empirically proving that it is the good function to optimise. Moreover, the channel assignation algorithm based on this optimisation presents an important throughput increase compared to the classical approaches.

6.8. Aggregation of temporal contact series into graph series

Participants: Christophe Crespelle, Eric Fleury, Yannick Léo.

We consider the problem of aggregating a temporal contact series into a series of graph. This consists in slicing time into time-windows of equal length and forming for each window the graph of the contacts occurred within it. The length chosen for the windows has a great impact on the properties of the graph series obtained. Then the key question that arises is: how one should choose the length of aggregation windows? In the master insternship of Yannick Léo (spring 2013), we designed a method to do so, by using the occupation rate of paths in the graph series. We have applied this method on several real-world data and obtained very good results. The method has also greatly beneficiated of a new notion of shortest dynamic paths that we developed during the master internship of Pierre-Alain Scribot (spring 2013).

6.9. Dynamic Contact Network Analysis in Hospital Wards

Participants: Christophe Crespelle, Eric Fleury, Lucie Martinet.

We analysed a huge and very precise trace of contact data collected during 6 months on the entire population of a rehabilitation hospital. We investigated the graph structure of the average daily contact network, and we unveiled striking properties of this structure in the considered hospital, as a very strong introversion of services, the key role of the contacts between patients and staff in connecting those introverted services all together, and very different pattern of contacts during one day between patients and staffs. The methodology we designed to lead these analysis is very general and can be applied for analysing any dynamic complex network where nodes are classified into groups. Those results are part of Lucie Martinet's PhD thesis.

6.10. A Linear-Time Algorithm for Computing the Prime Decomposition of a Directed Graph with Regard to the Cartesian Product

Participant: Christophe Crespelle.

We design the first linear-time algorithm for computing the prime decomposition of a digraph G with regard to the cartesian product. A remarkable feature of our solution is that it computes the decomposition of G from the decomposition of its underlying undirected graph, for which there exists a linear-time algorithm. First, this allows our algorithm to remain conceptually very simple and in addition, it provides new insight into the connexions between the directed and undirected versions of cartesian product of graphs [11]

6.11. Linear-time Constant-ratio Approximation Algorithm and Tight Bounds for the Contiguity of Co-graphs

Participant: Christophe Crespelle.

We consider a graph parameter called *contiguity* which aims at encoding a graph by a linear ordering of its vertices. The purpose is to obtain very compact encoding of a graph which still answers in optimal time to neighbourhood queries on the graph (*i.e.* list the neighbours of a given vertex). This allows to deal with very large instances of graphs by loading them entirely into the memory, without penalising the running time of algorithms treating those instances. We designed a linear time algorithm for computing a constant-ratio approximation of the contiguity of an arbitrary co-graph. Our algorithm does not only give an approximation of the parameter, but also provides an encoding of the co-graph realising this value [10]

6.12. Model for Time-Varying Graphs.

Participant: Éric Fleury.

We propose a novel model for representing finite discrete Time-Varying Graphs (TVGs). The major application of such a model is for the modelling and representation of dynamic networks. In our proposed model, an edge is able to connect a node u at a given time instant t_a to any other node v (u possibly equal to v) at any other time instant t_b (t_a possibly equal to t_b), leading to the concept that such an edge can be represented by an ordered quadruple of the form (u, t_a, v, t_b). Building upon this basic concept, our proposed model defines a TVG as an object $H = (V, E, T)$, where V is the set of nodes, $E \subseteq V \times T \times V \times T$ is the set of edges, and T is the finite set of time instants on which the TVG is defined. We show how key concepts, such as degree, path, and connectivity, are handled in our model. We also analyse the data structures used for the representation of dynamic networks built following our proposed model and demonstrate that, for most practical cases, the asymptotic memory complexity of our TVG representation model is determined by the cardinality of the set of edges. (See [20])

DIANA Team

6. New Results

6.1. Service Transparency

Participants: Chadi Barakat, Walid Dabbous, Maksym Gabielkov, Young-Hwan Kim, Arnaud Legout, Byungchul Park, Ashwin Rao, Riccardo Ravaioli, Damien Saucez, Thierry Turletti.

The Complete Picture of the Twitter Social Graph

We made an in-depth study of the macroscopic structure of the Twitter social graph unveiling the highways on which tweets propagate, the specific user activity associated with each component of this macroscopic structure, and the evolution of this macroscopic structure with time for the past 6 years. For this study, we crawled Twitter to retrieve all accounts and all social relationships (follow links) among accounts; the crawl completed in July 2012 with 505 million accounts interconnected by 23 billion links. Then, we presented a methodology to unveil the macroscopic structure of the Twitter social graph. This macroscopic structure consists of 8 components defined by their connectivity characteristics. Each component group users with a specific usage of Twitter. For instance, we identified components gathering together spammers, or celebrities. Finally, we introduced a method to approximate the macroscopic structure of the Twitter social graph in the past, validate this method using old datasets, and discuss the evolution of the macroscopic structure of the Twitter social graph during the past 6 years. This work is accepted in Sigmetrics'14 [23].

Meddle: Middleboxes for Increased Transparency and Control of Mobile Traffic

Meddle is a platform that relies on traffic indirection to diagnose mobile Internet traffic. Meddle is motivated by the absence of built-in support from ISPs and mobile OSes to freely monitor and control mobile Internet traffic; the restrictions imposed by mobile OSes and ISPs also make existing approaches impractical. Meddle overcomes these hurdles by relying on the native support for traffic indirection by mobile OSes. Specifically, Meddle proxies mobile Internet traffic through a software defined middleboxes configured for mobile traffic diagnosis. We use Meddle to tests the limits of the network perspective of mobile Internet traffic offered by traffic indirection. We use this perspective to characterize and control the behavior of mobile applications and provide a first look at ISP interference on mobile Internet traffic. We then performed controlled experiments on 100 popular iOS and Android applications to show how Meddle can be used to identify misbehavior and to block traffic causing this misbehavior. Unlike existing solutions, this activity can be performed without warranty voiding the device and activated on the fly on-demand. This work is done in the context of Aswhin Rao's PhD thesis [11] in collaboration with Northeastern University and Berkeley.

Understanding of modern web traffic

This recent years and with the advent of mobile devices, web traffic has changed and moved from static to dynamic generation. Interestingly, while it is well known that network protocols are intertwined in such a way the characteristics of a layer are affected by those of other layers, most of the measurement work done so far does not pay enough attention to this aspect. We then conducted a cross-layer measurement analysis that confronts all the layers from the very deep technological details to the very high level of users behaviors to shed new light on this issue. To support our study, we analysed an Internet packet traffic trace and showed how this cross-layer analysis approach can explain why TCP flows in mobile traffic are larger than usual. We are currently refining our study to characterises the discrepancies between the different network stack protocol implementations based on the mobile/non-mobile nature of the devices but also their operating system and version. This work is currently under submission.

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 assume that discrimination can take 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 performance strongly depends on the way routers reply to probe packets. We carried out large scale experiments to understand the way routers reply to our probes and we calibrated models to these replies. The next step will be to evaluate the performance of Chkdiff under these models, before making the tool public and available to the community. Chkdiff is currently the subject of a collaboration with I3S around the PhD thesis of Riccardo Ravaioli (funded by the Labex UCN@Sophia). The work is ongoing and will be submitted soon.

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 has been published in [18].

Packet Extraction Tool for Large Volume Network Traces

Network packet tracing has been used for many different purposes during the last few decades, such as network software debugging, networking performance analysis, forensic investigation, and so on. Meanwhile, the size of packet traces becomes larger, as the speed of network rapidly increases. Thus, to handle huge amounts of traces, we need not only more hardware resources, but also efficient software tools. However, traditional tools are inefficient at dealing with such big packet traces. We proposed pcapWT, an efficient packet extraction tool for large traces. PcapWT provides fast packet lookup by indexing an original trace using a Wavelet Tree structure. In addition, pcapWT supports multi-threading for avoiding synchronous I/O and blocking system calls used for file processing, and is particularly efficient on machines with SSD. PcapWT shows remarkable performance enhancements in comparison with traditional tools such as tcpdump and most recent tools such as pcapIndex in terms of index data size and packet extraction time. Our benchmark using large and complex traces shows that pcapWT reduces the index data size down below 1% of the volume of the original traces. Moreover, packet extraction performance is 20% better than with pcapIndex. Furthermore, when a small amount of packets are retrieved, pcapWT is hundreds of times faster than tcpdump. These results, done in collaboration within the CIRIC, have just been submitted to Computer Networks[34].

Impact of new transport protocols on BitTorrent performance

In the paper [27], we address the trade-off between the data plane efficiency and the control plane timeliness for the BitTorrent performance. We argue that loss-based congestion control protocols can fill large buffers, leading to a higher end-to-end delay, unlike low-priority or delay-based congestion control protocols. We perform experiments for both the uTorrent and mainline BitTorrent clients,

and we study the impact of uTP (a novel transport protocol proposed by BitTorrent) and several TCP congestion control algorithms (Cubic, New Reno, LP, Vegas and Nice) on the download completion time. Briefly, in case peers in the swarm all use the same congestion control algorithm, we observe that the specific algorithm has only a limited impact on the swarm performance. Conversely, when a mix of TCP congestion control algorithms coexists, peers employing a delay-based low-priority algorithm exhibit shorter completion time.

6.2. Open Network Architecture

Participants: Bruno Astuto Arouche Nunes, Chadi Barakat, Daniel Camara, Walid Dabbous, Lucia Guevgeozian Odizzio, Young-Hwan Kim, Mohamed Amine Larabi, Arnaud Legout, Emilio Mancini, Xuan-Nam Nguyen, Thierry Parmentelat, Alina Quereilhac, Damien Saucez, Julien Tribino, Thierry Turletti, Fréderic Urbani.

Delay Tolerant Networks

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 a content-centric dessemination algorithm for delay-tolerant networks, called for short CEDO, that distributes content to multiple receivers over a DTN. CEDO 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. CEDO 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 implemented CEDO over the CCNx protocol, which provides the basic tools for requesting, storing, and forwarding content. Detailed information on CEDO and the implementation work carried out herein can be found in this publication [22] and at the following web page: http://planete.inria.fr/Software/CEDO/.

Predicting nodes spatial node density in mobile ad-hoc networks

User mobility is of critical importance when designing mobile networks. In particular, "waypoint" mobility has been widely used as a simple way to describe how humans move. This paper introduces the first modeling framework to model waypoint-based mobility. The proposed framework is simple, yet general enough to model any waypoint-based mobility regimes. It employs first order ordinary differential equations to model the spatial density of participating nodes as a function of (1) the probability of moving between two locations within the geographic region under consideration, and (2) the rate at which nodes leave their current location. We validate our models against real user mobility recorded in GPS traces collected in three different scenarios. Moreover, we show that our modeling framework can be used to analyze the steady-state behavior of spatial node density resulting from a number of synthetic waypoint-based mobility regimes, including the widely used

Random Waypoint (RWP) model. Another contribution of the proposed framework is to show that using the well-known preferential attachment principle to model human mobility exhibits behavior similar to random mobility, where the original spatial node density distribution is not preserved. Finally, as an example application of our framework, we discuss using it to generate steady-state node density distributions to prime mobile network simulations. This work was done in collaboration with Dr. Katia Obraczka, from UC Santa Cruz, and was published in WINET [12].

Software Defined Networking in Heterogeneous Networked Environments

We worked on the exploration of the software defined networking paradigm to facilitate the implementation and large scale deployment of new network protocols and services in heterogeneous networked environments. Our activities related to this research thrust are described hereafter. We wrote a survey of the emerging field of Software-Defined Networking (SDN). SDN is currently attracting significant attention from both academia and industry. Its field is quite recent, yet growing at a very fast pace. Still, there are important research challenges to be addressed. We look at the history of programmable networks, from early ideas until recent developments. In particular we described the SDN architecture in detail as well as the OpenFlow standard. We presented current SDN implementations and testing platforms and examined network services and applications that have been developed based on the SDN paradigm. We concluded with a discussion of future directions enabled by SDN ranging from support for heterogeneous networks to Information Centric Networking (ICN). The survey will be published in 2014 in the IEEE Surveys and Tutorials journal [32].

We have also specified a number of use cases motivating the need for extending the SDN model to heterogeneous networked environments. Such environments consist of infrastructure-based and infrastructure-less networks. These specifications and use cases were summarized in a recent publication [19].

We have also implemented a Capacity Sharing platform by leveraging SDN in hybrid networked environments, i.e., environments that consist of infrastructure-based as well as infrastructureless networks. The proposed SDN-based framework provides flexible, efficient, and secure capacity sharing solutions in a variety of hybrid network scenarios. In the paper published at the Capacity Sharing Workshop CSWS 2013 [40], we identify the challenges raised by capacity sharing in hybrid networks, describe our framework in detail and how it addresses these challenges, and discuss implementation issues.

The aforementioned capacity sharing work is just one application and a preliminary of our longer term effort. We have also started to specify the H-SDN protocols based on the use cases mentioned above, including the capacity sharing use case. These efforts are part of a broader work where we propose a framework to enable the implementation and deployment of more generic H-SDN networks and applications. This framework contemplates important issues regarding H-SDN deployment, such as: security, increased scalability and performance by distribution of SDN control and seamless handover of mobile stations, to name a few. We have targeted Mobisys2014 as a venue for publishing our proposal and results regarding this topic [39].

Rule Placement in Software-Defined Networking

OpenFlow is a new communication standard that decouples control and data planes to simplify traffic management. More precisely, OpenFlow switches populate their forwarding tables by opportunistically querying a centralized controller for flows whose rules (i.e., forwarding actions) are not yet installed. However, the flexibility offered by this new paradigm comes at the expense of extra signaling overhead as, in practice, switches might not be able to store all rules in their local forwarding tables. The question of which rules to install then becomes essential. In our research, we leverage the fact that some flows are more important to manage than others, and thus construct an optimal placement problem of rules in OpenFlow switches that ensures the most valuable traffic is matched by its appropriate rules while respecting switches and links capacity constraints. The rest of the traffic with no installed rules follows a default, yet less appropriate, path within the network. We have

formulated and solved this optimisation problem in the case of realistic operational needs, and prove that the optimal placement of rules is NP-hard. The intrinsic complexity of the problem led us to design a greedy heuristic that we evaluated with two representative use cases: BGP multihoming and Access Control Lists. On one hand, the evaluation shows the versatility and the generality of the optimization problem, and on another hand, it demonstrates that heuristics with apparent simplicity are still efficient. We are now extending this work to support traffic dynamics and mobility. This work is currently under submission.

Information-Centric Networking and economical aspects

With the explosion of broadband Over-The-Top (OTT) services all around the world, the Internet is autonomously migrating toward overlay and incrementally deployable content distribution infrastructures. Information-Centric Networking (ICN) technologies are the natural candidates to more efficiently bind and distribute popular contents to users. However, the strategic incentives in exploiting ICN, for both users and ISPs, are much less understood to date. In this work, we shed light on how OTTs shall shape prices and discounts to motivate ICN usage, depending on their awareness over content distribution costs. Actually, the Internet ecosystem is fast and dynamic and new ideas can rapidly reach millions of users spread worldwide without having to rely on special involvement of intermediate transit networks. In this context, Over-The-Top broadband content providers can leverage their customer resources to allowing, from one hand, to improve access performance, and, from the other hand, to reduce operational costs the OTT provider would incur on by directly serving the customers. In this context, Information-Centric Networking appears as an adequate offloading technique, if incrementally deployed as an overlay network. This paper analyses the incentive compatibility in the adoption of a ICN overlay for OTT services and is, as of our knowledge, we are the first in addressing the topic by following a non-cooperative game theory reasoning, we believe adequate in its non-cooperative nature due to independency between the involved ICN stakeholders. Our analysis allows us to assess that the business model currently standing for legacy CDNs does not make strategic sense for ICN overlays and that, hover, it exists incentives for OTT customers to get involved in the distributions process via an ICN overlay reducing so server load. These unique specifications for the design of an ICN overlay for OTT content distribution do also have relevant implications for ICN protocol design. The OTT provider would need a form of control over the ICN overlay operations. We identify the usage of a OTT- set policy metric for ICN routing as the most appropriate way to ensure ICN users follow the equilibrium strategy suggested by our incentive compatibility framework. We highlight moreover the need of a scalable way of building and controlling ICN overlays over the legacy TCP/IP Internet to support related signaling, forwarding rule registration, and positive strategic behaviour.

Information-Centric Networking and rate control implications

Information-centric networking (ICN) leverages content demand redundancy and proposes in-network caching to reduce network and servers load and to improve quality of experience. We have studied the interaction between in-network caching of ICN and Additive Increase Multiplicative Decrease (AIMD) end-to-end congestion control with a focus on how bandwidth is shared, as a function of content popularity and caches provisioning. As caching shortens AIMD feedback loop, the download rate of AIMD is impacted. We earlier shed light on the potential negative impact of in-network caching on instantaneous throughput fairness. The work accomplished in 2013 precisely quantify the issue thanks to an analytic model based on Discriminatory Processor Sharing and real experiments, we observe that popular contents benefit from caching and realize shorter download times at the expense of unpopular contents which see their download times inflated by a factor bounded by $\frac{1}{1-\rho}$, where ρ is the network load. This bias can be removed by redefining congestion control to be delay independent or by over-provisioning link capacity at the edge so that to compensate for the greediness of popular contents. The experimentation study has been supported by the work of Ilaria Cianci internship on the CCN-Jocker emulator. This work is currently under submission.

Routing in Information-Centric Network

The idea behind Information-Centric Networking (ICN) is to omit the notion of host and location and use contents as direct routing and forwarding primitives, instead of IP addresses. This shift of paradigm allow ICN to natively offer in-network caching, i.e., to cache content on the path from content providers to requesters. Actually our studies shows a large spatial and temporal locality of contents amongst users in the same network which proves that in-network caching can achieve good overall performance. However, caching contents strictly on their paths is far from being optimal when paths are not shared among content consumers as contents may be replicated on routers so reducing the total volume of contents that can be cached. To overcome this limitation, we introduced the notion of off-path caching in [21] where we allocate content to well defined off-path caches within the network and deflect the traffic off the optimal path toward these caches that are spread across the network. Off-path caching improves the global hit ratio by efficiently utilizing the network-wide available caching capacity and permits to reduce egress links bandwidth usage.

Locator/Identifier Separation Protocol (LISP)

The future Internet has been a hot topic during the past decade and many approaches proposed towards this future Internet, ranging from incremental evolution to complete clean state ones, have been proposed. One of the proposition, LISP, advocates for the separation of the identifier and the locator roles of IP addresses to reduce BGP churn and BGP table size. Up to now, however, most studies concerning LISP have been theoretical and, in fact, little is known about the actual LISP deployment performance. We filled this gap through measurement campaigns carried out on the LISP Beta Network. More precisely, we evaluated the performance of the two key components of the infrastructure: the control plane (i.e., the mapping system) and the interworking (i.e., communication between LISP and non-LISP sites). Our measurements highlight that performance offered by the LISP interworking infrastructure is strongly dependent on BGP routing policies. If we exclude misconfigured nodes, the mapping system typically provides reliable performance and relatively low median mapping resolution delays. Although the bias is not very important, control plane performance favours USA sites as a result of its larger LISP user base but also because European infrastructure is unreliable. Finally, the LISP Map-versioning RFC mentioned in the last year activity report was published this year [33]. All details are reported in [17], [29].

Running Live CCNx Experiments on Wireless and Wired Testbeds with NEPI

CCNx has long left the early development stage where simulation and emulation frameworks, like ccnSim and mininet, were enough to validate new approaches and improvements. It has now reached a level of maturity which calls for evaluation in more realistic environments. If it is to be deployed in the wild Internet or even in private network settings, a framework that provides proper validation in comparable environments is required. For this purpose we demonstrate the capabilities of the NEPI framework to run CCNx experiments in realistic environments. NEPI can run CCNx experiments directly on Internet settings as well as wireless or wired private network environments. This framework allows to automate host con guration, software installation, result collection and to define execution sequence between applications. Furthermore, it provides the ability to conduct interactive experiments where researchers are free to modify the experiment scenario on the fly. These results were demonstrated at CCNxCon'2013 [38].

Evaluating costs of CCN overlays

We are currently involved in a collaboration with PARC (Palo Alto research center) regarding the evaluation of the CCN (Control Centric Networking) technology. Early results of this work were presented in the poster session at the CCNxConf 2013 meeting. In this work we present a set of scenarios to evaluate the performance of CCN overlays on top of the Internet, for worse case conditions. We used the NEPI experiment API to construct different overlay topologies on PlanetLab, for which we varied the topology configuration (e.g. number and degree of nodes), the CCN parameters (e.g. pipeline, cache usage, prefix routes) and the traffic patterns (e.g. single stream, prefix independent chunks). The objective of this study is to find correlations between these variables and the time to deliver content and the overlay network utilization. Our contribution is twofold. In one hand we provide a benchmark which can be used as reference for comparison of new CCNx

versions and for other ICN solutions, and as input traces for CCN simulations. In the other hand, we provide results that can be used to improve the CCNx implementation and that can help Internet providers or end users to better design CCN overlays to satisfy their needs. The work is still ongoing and will be submitted soon.

Enabling Iterative Development and Reproducible Evaluation of Network Protocols

Over the last two decades several efforts have been made to provide adequate experimental environments, aiming to ease the development of new network protocols and applications. These environments range from network simulators providing highly controllable evaluation conditions, to live testbeds providing realistic evaluation environment. While these different approaches foster network development in different ways, there is no simple way to gradually transit from one to another, or to combine their strengths to suit particular evaluation needs. We believe that enabling a gradual transition from a pure simulated environment to a pure realistic one, where the researcher can decide which aspects of the environment are realistic and which are controllable, allows improving network solutions by simplifying the problem analysis and resolution. We have designed a new network experimentation framework, called IDEV, where simulated and real components can be arbitrarily combined to build custom test environments, allowing refining and improving new protocols and applications implementations by gradually increasing the level of realism of the evaluation environment. Moreover, we proposed a testbed architecture specifically adapted to support the proposed concept, and discuss the design choices we made based on our previous experience in the area of network testbeds. These choices address key issues in network testbed development, such as ease of experimentation, experiment reproducibility, and testbed federation, to enable scaling the size of experiments beyond what a single testbed would allow. This work has been described in a paper that will be published in the Computer Networks journal in 2014, see [15].

Direct Code Execution: Revisiting Library OS Architecture for Reproducible Network Experiments

We proposed Direct Code Execution (DCE), a framework that dramatically increases the number of available protocol models and realism available for ns-3 simulations. DCE meets the goals recently proposed for fully reproducible networking research and runnable papers, with the added benefits of 1) the ability of completely deterministic reproducibility, 2) the scalability that simulation time dilation offers, 3) capabilities supporting automated code coverage analysis, and 4) improved debuggability via execution within a single address space. We reported on packet processing benchmark and showcased key features of the framework with different use cases. Then, we reproduced a previously published Multipath TCP (MPTCP) experiment and highlight how code coverage testing can be automated by showing results achieving 55-86% coverage of the MPTCP implementation. We also demonstrated how network stack debugging can be easily performed and reproduced across a distributed system. Our first benchmarks are promising and we believe this framework can benefit the network community by enabling realistic, reproducible experiments and runnable papers. This work has been published in the ACM CoNext conference 2013 [25], in Santa Barbara, CA, USA and will be published in IEEE Communication Magazine in 2014 [14]. DCE has been demonstrated at the ACM MSWiM conference at Barcelona, Spain in November 2013 [42] In the same context, we designed DCE Cradle, a framework that allows to use any features of the Linux kernel network stack with existing ns-3 applications. DCE Cradle uses DCE to address the brittleness of Network Simulation Cradle (NSC). We carefully designed DCE Cradle without breaking the existing functionality of DCE and ns-3 socket architecture by considering the gaps between the asynchronous ns-3 socket API and the general POSIX socket API. We validated the implementation of DCE Cradle with the behavior of TCP implementation in congested links, and then studied its performance by focusing on the simulation time and network scale. We showed that DCE Cradle is at most 1.3 times faster than NSC, while it is about 2.2 times slower than the ns-3 native stack. Then we showcased an actual implementation of the DCCP transport protocol to verify how easy it is to simulate a real implementation using DCE Cradle. We believe that this tool can highly benefit the network community by enabling more realistic evaluation of network

protocols. This work has been published in the ns-3 workshop in 2013 in Cannes and got the best paper award [26].

The ns-3 Consortium

We have founded in 2012 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. The Consoritum started his activities in March 2013. Two European institutions (Centre Tecnològic de Telecomunicacions de Catalunya - CTTC and INESC Porto)) and two Amercian universities (Georgia Tech and Bucknell) joined the consortium as Executive members in 2013. For more details see the consortium web page https://www.nsnam.org/consortium/.

Contiki over ns-3

This year we worked on the adaptation of Contiki OS over ns-3. Contiki is a popular, and highly optimized, operating system for sensor nodes. We developed a proof of concept adaptation layer that, even though simple and limited, was able to show that such interaction is indeed possible. The adaptation layer was capable of transferring data from different sensors using ns-3 to interconnect them. Sensor nodes were controlled by the ns-3 scheduler, respecting the ns-3 clock and executing over simulated time. In fact, the sensors were not even aware they were placed over a simulated network.

Federation of experimental testbeds

We are involved in the F-Lab (French ANR) 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 DIANA team and Princeton University, we kept contributing 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 2013, as a step forward to our contribution to the specification of the Aggregate Manager (AM) API v3, which is the control plane interface through which experimenters discover and reserve resources at testbeds, we have focused on coming up with a separate implementation of SFAWrap that supports AM API v3 and brings a more elaborate lifecycle for slices provisioning. Secondly, we implemented a AM API v2 to AM API v3 adapter, which represents the glue between the already existing AM API v2 compliant testbed drivers and the AM API v3 compliant interfaces of SFAWrap. The v2 to v3 adapter provides AM API v3 compatibility to already existing AM API v2-based testbed drivers until their authors find the time to adapt their driver for a native support of AM API v3 if they want to take full advantage of the new lifecycle. Thirdly, within the contexts of the formerly listed projects, and as a consequence of the growing need for testbeds federation, the providers of testbeds such as: BoneFire, SmartSantander decided to adopt SFAWrap in order to join the global federation of testbeds by exposing their testbeds through SFA. Thus, we had to provide to those partners a close support to achieve this goal. Finally, as for any kind of software development project, and due to the growing usage of SFAWrap, we had to be active on both operational and maintanace tasks. See [37] and [41] for more details. We also contributed, in the context of the Fed4FIRE project, to the definition and early implementation of an architecture for heterogeneous federation of future internet experimental facilities. The results of this work were presented at the FutureNetworkSummit 2013 conference. In this work, requirements involving different aspects of the federation of heterogeneous facilities where collected and analysed, and a multilayer architecture was proposed to address them. Our contribution mainly focuses on the experiment control plane of the federation architecture [28]. The experiment control plane involves the interface between the experimenter and the facilities, and

it covers tasks such as federation of the resource discovery, provisioning, reservation, configuration and deployment. The proposed architecture combines the use of SFA (Slice Federation Architecture) and OMF (cOntrol and Management Framework) into a common middle-ware that allows to federate resource control within an experiment across facilities.

DIONYSOS Project-Team

6. New Results

6.1. Quality of Experience

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, Gerardo Rubino, César Viho, Pantelis Frangoudis, Hyunhee Park, Kandaraj Piamrat.

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 that can be obtained from a panel of human observers, under a controlled subjective testing experiment, following an appropriate standard (which depends on the type of sequence or application). Efficiency means that our measuring tool can work in real time, if necessary. Observe that perceived quality is, in general, the main component of QoE when the application or service involves video and audio, or voice. PSQA works by analyzing the networking environment of the communication and some the technical characteristics of the latter. It works without any need to the original sequence (as such, it belongs to the family of *no-reference* techniques).

It must be pointed out that a PSQA measuring or monitoring module is network-dependent and application-dependent. Basically, for each specific networking technology, 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, typically for controlling purposes.

Learning tools. 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) for that purpose (see [74] for a general description), but recently, we 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) [71] with the goal of attacking the main limitations in training time for recurrent neural networks while introducing no significant disadvantages. Two RC models have been proposed independently and simultaneously under the name of *Liquid State Machine* (LSM) [73] and *Echo State Networks* (ESN) [71]. They constitute today one of the basic paradigms for Recurrent Neural Networks modeling [72]. 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 high-dimensional dynamical system that expand the input stream in a space of states. The learning part of the model is the parametric one. In [41] we propose a new learning tool which merges the capabilities of Random Neural Networks (RNNs) with those of RC models. We keep some of the nice features of RNNs with the ability of RC models 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. In [63], more results about the good behavior of our new tool are presented.

QoE for SVC. A recent video encoding scheme called Scalable Video Coding (SVC) provides the flexibility and the capability to adapt the video quality to varying network conditions and heterogeneous users. Last year, we started to look at the relations between the way SVC is used and the obtained perceived quality. This year we continued these efforts, together with exploring the use of QoE estimation tools for SVC video coding in network control. In [46] we evaluate different configurations for SVC-based adaptive streaming in terms of user QoE. The aim is to provide recommendations about the different rates to be used in order to create the video representation configuration. These results are part of the PhD [11]. In [25], we extended our previous work on SVC in DVB-T2, by proposing an analytical model to evaluate the performance of associating SVC with DVB-T2 and QoE. To do this, we developed a discrete time Markov Chain model which captures the

system evolution in terms of number of SVC layers that need to be decoded in order to increase user QoE. In [45], we introduced a new solution to be used by a DASH client for selecting the video representation. Our proposal relies on using the PTP synchronization protocol in order to estimate the end-to-end delays between the client and the server, and hence to correlate this information with network load. The correlation between delays and load was based on a fitting function.

In [54], we focus on SVC multicast over IEEE 802.11 networks. Traditionally, multicast uses the lowest modulation resulting in a video with only base quality even for users with good channel conditions. To optimize QoE, we propose to use multiple multicast sessions with different transmission rates for different SVC layers. The goal is to provide at least the multicast session with acceptable quality to users with bad channel conditions and to provide additional multicast sessions having SVC enhancement layers to users with better channel conditions. The selection of modulation rate for each SVC layer and for each multicast session is achieved with binary integer linear programming depending on network conditions with a goal to maximize global QoE. The results show that our algorithm maximizes global QoE by providing highest quality videos to users with good channel conditions and by guaranteeing at least acceptable QoE for all users.

VoIP. We continued to work on the perceptual quality of voice-based applications and services. In [17], we consider a well-known and widely used *full-reference* technique for measuring speech quality called PESQ, and we propose a learning-based tool for approximating PESQ output without any need for the original signal, following the same black-box parametric PSQA approach. The procedure uses the Echo State Networks previously mentioned.

In [48], we propose a new packet loss model that differentiates loss instances depending on their perceptual impact. In particular, the model captures the differences between short and long interruptions from the perceptual quality viewpoint. In some cases, the delays and their variation have a strong impact on the perceived quality. In [49] we explore the variability of packet delays on MANETs. For that purpose, a wide range of representative scenarios are defined and simulated. The gathered traces are then inspected from qualitative and quantitative perspectives. In [50], a Markovian model is proposed to capture these and other features of delays in the same class of mobile networks.

6.2. Network Economics

Participant: Bruno Tuffin.

The general field of network economics, analyzing the relationships between all actors of the digital economy, has been an important subject for years in the team.

A new book on the subject. We have published a book on this broad topic [61]. Presenting a balance of theory and practice, this up-to-date guide provides a comprehensive overview of the key issues in telecommunication network economics, as well as the mathematical models behind the solutions. These mathematical foundations enable the reader to understand the economic issues arising at this pivotal time in network economics, from business, research, and political perspectives. This is followed by a unique practical guide to current topics, including app stores, volume-based pricing, auctions for advertisements, search engine business models, the network neutrality debate, the relationship between mobile network operators and mobile virtual network operators, and the economics of security. The guide discusses all types of players in telecommunications, from users, to access and transit network providers; to service providers (including search engines, cloud providers or content delivery networks); to content providers, and regulatory bodies. The book is designed for graduate students, researchers, and industry practitioners working in telecommunications.

Research contributions in network economics during 2013 can be decomposed into the application of auction theory, cognitive networks, and network/search neutrality analysis.

Auction theory. In the next generation Internet, we have seen the convergence of multimedia services and Internet with the mobility of users. Vertical handover decision (VHD) algorithms are essential components of the mobility management architecture in mobile wireless networks. VHD algorithms help mobile users to choose the best mobile network to connect among available candidates. It also can help the network manager to optimize easily the limited resources shared among the network providers and the users. In [26], we formulate

VHD algorithm as a resource allocation problem for down-link transmission power in multiple W-CDMA networks and show how combinatorial double-sided auctions can be applied to this specific problem. The proposed pricing schemes make use of the signal interference to noise ratio (SINR), achievable data rates, power allocation at mobile networks, and monetary cost as decision criteria, and our model differentiates new calls and on-going communications to take into account that the last category has somewhat more importance. Several combinatorial double-sided auction are proposed to maximize the social welfare and /or to provide incentives for mobile users and mobile operators to be truth-telling in terms of valuation or cost. Finally, the economic properties of the different proposed pricing schemes are also studied by means of simulations.

Cognitive networks. Cognitive radio technologies for spectrum sharing have received an enormous interest from the research community for the last decade, and more recently from regulators and mobile operators. We have studied a cognitive radio network in [47] where primary operator and an entrant secondary operator compete for users. The system is modeled using queueing and game theories. The economic viability of supporting the secondary operator service using an opportunistic access to the spectrum owned by the primary operator is assessed. Against the benchmark of the primary operator operating as a monopolist, we show that the entry of the secondary operator is desirable from an efficiency perspective, since the carried traffic increases. For a range of parameter values, a lump sum payment can be designed so that the incumbent operator has an incentive to let the secondary operator enter. Additionally, the opportunistic access setting has been compared against a leasing-based alternative, and we have concluded that the former outperforms the latter in terms of efficiency and incentive.

Network/search neutrality analysis. Network neutrality is the topic of a vivid and very sensitive debate, in both the telecommunication and political worlds, because of its potential impact in everyday life. That debate has been raised by Internet Service Providers (ISPs), complaining that content providers (CPs) congest the network with insufficient monetary compensation, and threatening to impose side payments to CPs in order to support their infrastructure costs. 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. However, this is a typical situation that is condemned by ISPs, and, according to them, another reason of the non-neutrality need. We develop and analyze in [23] a model 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, 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.

The very related recently raised search neutrality debate questions the ranking methods implemented by search engines: when a search is performed, do they (or should they) display the web pages ordered according to the quality-of-experience (relevance) of the content? In [68], we analyze that question in a setting when content is offered for free, content providers making revenue through advertising. For content providers, determining the amount of advertising to add to their content is a crucial strategic decision. Modeling the trade-off between the revenue per visit and the attractiveness, we investigate the interactions among competing content providers as a non-cooperative game, and consider the equilibrium situations to compare the different ranking policies. Our results indicate that when the search engine is not involved with any high-quality content provider, then it is in its best interest to implement a neutral ranking, which also maximizes user perceived quality-of-experience and favors innovation. On the other hand, if the search engine controls some high-quality content, then favoring it in its ranking and adding more advertisement yields a larger revenue. This is not necessarily at the expense of user perceived quality, but drastically reduces the advertising revenues of the other content providers, hence reducing their chances to innovate.

6.3. Wireless and Mobile Networks

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, César Viho, Osama Arouk, Btissam Er-Rahmadi, Hyunhee Park, Kandaraj Piamrat.

We continue our activities around wireless and mobile networks, where we focus particularly on 4G networks as well as on a new mobile architecture known as mobile cloud.

LTE improvements. First part of our works concentrates on emerging applications and their impact on 4G networks. In [58], we proposed a solution to handle social network traffic, which is characterized by its elasticity and intensity in a short period of time. The proposed contribution is based on content detection systems such as Deep Packet Inspection (DPI) to identify traffic belonging to a group of users (sharing the same content) of a social network. Upon detecting the type of traffic, we proposed to control it by creating a multicast group. This would reduce the amount of traffic exchanged by switching from unicast communications to multicast communications. Another solution is to cache, at the geographically nearest base station, the shared content among users. Here we positioned ourselves in the case where the social network traffic comes from the same geographical region. We also investigated network decentralization in conjunction with the selective IP traffic offload approaches to handle such increased data traffic. We first devised different approaches based on a per-destination-domain-name basis, which offer operators a fine-grained control to determine whether a new IP connection should be offloaded or accommodated via the core network. Two of our solutions are based on Network Address Translation (NAT) named simple-NATing and twice-NATing, whereas a third one employs simple tunneling, and a fourth adopts multiple Access Point Names. We also proposed methods enabling user equipment devices to always have efficient packet data network connections [30]. Another aspect, we addressed is the gateway selection process, where in [59] we argue the need for other metrics to improve the gateway selection mechanisms in distributed mobile networks. We therefore proposed to consider the end-to-end connection and the service/application type as two important additional metrics in the selection of data anchor gateways in the context of the Evolved Packet System (EPS).

M2M. In [56], [32] we addressed another type of traffic that appeared these last years, namely Machine to machine communication or Machine Type Communication (MTC). Such traffic is known by its intensity and its impact on increasing congestion in both parts of the 4G networks, the Radio Access Network (RAN) and the core network parts. The main spirit of the proposed solutions is to proactively anticipate system overload by reducing the amount of MTC signaling messages exchanged in normal network operations. The first solution reduces the number of exchanged signaling messages when triggering MTC devices with low mobility. It enables direct triggering of MTC devices with low mobility by MTC-IWF (MTC InterWorking Function), without involving the MME (Mobility Management Entity). Second solution defines a method for controlling and anticipating network overload in case of an event/scenario whereby a mass of messages with some common Information Elements (IE) are to be exchanged on an interface between two nodes. The network overload control is achieved via dynamic creation of a profile characterizing the event/scenario and the common IEs.

Home networks. In-home wireless networks are now wide-spreading as today's home network is composed of at least one wireless network. The dramatic increase of traffic in such networks yields to difficulties in guaranteeing user experience especially for some specific services like IPTV. This is particularly complicated when using UDP at transport layer and traditional MAC protocol at link layer. Therefore, we investigated comparison of different combinations of transport and link layer performances for the delivery of IPTV. For validation, we use NS-3 and a realistic propagation model generated with a real house description. We analyze impact of link layer (with or without coordination) and transport layer (UDP or TCP). Then, we propose a combined solution using TCP over a coordinated MAC protocol (see [52]). The proposed solution can be easily deployed in real products and is compatible with existing devices.

Another part of our activities in wireless network are related to energy saving. Indeed, one of the biggest problem today in the wireless world is that wireless devices are battery driven, which reduce their operating lifetime. The experimental measurements we have achieved in [18], [42] revealed that operating system overhead causes a drop in performance and energy consumption properties as compared to the GPP in case of certain low video qualities. We propose, thus, a new approach for energy-aware processor switching (GPP or DSP) which takes into consideration the video quality. We show the pertinence of our solution in the context of adaptive video decoding and implement it on an embedded Linux operating system.

6.4. Future Networks

Participants: Yassine Hadjadj-Aoul, Adlen Ksentini, Leila Ghazzai, Jean-Michel Sanner.

Mobile cloud. One of the 5G-architecture visions considers the usage of cloud to build mobile networks and help in decentralizing mobile networks on demand, elastically, and in the most cost-efficient way. This concept of carrier cloud becomes of vital importance knowing that several cloud providers are distributing their cloud/network, globally deploying more regional data centers, to meet their ever-increasing business demands. As an important enabler of the carrier cloud concept, network function virtualization (NFV) is gaining great momentum among industries. NFV aims for decoupling the software part from the hardware part of a carrier network node, traditionally referring to a dedicated hardware, single service and single-tenant box, and that is using virtual hardware abstraction. Network functions become thus a mere code, runnable on a particular, preferably any, operating system and on top of a dedicated hardware platform. The ultimate objective is to run network functions as software in standard virtual machines (VMs) on top of a virtualization platform in a general-purpose multi-service multi-tenant node (e.g., Carrier Grade Blade Server) put into the cloud. In [31], we presented and detailed the Follow Me Cloud (FMC) concept, whereby mobile services hosted in federated clouds follow mobile users as they move and according to their needs. We then provided in [55] a detailed analytical model based on continuous time Markov chain which considers to evaluate the performance of FMC in terms of service migration cost and QoS gain for user. An efficient mobile cloud cannot be built without efficient algorithms for the placement of NFV over this federated cloud. In this vein, in [57] we argued the need for avoiding or minimizing the frequency of mobility gateway (S-GW) relocations and discussed how this gateway relocation avoidance can be reflected in an efficient network function placement algorithm for the realization of mobile cloud. The problem was modeled by an Integer Linear Problem and proved to be NP hard. Therefore, two heuristics were proposed for the creation of a NFV S-GW instance in the cloud.

SDN. We started an activity on Software Defined Networking (SDN), a recent idea proposed to handle network management problems. SDN are becoming an important issue with the ever-increasing network complexity. They are proposed as an alternative to the current architecture of the Internet, which cannot meet the supported services requirements such as Quality of Service/Experience (Qos/QoE), security and energy consumption. We particularly address the scalability issue by proposing a hierarchical controller-based architecture handling the whole control chain.

6.5. Interoperability assessment and improvement

Participants: César Viho, Anthony Baire, Nanxing Chen.

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. Therefore, in the recent period, we propose an interoperability testing methodology using a *passive* approach. It appeared more suitable for this distributed, unrliable and constrained environment brought by IoT. We have also developed a tool that implements this passive method. It has been used successfully to test CoAP implementations during the two CoAP Plugtest interoperability sessions on IoT protocols (CoAP and 6LoWPAN) organized by ETSI and IPSO Alliance. These contributions are published in [10].

6.6. Performance Evaluation of Distributed Systems

Participants: Bruno Sericola, Romaric Ludinard.

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 [38] and [60] 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. These results also led to the patent [70].

Robustness Analysis of Large Scale Distributed Systems. In the continuation of previous work which proposed an in-depth study of the dynamicity and robustness properties of large-scale distributed systems, in [15] we analyze 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.

Secure Uniform Sampling in Dynamic Systems. In [37], we consider the problem of achieving uniform node sampling in large scale systems in presence of a strong adversary. We first propose an omniscient strategy that processes on the fly an unbounded and arbitrarily biased input stream made of node identifiers exchanged within the system, and outputs a stream that preserves Uniformity and Freshness properties. We show through Markov chains analysis that both properties hold despite any arbitrary bias introduced by the adversary. We then propose a knowledge-free strategy and show through extensive simulations that this strategy accurately approximates the omniscient one. We also evaluate its resilience against a strong adversary by studying two representative attacks (flooding and targeted attacks). We quantify the minimum number of identifiers that the adversary must insert in the input stream to prevent uniformity. To our knowledge, such an analysis has never been proposed before.

6.7. Monte Carlo

Participants: Gerardo Rubino, Bruno Tuffin, Pablo Sartor Del Giudice.

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, for two reasins: the need in accelerating the occurrence of those events and in getting unbiased estimators of them 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. Rare event simulation has been reviewed in [22].

Multidimensional integrals. In [20], we present a versatile Monte Carlo method for estimating multidimensional integrals, with applications to rare-event probability estimation. The method combines two distinct and popular Monte Carlo simulation techniques, 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.

Static models. Static reliability analysis has been the topic of an extensive activity in the group for years. Exact evaluation of static network reliability parameters belongs to the NP-hard family and Monte Carlo simulation is therefore a relevant tool to provide estimations for them.

In [67], we first review a Recursive Variance Reduction (RVR) estimator which approaches the unreliability metric by recursively reducing the graph from the random choice of the first working link on selected cuts. We show that the method does not verify the bounded relative error (BRE) property as reliability of individual links goes to one, i.e., that the estimator is not robust in general to high reliability of links. We then propose to use the decomposition ideas of the RVR estimator in conjunction with the Importance Sampling technique.

Two new estimators are presented: the first one, called Balanced Recursive Decomposition estimator, chooses the first working link on cuts uniformly, while the second, called Zero-Variance Approximation Recursive Decomposition estimator, tries to mimic the estimator with variance zero for this technique. We show that in both cases the BRE property is verified and, moreover, that a Vanishing Relative Error property can be obtained for the Zero-Variance Approximation RVR under specific sufficient conditions. A numerical illustration of the power of the methods is provided on several benchmark networks.

The same problem is also analyzed in [19] by a novel method that exploits a generalized splitting (GS) algorithm. We show that the proposed GS algorithm can accurately estimate extremely small unreliabilities and we exhibit large examples where it performs much better than existing approaches. Remarkably, it is also flexible enough to dispense with the frequently made assumption of independent edge failures.

On the same type of model, we propose in [51] an adaptive parameterized method to approximate the zero-variance change of measure for the evaluation of static network reliability models, with links subject to failures. The method uses two rough approximations of the unreliability function, conditional on the states of any subset of links being fixed. One of these approximation, based on mincuts, under-estimates the true unknown unreliability, whereas the other one, based on minpaths, over-estimates it. Our proposed change of measure takes a convex linear combination of the two, estimates the optimal (graph-dependent) coefficient in this combination from pilot runs, and uses the resulting conditional unreliability approximation at each step of a dynamic Importance Sampling algorithm. This new scheme is more general and more flexible than a previously-proposed zero-variance approximation one, which is based on mincuts only and which was shown to be robust asymptotically when unreliabilities of individual links decrease toward zero. Our numerical examples show that the new scheme is often more efficient when the unreliabilities of the individual links are not so small but the overall unreliability is small because the system can fail in many ways. Part of these results are in the PhD [13].

In [43], we present a generalization of the above static models to cases for which the component failures are not independent. To model the dependence and also to develop effective simulation methods that estimate the system unreliability, we extend the static model into an auxiliary dynamic one where the components fail at random times, according to a Marshall-Olkin multivariate exponential distribution. We examine and compare different versions of this model and develop efficient unreliability estimation methods based on conditional Monte Carlo and on a generalized splitting methodology.

In [28], a different splitting algorithm is proposed for solving the same static problem, which is converted into a dynamic one by means of the Creation Process of Elperin, Gerbtsbakh and Lomonosov. The classic splitting technique is then applied, and the obtained results are explored through several numerical experiments. The relative error and the covering properties of the obtained estimator are particularly studied.

In [29], a generalization of the basic model is studied using Monte Carlo. The idea is that the system (the network) works when the terminal nodes are connected *by at least one path whose length is less than or equal to a given parameter* d. This is called Diameter Constrained Reliability. If the parameter d is greater than or equal to the longest path in the network (or between terminals), the problem is the classic one. The paper proposes a variance reduction technique for the estimation of the system's reliability in this setting. In [21], we analyze the particular case of $d = 2$ using exact techniques. These results are part of the thesis [14].

Finally, in [34] and [36] we made general presentations on the rare event problem in general, and on some of the team's results concerning the design of efficient techniques to analyze them.

6.8. Analytic models

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

New books about Markovian models and applications. The book [65] is the french version of the book [66]. Markov chains are a fundamental class of stochastic processes. They are the main modeling tool used in our team. They are widely used to solve problems in a large number of domains such as operations research, computer science, communication networks and manufacturing systems. The success of Markov chains is mainly due to their simplicity of use, the large number of available theoretical results and the quality

of algorithms developed for the numerical evaluation of many metrics of interest. The books present the theory of both discrete-time and continuous-time homogeneous Markov chains. They examine the explosion phenomenon, the Kolmogorov equations, the convergence to equilibrium and the passage time distributions to a state and to a subset of states. These results are applied to birth-and-death processes. A detailed study of the uniformization technique by means of Banach algebra results is also developed. This technique is used for the transient analysis of several queuing systems.

Another book entitled "Markov Chains and Dependability Theory" will be published soon by Cambridge University Press (see http://www.amazon.fr/Markov-Chains-Dependability-Theory-Gerardo/dp/1107007577/). Dependability metrics are omnipresent in every engineering field, from simple ones through to more complex measures combining performance and dependability aspects of systems. The book presents the mathematical basis of the analysis of these metrics in the most used framework, Markov models, describing both basic results and specialised techniques. It presents both discrete and continuous time Markov chains before focusing on dependability measures, which necessitate the study of Markov chains on subsets of states representing different user satisfaction levels for the modelled system. Topics covered include Markovian state lumping, analysis of sojourns on subset of states of Markov chains, analysis of most dependability metrics, fundamentals of performability analysis, and bounding and simulation techniques designed to evaluate dependability measures. The book is of interest to graduate students and researchers in all areas of engineering where the concepts of lifetime, repair duration, availability, reliability and risk are important.

Fluid models. In [53] and [44] we propose a new way of transporting video flows on a peer-to-peer architecture of the Bit-Torrent type. We analyze the performance obtained by our proposal by means of fluid views of the systems, that is, by representing them using differential equations. In [53] the basic idea is to select the downloading peers according to their progress in the downloading process: a given peer only sends chunks to other peers that are downloading at least roughly in the same "area" of the stream. The system is improved in [44] where the main resource (the available bandwidth) is distributed differently among the peers, giving some kind of priority to those nodes remaining more time connected.

In [39], we look at the problem of approximating Markovian views of the Machine Repaiman Model where life-times and repair times have Phase-type distributions, by differential equations. The machine population goes to infinity, and we analyze the properties of the limiting differential equation (once the Markovian sequence of models is properly scaled) and their relations with the initial models. In [63] we describe these results and other results concerning the same type of limiting processes, but concerning peer-to-peer networks. We discuss here the convergence aspects; the properties of the fluid models themselves are discussed in the two papers [53] and [44] mentioned before.

DYOGENE Project-Team

6. New Results

6.1. Ancillary service to the grid from deferrable loads: the case for intelligent pool pumps in Florida

Renewable energy sources such as wind and solar power have a high degree of unpredictability and time-variation, which makes balancing demand and supply challenging. One possible way to address this challenge is to harness the inherent flexibility in demand of many types of loads. In [28], we focus on pool pumps, and how they can be used to provide ancillary service to the grid for maintaining demand-supply balance. A Markovian Decision Process (MDP) model is introduced for an individual pool pump. A randomized control architecture is proposed, motivated by the need for decentralized decision making, and the need to avoid synchronization that can lead to large and detrimental spikes in demand. An aggregate model for a large number of pools is then developed by examining the mean field limit. A key innovation is an LTI-system approximation of the aggregate nonlinear model, with a scalar signal as the input and a measure of the aggregate demand as the output. This makes the approximation particularly convenient for control design at the grid level. Simulations are provided to illustrate the accuracy of the approximations and effectiveness of the proposed control approach.

6.2. Impact of Storage on the Efficiency and Prices in Real-Time Electricity Markets

In [19] we study the effect of energy-storage systems in dynamic real-time electricity markets. We consider that demand and renewable generation are stochastic, that real-time production is affected by ramping constraints, and that market players seek to selfishly maximize their profit. We distinguish three scenarios, depending on the owner of the storage system: (A) the supplier, (B) the consumer, or (C) a stand-alone player. In all cases, we show the existence of a competitive equilibrium when players are price-takers (they do not affect market prices). We further establish that under the equilibrium price process, players' selfish responses coincide with the social welfare-maximizing policy computed by a (hypothetical) social planner. We show that with storage the resulting price process is smoother than without. We determine empirically the storage parameters that maximize the players' revenue in the market. In the case of consumer-owned storage, or a stand-alone storage operator (scenarios B and C), we find that they do not match socially optimal parameters. We conclude that consumers and the stand-alone storage operator (but not suppliers) have an incentive to under-dimension their storage system. In addition, we determine the scaling laws of optimal storage parameters as a function of the volatility of demand and renewables. We show, in particular, that the optimal storage energy capacity scales as the volatility to the fourth power.

6.3. Risk-Aware SLA Negotiation

In order to assure Quality of Service (QoS) connectivity, Network Service Providers (NSPs) negotiate Service Level Agreements (SLAs). However, a committed SLA might fail to respect its QoS promises. In such a case, the customer is refunded. To maximize their revenues, the NSPs must deal with risks of SLA violations, which are correlated to their network capacities. Due to the complexity of the problem, we first study in [21], a system with one NSP provider and give a method to compute its risk-aware optimal strategy using (max; +)-algebras. Using the same method, we study the case where two NSPs collaborate and the case where they compete, and we derive the Price of Anarchy. This method provides optimal negotiation strategies but, when modeling customers' reaction to SLA failure, analytical results do not hold. Hence, we propose a learning framework that chooses the NSP risk-aware optimal strategy under failures capturing the impact of reputation. Finally, by simulation, we observe how the NSP can benefit from such a framework.

6.4. Impact of Rare Alarms on Event Correlation

Nowadays, telecommunication systems are growing more and more complex, generating a large amount of alarms that cannot be effectively managed by human operators. The problem is to detect significant combinations of alarms describing an issue in real-time. In [18], we present a powerful heuristic algorithm that constructs dependency graphs of alarm patterns. More precisely, it highlights patterns extracted from an alarm flow obtained from a learning process with a small footprint on network management system performance. This algorithm helps to detect issues in real-time by effectively delivering concise alarm patterns. Furthermore, it allows the proactive analysis of the functioning of a network by computing the general trends of this network. We evaluate our algorithm on an optical network alarm data set of an existing operator. We find similar results as the expert analysis performed for this operator by Alcatel-Lucent Customer Services.

6.5. Some Synchronization Issues in OSPF Routing

A routing protocol such as OSPF has a cyclic behavior to regularly update its view of the network topology. Its behavior is divided into periods. Each period produces a flood of network information messages. We observe a regular activity in terms of messages exchanges and filling of receive buffers in routers. [17] examines the consequences of possible overlap of activity between periods, leading to a buffer overflow. OSPF allows “out of sync" flows by considering an initial delay (phase). We study the optimum calculation of these offsets to reduce the load, while maintaining a short period to ensure a protocol reactive to topology changes. Such studies are conducted using a simulated Petri net model. A heuristic for determining initial delays is proposed. A core network in Germany serves as illustration.

6.6. Exact Worst-case Delay in FIFO-multiplexing Feed-forward Networks

In this paper we compute the actual worst-case end-to-end delay for a flow in a feed-forward network of FIFO-multiplexing service curve nodes, where flows are shaped by piecewise-affine concave arrival curves, and service curves are piecewise affine and convex. We show that the worst-case delay problem can be formulated as a mixed integer-linear programming problem, whose size grows exponentially with the number of nodes involved. Furthermore, we present approximate solution schemes to find upper and lower delay bounds on the worst-case delay. Both only require to solve just *one* linear programming problem, and yield bounds which are generally more accurate than those found in the previous work, which are computed under more restrictive assumptions.

6.7. Fast weak–KAM integrators for separable Hamiltonian systems

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.

6.8. Probabilistic cellular automata, invariant measures, and perfect sampling

A probabilistic cellular automaton (PCA) can be viewed as a Markov chain. The cells are updated synchronously and independently, according to a distribution depending on a finite neighborhood. In [9], we investigate the ergodicity of this Markov chain. A classical cellular automaton is a particular case of PCA. For a one-dimensional cellular automaton, we prove that ergodicity is equivalent to nilpotency, and is therefore undecidable. We then propose an efficient perfect sampling algorithm for the invariant measure of an ergodic PCA. Our algorithm does not assume any monotonicity property of the local rule. It is based on a bounding process which is shown to also be a PCA. Last, we focus on the PCA majority, whose asymptotic behavior is unknown, and perform numerical experiments using the perfect sampling procedure.

6.9. Density Classification on Infinite Lattices and Trees

Consider an infinite graph with nodes initially labeled by independent Bernoulli random variables of parameter p. In [7], 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 to the uniform configuration with only 0's if p<1/2, and only 1's if p>1/2. We present solutions to the problem on the regular grids of dimension d, for any d>1, and on the regular infinite trees. For the bi-infinite line, we propose some candidates that we back up with numerical simulations.

6.10. Semi-infinite paths of the radial spanning tree

In the paper [4], in collaboration with David Coupier and Viet Chi Tran of Lille 1, we study the semi-infinite paths of the radial spanning tree (RST) of a Poisson point process in the plane using Stochastic Geometry. We first show that the expectation of the number of intersection points between semi-infinite paths and the sphere with radius r grows sublinearly with r. Then, we prove that in each (deterministic) direction, there exists with probability one a unique semi-infinite path, framed by an infinite number of other semi-infinite paths of close asymptotic directions. The set of (random) directions in which there are more than one semi-infinite paths is dense in $[0, 2\pi)$. It corresponds to possible asymptotic directions of competition interfaces. We show that the RST can be decomposed in at most five infinite subtrees directly connected to the root. The interfaces separating these subtrees are studied and simulations are provided.

6.11. Generating functionals of random packing point processes

In the paper [45], we study the generating functionals of a class of random packing point processes of the Matérn type. Consider a symmetrical conflict relationship between the points of a point process. The Matérn type constructions provide a generic way of selecting a subset of this point process which is conflict-free. The simplest one consists in keeping only conflict-free points. There is however a wide class of Matérn type processes based on more elaborate selection rules and providing larger sets of selected points. The general idea being that if a point is discarded because of a given conflict, there is no need to discard other points with which it is also in conflict. The ultimate selection rule within this class is the so called Random Sequential Adsorption, where the cardinality of the sequence of conflicts allowing one to decide whether a given point is selected is *not* bounded. The present paper provides a sufficient condition on the span of the conflict relationship under which all the above point processes are well defined when the initial point process is Poisson. It then establishes, still in the Poisson case, a set of differential equations satisfied by the probability generating functionals of these Matérn type point processes. Integral equations are also given for the Palm distributions.

6.12. Clustering and percolation of point processes

We are interested in phase transitions in certain percolation models on point processes and their dependence on clustering properties of the point processes. In [5], we show that point processes with smaller void probabilities and factorial moment measures than the stationary Poisson point process exhibit non-trivial phase transition in the percolation of some coverage models based on level-sets of additive functionals of the point process. Examples of such point processes are determinantal point processes, some perturbed lattices, and more generally, negatively associated point processes. Examples of such coverage models are k-coverage in the Boolean model (coverage by at least k grains) and SINR-coverage (coverage if the signal-to-interference-and-noise ratio is large). In particular, we answer in affirmative the hypothesis of existence of phase transition in the percolation of k-faces in the C̆ech simplicial complex (also called clique percolation) on point processes which cluster less than the Poisson process. We also construct a Cox point process, which is "more clustered" than the Poisson point process and whose Boolean model percolates for arbitrarily small radius. This shows that clustering (at least, as detected by our specific tools) does not always "worsen" percolation, as well as that upper-bounding this clustering by a Poisson process is a necessary assumption for the phase transition to hold.

6.13. Using Poisson processes to model lattice cellular networks

An almost ubiquitous assumption made in the stochastic-analytic approach to study of the quality of user-service in cellular networks is Poisson distribution of base stations, often completed by some specific assumption regarding the distribution of the fading (e.g. Rayleigh). The former (Poisson) assumption is usually (vaguely) justified in the context of cellular networks, by various irregularities in the real placement of base stations, which ideally should form a lattice (e.g. hexagonal) pattern. In the first part of [14] we provide a different and rigorous argument justifying the Poisson assumption under sufficiently strong log-normal shadowing observed in the network, in the evaluation of a natural class of the typical-user service-characteristics (including path-loss, interference, signal-to-interference ratio, spectral efficiency). Namely, we present a Poisson-convergence result for a broad range of stationary (including lattice) networks subject to log-normal shadowing of increasing variance. We show also for the Poisson model that the distribution of all these typical-user service characteristics does not depend on the particular form of the additional fading distribution. Our approach involves a mapping of 2D network model to 1D image of it "perceived" by the typical user. For this image we prove our Poisson convergence result and the invariance of the Poisson limit with respect to the distribution of the additional shadowing or fading. Moreover, in the second part of the paper we present some new results for Poisson model allowing one to calculate the distribution function of the SINR in its whole domain. We use them to study and optimize the mean energy efficiency in cellular networks.

6.14. Compactification of the Action of a Point-Shift on the Palm Probability of a Point Process

In collaboration with Mir-Omid Haji-Mirsadeghi (Sharif University, Iran) [50], we analyzed the compactification of Palm probabilities by the action of a point–shift. A point-shift maps, in a translation invariant way, each point of a stationary point process Φ to some point of Φ. The initial motivation of this paper is the construction of probability measures, defined on the space of counting measures with an atom at the origin, which are left invariant by a given point-shift f. The point-shift probabilities of Φ are defined from the action of the semigroup of point-shift translations on the space of Palm probabilities, and more precisely from the compactification of the orbits of this semigroup action. If the point-shift probability is uniquely defined, and if f is continuous with respect to the vague topology, then the point-shift probability of Φ provides a solution to the initial question. Point-shift probabilities are shown to be a strict generalization of Palm probabilities: when the considered point-shift f is bijective, the point-shift probability of Φ boils down to the Palm probability of Φ. When it is not bijective, there exist cases where the point-shift probability of Φ is the law of Φ under the Palm probability of some stationary thinning Ψ of Φ. But there also exist cases where the point-shift probability of Φ is singular w.r.t. the Palm probability of Φ and where, in addition, it cannot be the law of Φ under the Palm probability of any stationary point process Ψ jointly stationary with Φ. The paper also gives a criterium of existence of the point-shift probabilities of a stationary point process and discusses uniqueness. The results are illustrated through several examples.

6.15. A Stochastic Geometry Framework for Analyzing Pairwise-Cooperative Cellular Networks

Cooperation in cellular networks has been recently suggested as a promising scheme to improve system performance, especially for cell-edge users. In [34], we use stochastic geometry to analyze cooperation models where the positions of Base Stations (BSs) follow a Poisson point process distribution and where Voronoi cells define the planar areas associated with them. For the service of each user, either one or two BSs are involved. If two, these cooperate by exchange of user data and channel related information with conferencing over some backhaul link. Our framework generally allows variable levels of channel information at the transmitters. In this paper we investigate the case of limited channel state information for cooperation (channel phase, second neighbour interference), but not the fully adaptive case which would require considerable feedback. The total per-user transmission power is further split between the two transmitters and a common message is encoded. The decision for a user to choose service with or without cooperation is directed by a family of

geometric policies depending on its relative position to its two closest base stations. An exact expression of the network coverage probability is derived. Numerical evaluation allows one to analyze significant coverage benefits compared to the non-cooperative case. As a conclusion, cooperation schemes can improve system performance without exploitation of extra network resources.

6.16. SINR-based k-coverage probability in cellular networks with arbitrary shadowing

In [20], we give numerically tractable, explicit integral expressions for the distribution of the signal-to-interference-and-noise-ratio (SINR) experienced by a typical user in the down-link channel from the k-th strongest base stations of a cellular network modelled by Poisson point process on the plane. Our signal propagation-loss model comprises of a power-law path-loss function with arbitrarily distributed shadowing, independent across all base stations, with and without Rayleigh fading. Our results are valid in the whole domain of SINR, in particular for $SINR < 1$, where one observes multiple coverage. In this latter aspect our paper complements previous studies reported in [55].

6.17. Equivalence and comparison of heterogeneous cellular networks

In [15], we consider a general heterogeneous network in which, besides general propagation effects (shadowing and/or fading), individual base stations can have different emitting powers and be subject to different parameters of Hata-like path-loss models (path-loss exponent and constant) due to, for example, varying antenna heights. We assume also that the stations may have varying parameters of, for example, the link layer performance (SINR threshold, etc). By studying the *propagation processes* of signals received by the typical user from all antennas marked by the corresponding antenna parameters, we show that seemingly different heterogeneous networks based on Poisson point processes can be equivalent from the point of view a typical user. These neworks can be replaced with a model where all the previously varying propagation parameters (including path-loss exponents) are set to constants while the only trade-off being the introduction of an isotropic base station density. This allows one to perform analytic comparisons of different network models via their isotropic representations. In the case of a constant path-loss exponent, the isotropic representation simplifies to a homogeneous modification of the constant intensity of the original network, thus generalizing a previous result showing that the propagation processes only depend on one moment of the emitted power and propagation effects. We give examples and applications to motivate these results and highlight an interesting observation regarding random path-loss exponents.

6.18. How user throughput depends on the traffic demand in large cellular networks: a typical cell analysis and real network measurements

In [40], we assume a space-time Poisson process of call arrivals on the infinite plane, independently marked by data volumes and served by a cellular network modeled by an infinite ergodic point process of base stations. Each point of this point process represents the location of a base station that applies a processor sharing policy to serve users arriving in its vicinity, modeled by the Voronoi cell, possibly perturbed by some random signal propagation effects. User service rates depend on their signal-to-interference-and-noise ratios with respect to the serving station. Little's that allows to express the mean user throughput in any region of this network model as the ratio of the mean traffic demand to the steady-state mean number of users in this region. Using ergodic arguments and the Palm theoretic formalism, we define a global mean user throughput in the cellular network and prove that it is equal to the ratio of mean traffic demand to the mean number of users in the steady state of the "typical cell" of the network. Here, both means account for double averaging: over time and network geometry, and can be related to the per-surface traffic demand, base-station density and the spatial distribution of the signal-to-interference-and-noise ratio. This latter accounts for network irregularities, shadowing and cell dependence via some cell-load equations. Inspired by the analysis of the typical cell, we propose also a simpler, approximate, but fully analytic approach, called the mean cell approach. The key quantity explicitly calculated in this approach is the cell load. In analogy to the load factor of the (classical) M/G/1 processor

sharing queue, it characterizes the stability condition, mean number of users and the mean user throughput. We validate our approach comparing analytical and simulation results for Poisson network model to real-network measurements.

6.19. Analysis of a Proportionally Fair and Locally Adaptive spatial Aloha in Poisson Networks

The proportionally fair sharing of the capacity of a Poisson network using Spatial-Aloha leads to closed-form performance expressions in two extreme cases: (1) the case without topology information, where the analysis boils down to a parametric optimization problem leveraging stochastic geometry; (2) the case with full network topology information, which was recently solved using shot-noise techniques. In [37], we show that there exists a continuum of adaptive controls between these two extremes, based on local stopping sets, which can also be analyzed in closed form. We also show that these control schemes are implementable, in contrast to the full information case which is not. As local information increases, the performance levels of these schemes are shown to get arbitrarily close to those of the full information scheme. The analytical results are combined with discrete event simulation to provide a detailed evaluation of the performance of this class of medium access controls.

6.20. Optimal Rate sampling in 802.11 Systems

In 802.11 systems, Rate Adaptation (RA) is a fundamental mechanism allowing transmitters to adapt the coding and modulation scheme as well as the MIMO transmission mode to the radio channel conditions, and in turn, to learn and track the (mode, rate) pair providing the highest throughput. So far, the design of RA mechanisms has been mainly driven by heuristics. In contrast, in [42], we rigorously formulate such design as an online stochastic optimisation problem. We solve this problem and present ORS (Optimal Rate Sampling), a family of (mode, rate) pair adaptation algorithms that provably learn as fast as it is possible the best pair for transmission. We study the performance of ORS algorithms in both stationary radio environments where the successful packet transmission probabilities at the various (mode, rate) pairs do not vary over time, and in non-stationary environments where these probabilities evolve. We show that under ORS algorithms, the throughput loss due to the need to explore sub-optimal (mode, rate) pairs does not depend on the number of available pairs, which is a crucial advantage as evolving 802.11 standards offer an increasingly large number of (mode, rate) pairs. We illustrate the efficiency of ORS algorithms (compared to the state-of-the-art algorithms) using simulations and traces extracted from 802.11 test-beds.

6.21. Flooding in Weighted Sparse Random Graphs

In [3], we study the impact of edge weights on distances in sparse random graphs. We interpret these weights as delays and take them as independent and identically distributed exponential random variables. We analyze the weighted flooding time defined as the minimum time needed to reach all nodes from one uniformly chosen node and the weighted diameter corresponding to the largest distance between any pair of vertices. Under some standard regularity conditions on the degree sequence of the random graph, we show that these quantities grow as the logarithm of n when the size of the graph n tends to infinity. We also derive the exact value for the prefactor. These results allow 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.

6.22. Viral Marketing On Configuration Model

In [38], we consider propagation of influence on a Configuration Model, where each vertex can be influenced by any of its neighbours but in its turn, it can only influence a random subset of its neighbours. Our (enhanced) model is described by the total degree of the typical vertex, representing the total number of its neighbours and the transmitter degree, representing the number of neighbours it is able to influence. We give a condition

involving the joint distribution of these two degrees, which if satisfied would allow with high probability the influence to reach a non-negligible fraction of the vertices, called a *big (influenced) component*, provided that the source vertex is chosen from a set of *good pioneers*. We show that asymptotically the big component is essentially the same, regardless of the good pioneer we choose, and we explicitly evaluate the asymptotic relative size of this component. Finally, under some additional technical assumption we calculate the relative size of the set of good pioneers. The main technical tool employed is the "fluid limit" analysis of the joint exploration of the configuration model and the propagation of the influence up to the time when a big influenced component is completed. This method was introduced in [59] to study the giant component of the configuration model. Using this approach we study also a reverse dynamic, which traces all the possible sources of influence of a given vertex, and which by a new "duality" relation allows to characterise the set of good pioneers.

6.23. Pioneers of Influence Propagation in Social Networks

With the growing importance of corporate viral marketing campaigns on online social networks, the interest in studies of influence propagation through networks is higher than ever. In a viral marketing campaign, a firm initially targets a small set of pioneers and hopes that they would influence a sizeable fraction of the population by diffusion of influence through the network. In general, any marketing campaign might fail to go viral in the first try. As such, it would be useful to have some guide to evaluate the effectiveness of the campaign and judge whether it is worthy of further resources, and in case the campaign has potential, how to hit upon a good pioneer who can make the campaign go viral.

In [43], we present a diffusion model developed by enriching the generalized random graph (a.k.a. configuration model) to provide insight into these questions. We offer the intuition behind the results on this model, rigorously proved in [38], and illustrate them here by taking examples of random networks having prototypical degree distributions — Poisson degree distribution, which is commonly used as a kind of benchmark, and Power Law degree distribution, which is normally used to approximate the real-world networks. On these networks, the members are assumed to have varying attitudes towards propagating the information. We analyze three cases, in particular — (1) Bernoulli transmissions, when a member influences each of its friend with probability p; (2) Node percolation, when a member influences all its friends with probability p and none with probability $1-p$; (3) Coupon-collector transmissions, when a member randomly selects one of his friends K times with replacement.

We assume that the configuration model is the closest approximation of a large online social network, when the information available about the network is very limited. The key insight offered by this study from a firm's perspective is regarding how to evaluate the effectiveness of a marketing campaign and do cost-benefit analysis by collecting relevant statistical data from the pioneers it selects. The campaign evaluation criterion is informed by the observation that if the parameters of the underlying network and the campaign effectiveness are such that the campaign can indeed reach a significant fraction of the population, then the set of good pioneers also forms a significant fraction of the population. Therefore, in such a case, the firms can even adopt the naïve strategy of repeatedly picking and targeting some number of pioneers at random from the population. With this strategy, the probability of them picking a good pioneer will increase geometrically fast with the number of tries.

6.24. Peer-to-Peer Networks

In [12], in collaboration with I. Norros (VTT, Finland) and F. Mathieu (Bell Labs), we propose a new model for peer-to-peer networking which takes the network bottlenecks into account beyond the access. This model can cope with key features of P2P networking like degree or locality constraints together with the fact that distant peers often have a smaller rate than nearby peers. Using a network model based on rate functions, we give a closed form expression of peers download performance in the system's fluid limit, as well as approximations for the other cases. Our results show the existence of realistic settings for which the average download time is a decreasing function of the load, a phenomenon that we call super-scalability.

6.25. Stability of the bipartite matching model

In [8], we consider the bipartite matching model of customers and servers introduced by Caldentey, Kaplan and Weiss (2009). Customers and servers play symmetrical roles. There are finite sets C and S of customer and server classes, respectively. Time is discrete and at each time step one customer and one server arrive in the system according to a joint probability measure μ on C$\times$ S, independently of the past. Also, at each time step, pairs of matched customers and servers, if they exist, depart from the system. Authorized em matchings are given by a fixed bipartite graph (C, S, E$\subset$ C $\times$ S). A matching policy is chosen, which decides how to match when there are several possibilities. Customers/servers that cannot be matched are stored in a buffer. The evolution of the model can be described by a discrete-time Markov chain. We study its stability under various admissible matching policies, including ML (match the longest), MS (match the shortest), FIFO (match the oldest), RANDOM (match uniformly), and PRIORITY. There exist natural necessary conditions for stability (independent of the matching policy) defining the maximal possible stability region. For some bipartite graphs, we prove that the stability region is indeed maximal for any admissible matching policy. For the ML policy, we prove that the stability region is maximal for any bipartite graph. For the MS and PRIORITY policies, we exhibit a bipartite graph with a non-maximal stability region.

6.26. Matchings on infinite graphs

Elek and Lippner (Proc. Am. Math. Soc. 138(8), 2939–2947, 2010) showed that the convergence of a sequence of bounded-degree graphs implies the existence of a limit for the proportion of vertices covered by a maximum matching. In [6], we provide a characterization of the limiting parameter via a local recursion defined directly on the limit of the graph sequence. Interestingly, the recursion may admit multiple solutions, implying non-trivial long-range dependencies between the covered vertices. We overcome this lack of correlation decay by introducing a perturbative parameter (temperature), which we let progressively go to zero. This allows us to uniquely identify the correct solution. In the important case where the graph limit is a unimodular Galton–Watson tree, the recursion simplifies into a distributional equation that can be solved explicitly, leading to a new asymptotic formula that considerably extends the well-known one by Karp and Sipser for Erdős-Rényi random graphs.

6.27. Double-hashing thresholds via local weak convergence.

A lot of interest has recently arisen in the analysis of multiple-choice "cuckoo hashing" schemes. In this context, a main performance criterion is the load threshold under which the hashing scheme is able to build a valid hashtable with high probability in the limit of large systems; various techniques have successfully been used to answer this question (differential equations, combinatorics, cavity method) for increasing levels of generality of the model. However, the hashing scheme analysed so far is quite utopic in that it requires to generate a lot of independent, fully random choices. Schemes with reduced randomness exists, such as "double hashing", which is expected to provide similar asymptotic results as the ideal scheme, yet they have been more resistant to analysis so far. In [22], we point out that the approach via the cavity method extends quite naturally to the analysis of double hashing and allows to compute the corresponding threshold. The path followed is to show that the graph induced by the double hashing scheme has the same local weak limit as the one obtained with full randomness.

6.28. Convergence of multivariate belief propagation, with applications to cuckoo hashing and load balancing

[23] is motivated by two applications, namely generalizations of cuckoo hashing, a computationally simple approach to assigning keys to objects, and load balancing in content distribution networks, where one is interested in determining the impact of content replication on performance. These two problems admit a common abstraction: in both scenarios, performance is characterized by the maximum weight of a generalization of a matching in a bipartite graph, featuring node and edge capacities. Our main result is a law of large numbers characterizing the asymptotic maximum weight matching in the limit of large bipartite random graphs, when

the graphs admit a local weak limit that is a tree. This result specializes to the two application scenarios, yielding new results in both contexts. In contrast with previous results, the key novelty is the ability to handle edge capacities with arbitrary integer values. An analysis of belief propagation algorithms (BP) with multivariate belief vectors underlies the proof. In particular, we show convergence of the corresponding BP by exploiting monotonicity of the belief vectors with respect to the so-called upshifted likelihood ratio stochastic order. This auxiliary result can be of independent interest, providing a new set of structural conditions which ensure convergence of BP.

6.29. Bypassing correlation decay for matchings with an application to XORSAT

Many combinatorial optimization problems on sparse graphs do not exhibit the correlation decay property. In such cases, the cavity method remains a sophisticated heuristic with no rigorous proof. In [24], we consider the maximum matching problem which is one of the simplest such example. We show that monotonicity properties of the problem allows us to define solutions for the cavity equations. More importantly, we are able to identify the 'right' solution of these equations and then to compute the asymptotics for the size of a maximum matching. The results for finite graphs are self-contained. We give references to recent extensions making use of the notion of local weak convergence for graphs and the theory of unimodular networks.

As an application, we consider the random XORSAT problem which according to the physics literature has a 'one-step replica symmetry breaking' (1RSB) glass phase. We derive new bounds on the satisfiability threshold valid for general graphs (and conjectured to be tight).

6.30. Sublinear-Time Algorithms for Monomer-Dimer Systems on Bounded Degree Graphs

For a graph G, let $Z(G,\lambda)$ be the partition function of the monomer-dimer system defined by $\sum_k m_k(G)\lambda^k$, where $m_k(G)$ is the number of matchings of size k in G. In [27], we consider graphs of bounded degree and develop a sublinear-time algorithm for estimating $\log Z(G,\lambda)$ at an arbitrary value $\lambda > 0$ within additive error ϵn with high probability. The query complexity of our algorithm does not depend on the size of G and is polynomial in $1/\epsilon$, and we also provide a lower bound quadratic in $1/\epsilon$ for this problem. This is the first analysis of a sublinear-time approximation algorithm for a $\#P$-complete problem. Our approach is based on the correlation decay of the Gibbs distribution associated with $Z(G,\lambda)$. We show that our algorithm approximates the probability for a vertex to be covered by a matching, sampled according to this Gibbs distribution, in a near-optimal sublinear time. We extend our results to approximate the average size and the entropy of such a matching within an additive error with high probability, where again the query complexity is polynomial in $1/\epsilon$ and the lower bound is quadratic in $1/\epsilon$. Our algorithms are simple to implement and of practical use when dealing with massive datasets. Our results extend to other systems where the correlation decay is known to hold as for the independent set problem up to the critical activity.

6.31. Reconstruction in the Labeled Stochastic Block Model

The labeled stochastic block model is a random graph model representing networks with community structure and interactions of multiple types. In its simplest form, it consists of two communities of approximately equal size, and the edges are drawn and labeled at random with probability depending on whether their two endpoints belong to the same community or not.

It has been conjectured that this model exhibits a phase transition: reconstruction (i.e. identification of a partition positively correlated with the true partition into the underlying communities) would be feasible if and only if a model parameter exceeds a threshold.

In [25], we prove one half of this conjecture, i.e., reconstruction is impossible when below the threshold. In the converse direction, we introduce a suitably weighted graph. We show that when above the threshold by a specific constant, reconstruction is achieved by (1) minimum bisection, and (2) a spectral method combined with removal of nodes of high degree.

6.32. Spectrum Bandit Optimisation

In [26], we consider the problem of allocating radio channels to links in a wireless network. Links interact through interference, modelled as a conflict graph (i.e., two interfering links cannot be simultaneously active on the same channel). We aim at identifying the channel allocation maximizing the total network throughput over a finite time horizon. Should we know the average radio conditions on each channel and on each link, an optimal allocation would be obtained by solving an Integer Linear Program (ILP). When radio conditions are unknown a priori, we look for a sequential channel allocation policy that converges to the optimal allocation while minimizing on the way the throughput loss or *regret* due to the need for exploring sub-optimal allocations. We formulate this problem as a generic linear bandit problem, and analyze it first in a stochastic setting where radio conditions are driven by a stationary stochastic process, and then in an adversarial setting where radio conditions can evolve arbitrarily. We provide, in both settings, algorithms whose regret upper bounds outperform those of existing algorithms for linear bandit problems.

6.33. Randomized Consensus with Attractive and Repulsive Links

In [29], we study convergence properties of a randomized consensus algorithm over a graph with both attractive and repulsive links. At each time instant, a node is randomly selected to interact with a random neighbor. Depending on if the link between the two nodes belongs to a given subgraph of attractive or repulsive links, the node update follows a standard attractive weighted average or a repulsive weighted average, respectively. The repulsive update has the opposite sign of the standard consensus update. In this way, it counteracts the consensus formation and can be seen as a model of link faults or malicious attacks in a communication network, or the impact of trust and antagonism in a social network. Various probabilistic convergence and divergence conditions are established. A threshold condition for the strength of the repulsive action is given for convergence in expectation: when the repulsive weight crosses this threshold value, the algorithm transits from convergence to divergence. An explicit value of the threshold is derived for classes of attractive and repulsive graphs. The results show that a single repulsive link can sometimes drastically change the behavior of the consensus algorithm. They also explicitly show how the robustness of the consensus algorithm depends on the size and other properties of the graphs.

6.34. Continuous-time Distributed Optimization of Homogenous Dynamics

This paper explores the fundamental properties of distributed minimization of a sum of functions with each function only known to one node, and a pre-specified level of node knowledge and computational capacity. We define the optimization information each node receives from its objective function, the neighboring information each node receives from its neighbors, and the computational capacity each node can take advantage of in controlling its state. It is proven that there exist a neighboring information way and a control law that guarantee global optimal consensus if and only if the solution sets of the local objective functions admit a nonempty intersection set for fixed strongly connected graphs. Then we show that for any tolerated error, we can find a control law that guarantees global optimal consensus within this error for fixed, bidirectional, and connected graphs under mild conditions. For time-varying graphs, we show that optimal consensus can always be achieved as long as the graph is uniformly jointly strongly connected and the nonempty intersection condition holds. The results illustrate that nonempty intersection for the local optimal solution sets is a critical condition for successful distributed optimization for a large class of algorithms.

6.35. Two-target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards

In [16], we consider an infinite-armed bandit problem with Bernoulli rewards. The mean rewards are independent, uniformly distributed over $[0, 1]$. Rewards 0 and 1 are referred to as a success and a failure, respectively. We propose a novel algorithm where the decision to exploit any arm is based on two successive targets, namely, the total number of successes until the first failure and until the first m failures, respectively, where m is a fixed parameter. This two-target algorithm achieves a long-term average regret in $\sqrt{2n}$ for a

large parameter m and a known time horizon n. This regret is optimal and strictly less than the regret achieved by the best known algorithms, which is in $2\sqrt{n}$. The results are extended to any mean-reward distribution whose support contains 1 and to unknown time horizons. Numerical experiments show the performance of the algorithm for finite time horizons.

FUN Project-Team

5. New Results

5.1. Routing in FUN

Participants: Thierry Delot, Tony Ducrocq, Nicolas Gouvy, Nathalie Mitton, Enrico Natalizio, David Simplot-Ryl, Tahiry Razafindralambo, Dimitrios Zormpas.

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. Georouting assumes that every node is aware of its location, the one of its neighbors and of the destination(s).

In this context, we first propose the first k-anycating georouting protocol, ie in which a node wishes to send a message to k sinks in the network [13]. Then, we tried to relax some of the assumptions. For instance in [12], we introduce HECTOR which is the first position based routing protocol which relies on virtual positions, is energy-aware and guarantees the data delivery.

In [46], [21], we assume that only a part of nodes is aware of its position and proposes a hybrid approach between position-based greedy approach and traditional on-demand routing. Indeed, geographic routing protocols show good properties for WSNs. They are stateless, local and scalable. However they require that each node of the network is aware of its own position. While it may be possible to equip each node with GPS receiver, even if it is costly, there are some issues and receiving a usable GPS signal may be difficult in some situations. For these reasons, we propose a geographic routing algorithm, called HGA, able to take advantages of position informations of nodes when available but also able to continue the routing in a more traditional way if position information is not available. We show with simulations that our algorithm offers an alternative solution to classical routing algorithm (non-geographic) and offers better performances for network with a density above 25 and more than 5% of nodes are aware of their position. [46] analyses the impact of nodes topology on network performances. We show that different topologies can lead to a difference of up to 25% on delivery ratio and average route length and more than 100% on overall cost of transmissions.

In [24], [25], [26], [3], we consider that nodes are able to move by themselves and we try to take advantage of this feature to improve the network performance. In sensor networks, there is often more than one sensor which reports an event to the sink in WSN. In existing solutions, this leads to oscillation of nodes which belong to different routes and their premature death. Experiments show that the need of a routing path merge solution is high. As a response, [24], [25] introduce the first routing protocol which locates and uses paths crossing to adapt the topology to the network traffic in a fully localized way while still optimizing energy efficiency. Furthermore the protocol makes the intersection to move away from the destination, getting closer to the sources, allowing higher data aggregation and energy saving. Our approach outperforms existing solutions and extends network lifetime up to 37%.

Using nodes location, position-based routing protocols generally apply a greedy routing that makes a sensor forward data to route to one of its neighbors in the forwarding direction of the destination. If this greedy step fails, the routing protocol triggers a recovery mechanism. Such recovery mechanisms are mainly based on graph planarization and face traversal or on a tree construction. Nevertheless real-world network planarization is very difficult due to the dynamic nature of wireless links and trees are not so robust in such dynamic environments. Recovery steps generally provoke huge energy overhead with possibly long inefficient paths. In [26], we propose to take advantage of the introduction of controlled mobility to reduce the triggering of a recovery process. We propose Greedy Routing Recovery (GRR) routing protocol. GRR enhances greedy routing energy efficiency as it adapts network topology to the network activity. Furthermore GRR uses controlled mobility to relocate nodes in order to restore greedy and reduce energy consuming recovery step triggering. Simulations demonstrate that GRR successfully bypasses topology holes in more than 72% of network topologies avoid- ing calling to expensive recovery steps and reducing energy consumption while preserving network connectivity.

[31] relaxes the assumption that nodes are aware of their neighbors and considers that dynamic energy sources could be available. It introduces MEGAN (Mobility assisted Energy efficient Georouting in energy harvesting Actuator and sensor Networks), a beacon-less protocol that uses controlled mobility, and takes account of the energy consumption and the energy harvesting to select next hop. MEGAN aims at prolonging the overall network lifetime rather than reducing the energy consumption over a single path. When node s needs to send a message to the sink d, it first computes the ideal position of the forwarder node based on available and needed energy, and then broadcasts this data. Every node within the transmission range of s in the forward direction toward d will start a backoff timer. The backoff time is based on its available energy and on its distance from the ideal position. The first node whose backoff timer goes o is the forwarder node. This node informs its neighbor- hood and then moves toward the ideal position. If, on its route, it finds a good spot for energy harvesting, it will actually stop its movement and forward the original message by using MEGAN, which will run on all the intermediate nodes until the destination is reached. Simulations show that MEGAN reduces energy consumption up to 50% compared to algorithms where mobility and harvesting capabilities are not exploited.

Additionally, according to a wide range of studies, (Informatics Technologies) IT should become a key facilitator in establishing primary education, reducing mortality and supporting commercial initiatives in Least Developed Countries (LDCs). The main barrier to the development of IT services in these regions is not only the lack of communication facilities, but also the lack of consistent information systems, security procedures, economic and legal support, as well as political commitment. In [18], we propose the vision of an infrastructureless data platform well suited for the development of innovative IT services in LDCs. We propose a participatory approach, where each individual implements a small subset of a complete information system thanks to highly secure, portable and low-cost personal devices as well as opportunistic networking, without the need of any form of infrastructure. [18] reviews the technical challenges that are specific to this approach.

5.2. Self-organization

Participants: Tony Ducrocq, Nathalie Mitton, David Simplot-Ryl, Isabelle Simplot-Ryl.

Self-organization encompasses several mechanisms. 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

HELLO protocol or neighborhood discovery is essential in wireless ad hoc networks. It makes the rules for nodes to claim their existence/aliveness. In the presence of node mobility, no x optimal HELLO frequency and optimal transmission range exist to maintain accurate neighborhood tables while reducing the energy consumption and bandwidth occupation. Thus a Turnover based Frequency and transmission Power Adaptation algorithm (TFPA) is presented in [27]. The method enables nodes in mobile networks to dynamically adjust both their HELLO frequency and transmission range depending on the relative speed. In TFPA, each node monitors its neighborhood table to count new neighbors and calculate the turnover ratio. The relationship between relative speed and turnover ratio is formulated and optimal transmission range is derived according to battery consumption model to minimize the overall transmission energy. By taking advantage of the theoretical analysis, the HELLO frequency is adapted dynamically in conjunction with the transmission range to maintain accurate neighborhood table and to allow important energy savings. The algorithm is simulated and compared to other state-of-the-art algorithms. The experimental results demonstrate that the TFPA algorithm obtains high neighborhood accuracy with low HELLO frequency (at least 11% average reduction) and with the lowest energy consumption. Besides, the TFPA algorithm does not require any additional GPS-like device to estimate the relative speed for each node, hence the hardware cost is reduced.

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 [11], 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 the best of our knowledge, Local Hypocomb is the first strictly-localized, degree-bounded planar graph computed using merely 1-hop neighbor position information. We present a construction algorithm for these graphs and analyze its time complexity. Hypocomb family graphs are promising for wireless ad hoc networking. We report our numerical results on their average degree and their impact on FACE [49] routing. We discuss their potential applications and some open problems.

5.2.3. 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 [8], [22], [1], 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.

On another approach, [34] considers the Slepian-Wolf coding based data aggregation problem and the corresponding dependable clustering problem in WSN. A dependable Slepian-Wolf coding based clustering (DSWC) algorithm is proposed to provide dependable clustering against cluster-head failures. The proposed D-SWC algorithm attempts to elect a primary cluster head and a backup cluster head for each cluster member during clustering so that once a failure occurs to the primary cluster head the cluster members within the failed cluster can promptly switchover to the backup cluster head and thus recover the connectivity of the failed cluster to the data sink without waiting for the next-round clustering to be performed. Simulation results show that the DSWC algorithm can effectively increase the amount of data transmitted to the data sink as compared with an existing nondependable clustering algorithm for Slepian-Wolf coding based data aggregation in WSNs.

5.3. Controlled mobility

Participants: Milan Erdelj, Valeria Loscri, Kalypso Magklara, Karen Miranda, Enrico Natalizio, Jean Razafimandimby Anjalalaina, Tahiry Razafindralambo, David Simplot-Ryl, Dimitrios Zormpas.

Controlled mobility [5] is a new paradigm that leads to a set of great new challenges.

5.3.1. Target coverage

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 [15], 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.

[47], [28] assume that these targets to cover are dynamic. We assume that no knowledge about either event position or duration is given a priori. Nonetheless, the events need to be monitored and covered thanks to mobile wireless sensors. Thus, mobile sensors have to discover the events and move towards a new Zone of Interest (ZoI) when the previous monitored event is over. An efficient, distributed and localized solution of this problem would be immediately exploitable by several applications domains, such as environmental, civil, etc. We propose two novel approaches to deal with dynamic event coverage. The first one is a modified version of the PSO, where particles (mobile sensors, nodes or devices in the following) update their velocity by using only local information coming from their neighbors. In practice, the velocity update is performed by considering neighbors' sensed events. Our distributed version of PSO is integrated with a distributed version of the Virtual Force Algorithm (VFA). Virtual Force approach has the ability to "position" nodes with no overlap, by using attractive and repulsive forces based on the distance between nodes. The other proposed algorithm is a distributed implementation of the VFA by itself. Both techniques are able to reach high levels of coverage and show a satisfying reactivity when the ZoI changes. This output parameter is measured as the capability for the sensors to "follow" a sequence of events happening in different ZoIs. The effectiveness of our techniques is shown through a series of simulations and comparisons with the classical centralized VFA.

On another approach consists in using flying drone to cover this set of targets. [39] focuses on the energy efficiency problem where camera equipped flying drones are able to detect and follow mobile events that happen on the ground. We give a mathematical formulation of the problem of minimizing the total energy consumption of a fleet of drones when coverage of all events is required. Due to the extremely high complexity of the binary optimization problem, the optimum solution cannot be obtained even for small instances. On the contrary, we present LAS, a localized solution for the aforementioned problem which takes into account the ability of the drones to fly at lower altitudes in order to conserve energy. We simulate LAS and we compare its performance to a centralized algorithm and to an approach that uses static drones to cover all the terrain. Our findings show that LAS performs similar to the centralized algorithm, while it outperforms the static approach by up to 150% in terms of consumed energy. Finally, the simulation results show that LAS is very sustainable in presence of communication errors.

5.3.2. Multiple 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.

The problems of multiple PoI discovery, coverage and data report are still solved separately and there are no works that combine the aforementioned problems into a single deployment scheme. In [9], [2], we present a novel approach for mobile sensor deployment, where we combine multiple PoI discovery and coverage with network connectivity preservation in order to capture the dynamics of the monitored area. Furthermore, we derive analytical expressions for circular movement parameters and examine the performance of our approach through extensive simulation campaigns.

[10] addresses the problem of autonomous deployment of mobile sensors that need to cover a predefined PoI with a connectivity constraint. In our algorithm, 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.

5.3.3. Robot cooperation

The concept of autonomous mobile agents gets a lot of attention in the domain of WSN or wireless sensor and actuator networks (WSAN). Multiple robots that coordinate or cooperate with other sensors, robots or human operator, allow the WSN/WSAN to perform tasks that are far beyond the scope of single robot unit. In[23], we describe the robot middleware architecture that allows networked multi-robot control and data acquisition in the context of wireless sensor networks. Furthermore, we present three examples of robot network deployment and illustrate the proposed architecture usability: the robotic network deployment with the goal of covering the Point of Interest, adaptable multi-hop video transmission scenario, and the case of obtaining the energy consumption during the deployment.

5.3.4. Substitution networks

A substitution network [4] is a temporary network that will be deployed to support a base network in trouble and help it to provide the best service.

WSN are widely deployed nowadays on a large variety of applications. The major goal of a WSN is to collect information about a set of phenomena. Such process is non trivial since batteries' life is limited and thus wireless transmissions as well as computing operations must be minimized. A common task in WSNs is to estimate the sensed data and to spread the estimated samples over the network. Thus, time series estimation mechanisms are vital on this type of processes so as to reduce data transmission. In [30], we assume a single-hop clustering mechanism in which sensor nodes are grouped into clusters and communicate with a sink through a single hop. We propose a couple of autoregressive mechanisms to predict local sensed samples in order to reduce wireless data communication. We compare our proposal with a model called EEE that has been previously proposed in the literature. We prove the efficiency of our algorithms with real samples publicly available and show that they outperform the EEE mechanism.

In [32], we propose an algorithm to efficiently (re)-deploy the wireless mobile routers of a substitution network by considering the energy consumption, a fast deployment scheme and a mix of the network metric. We consider a scenario where we have two routers in a fixed network and where connectivity must be restored between those two routers with a wireless mobile router. The main objective of the wireless mobile router is to increase the communication performance such as the throughput by acting as relay node between the two routers of the fixed network. We present a fast, adaptive and localized approach which takes into account different network metrics such as Received Signal Strength (RSS), Round-Trip Time (RTT) and the Transmission Rate, between the wireless mobile router and the two routers of the fixed network. Our method ameliorates the performance of our previous approach from the literature by shortening the deployment time, increasing the throughput, and consuming less energy in some specific cases.

5.4. Security

Participants: Nathalie Mitton, Enrico Natalizio.

[19] deals with the energy efficient issue of cryptographic mechanisms used for secure communication between devices in wireless sensor networks. Since these devices are mainly targeted for low power consumption appliances, there is an effort for optimization of any aspects needed for regular sensor operation. On a basis of utilization of hardware cryptographic accelerators integrated in microcontrollers, this article provides the comparison between software and hardware solutions. Proposed work examines the problems and solutions for implementation of security algorithms for WSN devices. Because the speed of hardware accelerator should

be much higher than the software implementation, there are examination tests of energy consumption and validation of performance of this feature. Main contribution of the article is real testbed evaluation of the time latency and energy requirements needed for securing the communication. In addition, global evaluation for all important network communication parameters like throughput, delay and delivery ratio are also provided.

The Internet of Things (IoT) will enable objects to become active participants of everyday activities. Introducing objects into the control processes of complex systems makes IoT security very difficult to address. Indeed, the Internet of Things is a complex paradigm in which people interact with the technological ecosystem based on smart objects through complex processes. The interactions of these four IoT components, person, intelligent object, technological ecosystem, and process highlight a systemic and cognitive dimension within security of the IoT. The interaction of people with the technological ecosystem requires the protection of their privacy. Similarly, their interaction with control processes requires the guarantee of their safety. Processes must ensure their reliability and realize the objectives for which they are designed. We believe that the move towards a greater autonomy for objects will bring the security of technologies and processes and the privacy of individuals into sharper focus. Furthermore, in parallel with the increasing autonomy of objects to perceive and act on the environment, IoT security should move towards a greater autonomy in perceiving threats and reacting to attacks, based on a cognitive and systemic approach. In [33], we will analyze the role of each of the mentioned actors in IoT security and their relationships, in order to highlight the research challenges and present our approach to these issues based on a holistic vision of IoT security.

5.5. RFID

Participants: Ibrahim Amadou, Nathalie Mitton.

Mitigating reader-to-reader collisions is one of the principal challenges in a large-scale dynamic RFID system with a number of readers deployed in order to maximize the system performance (i.e., throughput, fairness and latency). In prior works, contention-based and activity scheduling medium access control (MAC) protocols are commonly used approaches to reduce such problems. Existing protocols typically perform worse in a large-scale RFID dynamic system and require more additional components or are based on unrealistic assumptions. So far, many research efforts have been made to improve the performance or the reliability of Carrier Sense Multiple Access (CSMA) techniques for Mobile Ad-Hoc Networks (MANETs) by using an adaptive Backoff schemes. In [17], we look at these well known solutions that proved their efficiency in high congestion wireless networks. We evaluate the performance and characterize these solutions when they are used to reserve the wireless channel through broadcasting message for reader-to-tag communication. Based on the application requirements, we study their capacity to mitigate collisions, the channel access latency, the average number of successful requests sent per reader and the fairness index in the context of RFID networks.

5.6. Data collection and aggregation

Participant: Nathalie Mitton.

Named Data Networking (NDN) is a new promising paradigm for content retrieval and distribution in the future Internet. NDN communication is driven by data consumers that broadcast Interest packets to require named contents. The requests are forwarded towards the source(s) by directly using content names (instead of IP addresses), while in-network caching is used to improve delivery performance. NDN shows many similarities with data-centric models defined for wireless sensor networks (WSNs), e.g., directed diffusion. In addition, NDN defines a new complete communication framework with innovative naming and security schemes and novel routing and transport strategies. This clearly opens new perspectives in the design and development of sensor networks, which can benefit of the NDN framework to better support different kinds of applications and services. In [16], we explore the potentialities of NDN applied to WSNs and propose enhanced delivery strategies inspired by the directed diffusion scheme to be deployed in the NDN framework. Performance of a plain NDN scheme and of our enhanced solution is evaluated through the ndnSIM simulator. Achieved results confirm the viability of a NDN-like approach over WSNs and the better efficiency and effectiveness of the proposed solution compared to a plain NDN.

[38] considers the Slepian-Wolf coding based energy-minimization rate allocation problem in a WSN and propose a distributed rate allocation algorithm to solve the problem. The proposed distributed algorithm is based on an existing centralized rate allocation algorithm which has a high computational complexity. To reduce the computational complexity of the centralized algorithm and make the rate allocation performable in a distributed manner, we make necessary modifications to the centralized algorithm by reducing the number of sets in calculating the average energy consumption cost and limiting the number of conditional nodes that a set can use. Simulation results show that the proposed distributed algorithm can significantly reduce the computational time when compared with the existing centralized algorithm at the cost of the overall energy consumption for data transmission and the total amount of data transmitted in the network.

5.7. VANET

Participant: Nathalie Mitton.

Routing is a critical issue in vehicular ad hoc networks (VANETs). This paper considers the routing issue in both vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communications in VANETs, and proposes a Moving dirEction and DestinAtion Location based routing (MEDAL) algorithm for supporting V2V and V2I communications. MEDAL [36] takes advantage of both the moving directions of vehicles and the destination location to select a neighbor vehicle as the next hop for forwarding data. Unlike most existing routing algorithms, it only uses a HELLO message to obtain or update routing information without using other control messages, which largely reduces the number of control messages used in routing. Simulation results show that MEDAL can significantly improve the packet delivery ratio of the network as compared with the well-known Ad hoc On-demand Distance Vector Routing (AODV) algorithm.

5.8. Industrial Applications

Participants: Milan Erdelj, Nathalie Mitton, Enrico Natalizio.

The collaborative nature of industrial wireless sensor networks (IWSNs) brings several advantages over traditional wired industrial monitoring and control systems, including self-organization, rapid deployment, flexibility, and inherent intelligent processing. In this regard, IWSNs play a vital role in creating more reliable, efficient, and productive industrial systems, thus improving companies' competitiveness in the marketplace. Industrial Wireless Sensor Networks: Applications, Protocols, and Standards [42] examines the current state of the art in industrial wireless sensor networks and outlines future directions for research.

GANG Project-Team

5. New Results

5.1. Understanding graph representations

5.1.1. Connected graph searching

5.1.1.1. Computing H-Joins with Application to 2-Modular Decomposition

Participants: Michel Habib, Antoine Mamcarz, Fabien de Montgolfier.

We present in [10], a general framework to design algorithms that compute H-join. For a given bipartite graph H, we say that a graph G admits a H-join decomposition or simply a H-join, if the vertices of G can be partitioned in |H| parts connected as in H. This graph H is a kind of pattern, that we want to discover in G. This framework allows us to present fastest known algorithms for the computation of P 4-join (aka N-join), P 5-join (aka W-join), C 6-join (aka 6-join). We also generalize this method to find a homogeneous pair (also known as 2-module), a pair M 1,M 2 such that for every vertex x$\notin$(M 1$\cup$M 2) and i$\in$1,2, x is either adjacent to all vertices in M i or to none of them. First used in the context of perfect graphs (Chvátal and Sbihi in Graphs Comb. 3:127-139, 1987), it is a generalization of splits (a.k.a. 1-joins) and of modules. The algorithmics to compute them appears quite involved. In this paper, we describe an O(mn 2)-time algorithm computing all maximal homogeneous pairs of a graph, which not only improves a previous bound of O(mn 3) for finding only one pair (Everett et al. in Discrete Appl. Math. 72:209-218, 1997), but also uses a nice structural property of homogenous pairs, allowing to compute a canonical decomposition tree for sesquiprime graphs (i.e., graphs G having no module and such that for every vertex v$\in$G, G-v also has no module).

5.1.1.2. Algorithmic Aspects of Switch Cographs

Participants: Vincent Cohen-Addad, Michel Habib, Fabien de Montgolfier.

The paper [27], introduces the notion of involution module, the first generalization of the modular decomposition of 2-structure which has a unique linear-sized decomposition tree. We derive an $O(n^2)$ decomposition algorithm and we take advantage of the involution modular decomposition tree to state several algorithmic results. Cographs are the graphs that are totally decomposable w.r.t modular decomposition. In a similar way, we introduce the class of switch cographs, the class of graphs that are totally decomposable w.r.t involution modular decomposition. This class generalizes the class of cographs and is exactly the class of (Bull, Gem, Co-Gem, C_5)-free graphs. We use our new decomposition tool to design three practical algorithms for the maximum cut, vertex cover and vertex separator problems. The complexity of these problems was still unknown for this class of graphs. This paper also improves the complexity of the maximum clique, the maximum independant set, the chromatic number and the maximum clique cover problems by giving efficient algorithms, thanks to the decomposition tree. Eventually, we show that this class of graphs has Clique-Width at most 4 and that a Clique-Width expression can be computed in linear time.

5.1.1.3. LDFS-Based Certifying Algorithm for the Minimum Path Cover Problem on Cocomparability Graphs

Participants: Derek Corneil, Dalton Barnaby, Michel Habib.

For graph $G(V, E)$, a minimum path cover (MPC) is a minimum cardinality set of vertex disjoint paths that cover V (i.e., every vertex of G is in exactly one path in the cover). This problem is a natural generalization of the Hamiltonian path problem. Cocomparability graphs (the complements of graphs that have an acyclic transitive orientation of their edge sets) are a well studied subfamily of perfect graphs that includes many popular families of graphs such as interval, permutation, and cographs. Furthermore, for every cocomparability graph G and acyclic transitive orientation of the edges of $\overline{G}$ there is a corresponding poset P_G; it is easy to see that an MPC of G is a linear extension of P_G that minimizes the bump number of P_G. Although there are directly graph-theoretical MPC algorithms (i.e., algorithms that do not rely on poset formulations) for various subfamilies of cocomparability graphs, notably interval graphs, until now all MPC algorithms for

cocomparability graphs themselves have been based on the bump number algorithms for posets. In this paper [5], we present the first directly graph-theoretical MPC algorithm for cocomparability graphs; this algorithm is based on two consecutive graph searches followed by a certifying algorithm. Surprisingly, except for a lexicographic depth first search (LDFS) preprocessing step, this algorithm is identical to the corresponding algorithm for interval graphs. The running time of the algorithm is $O(\min(n^2, n + \text{mloglogn}))$, with the nonlinearity coming from LDFS.

5.1.1.4. Easy identification of generalized common and conserved nested intervals

Participants: Fabien de Montgolfier, Mathieu Raffinot, Irena Rusu.

In the paper [28], we explain how to easily compute gene clusters, formalized by classical or generalized nested common or conserved intervals, between a set of K genomes represented as K permutations. A $b-$nested common (resp. conserved) interval I of size $|I|$ is either an interval of size 1 or a common (resp. conserved) interval that contains another b-nested common (resp. conserved) interval of size at least $|I| - b$. When $b = 1$, this corresponds to the classical notion of nested interval. We exhibit two simple algorithms to output all b-nested common or conserved intervals between K permutations in $O(Kn + \text{nocc})$ time, where nocc is the total number of such intervals. We also explain how to count all b-nested intervals in O(Kn) time. New properties of the family of conserved intervals are proposed to do so.

5.1.1.5. On computing the diameter of real-world undirected graphs

Participants: Pierluigi Crescenzi, Roberto Grossi, Michel Habib, Leonardo Lanzi, Andrea Marino.

We propose in [2],a new algorithm for the classical problem of computing the diameter of undirected unweighted graphs, namely, the maximum distance among all the pairs of nodes, where the distance of a pair of nodes is the number of edges contained in the shortest path connecting these two nodes. Although its worst-case complexity is O(nm) time, where n is the number of nodes and m is the number of edges of the graph, we experimentally show that our algorithm works in O(m) time in practice, requiring few breadth-first searches to complete its task on almost 200 real-world graphs.

5.1.1.6. Toward more localized local algorithms: removing assumptions concerning global knowledge

Participants: Amos Korman, Jean-Sébastien Sereni, Laurent Viennot.

Numerous sophisticated local algorithm were suggested in the literature for various fundamental problems. Notable examples are the MIS and $(\Delta + 1)$-coloring algorithms by Barenboim and Elkin, by Kuhn, and by Panconesi and Srinivasan, as well as the $o(\Delta^2)$-coloring algorithm by Linial. Unfortunately, most known local algorithms (including, in particular, the aforementioned algorithms) are *non-uniform*, that is, they assume that all nodes know good estimations of one or more global parameters of the network, e.g., the maximum degree Δ or the number of nodes n. This paper [11], provides a rather general method for transforming a non-uniform local algorithm into a *uniform* one. Furthermore, the resulting algorithm enjoys the same asymptotic running time as the original non-uniform algorithm. Our method applies to a wide family of both deterministic and randomized algorithms. Specifically, it applies to almost all of the state of the art non-uniform algorithms regarding MIS and Maximal Matching, as well as to many results concerning the coloring problem. (In particular, it applies to all aforementioned algorithms.) To obtain our transformations we introduce a new distributed tool called *pruning* algorithms, which we believe may be of independent interest.

5.1.2. Self-organizing Flows in 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 this paper [24], 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.

5.2. Large Scale Networks Performance and Modeling

5.2.1. Can P2P Networks be Super-Scalable?

Participants: François Baccelli, Fabien Mathieu, Ilkka Norros, Rémi Varloot.

We propose in [14], a new model for peer-to-peer networking which takes the network bottlenecks into account beyond the access. This model can cope with key features of P2P networking like degree or locality constraints together with the fact that distant peers often have a smaller rate than nearby peers. Using a network model based on rate functions, we give a closed form expression of peers download performance in the system's fluid limit, as well as approximations for the other cases. Our results show the existence of realistic settings for which the average download time is a decreasing function of the load, a phenomenon that we call super-scalability.

5.2.2. Contenu généré par les utilisateurs : une étude sur DailyMotion

Participants: Yannick Carlinet, The Dang Huynh, Bruno Kauffmann, Fabien Mathieu, Ludovic Noirie, Sébastien Tixeuil.

Actuellement, une large part du trafic Internet vient de sites de "User-Generated Content" (UGC). Comprendre les caractéristiques de ce trafic est important pour les opérateurs (dimensionnement réseau), les fournisseurs (garantie de la qualité de service) et les équipementiers (conception d'équipements adaptés). Dans ce contexte, nous proposons [15], d'analyser et de modéliser des traces d'usage du site DailyMotion.

5.2.3. Rumor Spreading in Random Evolving Graphs

Participants: Andrea Clementi, Pierluigi Crescenzi, Carola Doerr, Pierre Fraigniaud, Isopi Marco, Alessandro Panconesi, Pasquale Francesco, Silvestri Riccardo.

In [13], we aim at analyzing the classical information spreading "push" protocol in *dynamic* networks. We consider the *edge-Markovian* evolving graph model which captures natural temporal dependencies between the structure of the network at time t, and the one at time $t+1$. Precisely, a non-edge appears with probability p, while an existing edge dies with probability q. In order to fit with real-world traces, we mostly concentrate our study on the case where $p = \Omega(\frac{1}{n})$ and q is constant. We prove that, in this realistic scenario, the "push" protocol does perform well, completing information spreading in $O(\log n)$ time steps, w.h.p., even when the network is, w.h.p., disconnected at every time step (e.g., when $p \ll \frac{\log n}{n}$). The bound is tight. We also address other ranges of parameters p and q (e.g., $p+q=1$ with arbitrary p and q, and $p = \Theta\left(\frac{1}{n}\right)$ with arbitrary q). Although they do not precisely fit with the measures performed on real-world traces, they can be of independent interest for other settings. The results in these cases confirm the positive impact of dynamism.

5.3. Complexity issues in distributed graph algorithms

5.3.1. What can be decided locally without identifiers?

Participants: Pierre Fraigniaud, Mika Göös, Amos Korman, Jukka Suomela.

Do unique node identifiers help in deciding whether a network G has a prescribed property P? We study this question in the context of distributed local decision, where the objective is to decide whether $G \in P$ by having each node run a constant-time distributed decision algorithm. If $G \in P$, all the nodes should output yes; if $G \notin P$, at least one node should output no. A recent work (Fraigniaud et al., OPODIS 2012) studied the role of identifiers in local decision and gave several conditions under which identifiers are not needed. In this article [21], we answer their original question. More than that, we do so under all combinations of the following two critical variations on the underlying model of distributed computing: (B): the size of the identifiers is bounded by a function of the size of the input network; as opposed to ($\neg B$): the identifiers are unbounded. (C): the nodes run a computable algorithm; as opposed to ($\neg C$): the nodes can compute any, possibly uncomputable function. While it is easy to see that under ($\neg B, \neg C$) identifiers are not needed, we show that under all other combinations there are properties that can be decided locally if and only if identifiers are present. Our constructions use ideas from classical computability theory.

5.3.2. Local Distributed Decision

Participants: Pierre Fraigniaud, Amos Korman, David Peleg.

A central theme in distributed network algorithms concerns understanding and coping with the issue of locality. Inspired by sequential complexity theory, we focus on a complexity theory for distributed decision problems. In the context of locality, solving a decision problem requires the processors to independently inspect their local neighborhoods and then collectively decide whether a given global input instance belongs to some specified language. Our paper [7], introduces several classes of distributed decision problems, proves separation among them and presents some complete problems. More specifically, we consider the standard LOCAL model of computation and define LD (for local decision) as the class of decision problems that can be solved in constant number of communication rounds. We first study the intriguing question of whether randomization helps in local distributed computing, and to what extent. Specifically, we define the corresponding randomized class BPLD, and ask whether LD=BPLD. We provide a partial answer to this question by showing that in many cases, randomization does not help for deciding hereditary languages. In addition, we define the notion of local many-one reductions, and introduce the (nondeterministic) class NLD of decision problems for which there exists a certificate that can be verified in constant number of communication rounds. We prove that there exists an NLD-complete problem. We also show that there exist problems not in NLD. On the other hand, we prove that the class NLD#n, which is NLD assuming that each processor can access an oracle that provides the number of nodes in the network, contains all (decidable) languages. For this class we provide a natural complete problem as well.

5.3.3. Locality and checkability in wait-free computing

Participants: Pierre Fraigniaud, Sergio Rajsbaum, Travers Corentin.

The paper [9],studies notions of locality that are inherent to the specification of distributed tasks, and independent of the computing model, by identifying fundamental relationships between the various scales of computation, from the individual process to the whole system. A locality property called *projection-closed* is identified. This property completely characterizes tasks that are wait-free *checkable*, where a task $T = (\mathcal{I}, \mathcal{O}, \Delta)$ is said to be checkable if there exists a distributed algorithm that, given $s \in \mathcal{I}$ and $t \in \mathcal{O}$, determines whether $t \in \Delta(s)$, i.e., whether t is a valid output for s according to the specification of T. Projection-closed tasks are proved to form a rich class of tasks. In particular, determining whether a projection-closed task is wait-free solvable is shown to be undecidable. A stronger notion of locality is identified by considering tasks whose outputs "look identical" to the inputs at every process: a task $T = (\mathcal{I}, \mathcal{O}, \Delta)$ is said to be *locality-preserving* if $\mathcal{O}$ is a covering complex of $\mathcal{I}$. We show that this topological property yields obstacles for wait-free solvability different in nature from the classical impossibility results. On the other hand, locality-preserving tasks are projection-closed, and thus they are wait-free checkable. A classification of locality-preserving tasks in term of their relative computational power is provided. This is achieved by defining a correspondence between subgroups of the *edgepath* group of an input complex and locality-preserving tasks. This correspondence enables to demonstrate the existence of hierarchies of locality-preserving tasks, each one containing, at the top, the universal task (induced by the universal covering complex), and, at the bottom, the trivial identity task.

5.3.4. Delays Induce an Exponential Memory Gap for Rendezvous in Trees

Participants: Pierre Fraigniaud, Pelc Andrzej.

The aim of rendezvous in a graph is meeting of two mobile agents at some node of an unknown anonymous connected graph. In this paper [8], we focus on rendezvous in trees, and, analogously to the efforts that have been made for solving the exploration problem with compact automata, we study the size of memory of mobile agents that permits to solve the rendezvous problem deterministically. We assume that the agents are identical, and move in synchronous rounds. We first show that if the delay between the starting times of the agents is *arbitrary*, then the lower bound on memory required for rendezvous is $\Omega(\log n)$ bits, even for the line of length n. This lower bound meets a previously known upper bound of $O(\log n)$ bits for rendezvous in arbitrary graphs of size at most n. Our main result is a proof that the amount of memory needed for rendezvous *with simultaneous start* depends essentially on the number ℓ of leaves of the tree, and is exponentially less impacted by the number n of nodes. Indeed, we present two identical agents with $O(\log \ell + \log\log n)$ bits of memory that solve the rendezvous problem in all trees with at most n nodes and at most ℓ leaves. Hence, for the class of trees with polylogarithmically many leaves, there is an exponential gap in minimum memory size needed for rendezvous between the scenario with arbitrary delay and the scenario with delay zero. Moreover, we show that our upper bound is optimal by proving that $\Omega(\log \ell + \log\log n)$ bits of memory are required for rendezvous, even in the class of trees with degrees bounded by 3.

5.3.5. On the Manipulability of Voting Systems: Application to Multi-Operator Networks

Participants: François Durand, Fabien Mathieu, Ludovic Noirie.

Internet is a large-scale and highly competitive economic ecosystem. In order to make fair decisions, while preventing the economic actors from manipulating the natural outcome of the decision process, game theory is a natural framework, and voting systems represent an interesting alternative that, to our knowledge, has not yet being considered. They allow competing entities to decide among different options. In this paper [20], we investigate their use for end-to-end path selection in multi-operator networks, analyzing their manipulability by tactical voting and their economic efficiency.We show that Instant Runoff Voting is much more efficient and resistant to tactical voting than the natural system which tries to get the economic optimum.

5.4. Communication and Fault Tolerance in Distributed Networks

5.4.1. Linear Space Bootstrap Communication Schemes

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Eli Gafni, Sergio Rajsbaum.

We consider in [18], a system of n processes with ids not a priori known, that are drawn from a large space, potentially unbounded. How can these n processes communicate to solve a task? We show that n a priori allocated Multi-Writer Multi-Reader (MWMR) registers are both needed and sufficient to solve any read-write wait free solvable task. This contrasts with the existing possible solution borrowed from adaptive algorithms that require $\Theta(n^2)$ MWMR registers. To obtain these results, the paper shows how the processes can non blocking emulate a system of n Single-Writer Multi-Reader (SWMR) registers on top of n MWMR registers. It is impossible to do such an emulation with $n-1$ MWMR registers. Furthermore, we want to solve a sequence of tasks (potentially infinite) that are sequentially dependent (processes need the previous task's outputs in order to proceed to the next task). A non blocking emulation might starve a process forever. By doubling the space complexity, using $2n-1$ rather than just n registers, the computation is wait free rather than non blocking.

5.4.2. Black Art: Obstruction-Free k-set Agreement with |MWMR registers| < |proccesses|

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Eli Gafni, Sergio Rajsbaum.

When n processes communicate by writing to and reading from $k < n$ MWMR registers the "communication bandwidth" precludes emulation of SWMR system, even non-blocking.

Nevertheless, recently a positive result was shown that such a system either wait-free or obstruction-free can solve an interesting one-shot task. This paper demonstrates another such result. It shows that $(n-1)-$set agreement can be solved obstruction-free with merely 2 MWMR registers. Achieving $k-$set agreement with $n-k+1$ registers is a challenge. In [17], we make the first step toward it by showing $k-$set agreement with $2(n-k)$ registers.

5.4.3. Adaptive Register Allocation with a Linear Number of Registers

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Eli Gafni, Leslie Lamport.

In [16], we give an adaptive algorithm in which processes use multi-writer multi- reader registers to acquire exclusive write access to their own single-writer, multi-reader registers. It is the first such algorithm that uses a number of registers linear in the number of participating processes. Previous adaptive algorithms require at least $\Theta(n^{3/2})$ registers

5.4.4. Uniform Consensus with Homonyms and Omission Failures

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Hung Tran-The.

In synchronous message passing models in which some processes may be homonyms, i.e. may share the same id, we consider the consensus problem. Many results have already been proved concerning Byzantine failures in models with homonyms, we complete in [19], the picture with crash and omission failures.

Let n be the number of processes, t the number of processes that may be faulty ($t < n$) and $l(1 \leq l \leq n)$ the number of identifiers. We prove that for crash failures and send-omission failures, uniform consensus is solvable even if $l = 1$, that is with fully anonymous processes for any number of faulty processes.

Concerning omission failures, when the processes are numerate, i.e. are able to count the number of copies of identical messages they received in each round, uniform consensus is solvable even for fully anonymous processes for $n > 2t$. If processes are not numerate, uniform consensus is solvable if and only if $l > 2t$.

All the proposed protocols are optimal both in the number of communication steps needed, and in the number of processes that can be faulty.

All these results show, (1) that identifiers are not useful for crash and send-omission failures or when processes are numerate, (2) for general omission or for Byzantine failures the number of different ids becomes significant.

5.4.5. Byzantine agreement with homonyms

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Rachid Guerraoui, Anne-Marie Kermarrec, Hung Tran-The.

So far, the distributed computing community has either as- sumed that all the processes of a distributed system have distinct identifiers or, more rarely, that the processes are anonymous and have no identifiers. These are two extremes of the same general model: namely, n processes use l dif- ferent authenticated identifiers, where $1 \leq l \leq n$. In this paper [3], we ask how many identifiers are actually needed to reach agreement in a distributed system with t Byzantine processes. We show that having $3t + 1$ identifiers is necessary and sufficient for agreement in the synchronous case but, more sur- prisingly, the number of identifiers must be greater than $(n + 3t)/2$ in the partially synchronous case. This demonstrates two differences from the classical model (which has l = n): there are situations where relaxing synchrony to partial syn- chrony renders agreement impossible; and, in the partially synchronous case, increasing the number of correct processes can actually make it harder to reach agreement. The im- possibility proofs use the fact that a Byzantine process can send multiple messages to the same recipient in a round. We show that removing this ability makes agreement easier: then, $t + 1$ identifiers are sufficient for agreement, even in the partially synchronous model.

5.4.6. Byzantine agreement with homonyms in synchronous systems

Participants: Carole Delporte-Gallet, Hugues Fauconnier, Hung Tran-The.

We consider in [4], the Byzantine agreement problem 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 $l \frac{(n-r)t}{n-t-\min(t,r)}$ 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.

5.4.7. *Convergence of the D-iteration algorithm: convergence rate and asynchronous distributed scheme*

Participants: Dohy Hong, Fabien Mathieu, Gérard Burnside.

In this paper [25], we define the general framework to describe the diffusion operators associated to a positive matrix. We define the equations associated to diffusion operators and present some general properties of their state vectors. We show how this can be applied to prove and improve the convergence of a fixed point problem associated to the matrix iteration scheme, including for distributed computation framework. The approach can be understood as a decomposition of the matrix-vector product operation in elementary operations at the vector entry level.

5.5. Discrete Optimization Algorithms

5.5.1. *Shrinking Maxima, Decreasing Costs: New Online Packing and Covering Problems*

Participants: Pierre Fraigniaud, Magnús M. Halldórsson, Boaz Patt-Shamir, Dror Rawitz, Adi Rosén.

We consider in [23], two new variants of online integer programs that are duals. In the packing problem we are given a set of items and a collection of knapsack constraints over these items that are revealed over time in an online fashion. Upon arrival of a constraint we may need to remove several items (irrevocably) so as to maintain feasibility of the solution. Hence, the set of packed items becomes smaller over time. The goal is to maximize the number, or value, of packed items. The problem originates from a buffer-overflow model in communication networks, where items represent information units broken into multiple packets. The other problem considered is online covering: There is a universe to be covered. Sets arrive online, and we must decide for each set whether we add it to the cover or give it up. The cost of a solution is the total cost of sets taken, plus a penalty for each uncovered element. The number of sets in the solution grows over time, but its cost goes down. This problem is motivated by team formation, where the universe consists of skills, and sets represent candidates we may hire. The packing problem was introduced for the special case where the matrix is binary; in this paper we extend the solution to general matrices with non-negative integer entries. The covering problem is introduced in this paper; we present matching upper and lower bounds on its competitive ratio.

5.5.2. *Generalized Subdifferentials of the Sign Change Counting Function*

Participants: Dominique Fortin, Ider Tseveendorj.

A natural generalization of piecewise linear approximation of non convex problems relies on piecewise convex approximation; along the way to solve the piecewise convex maximization problem [30] both effectively and efficiently, optimality conditions have to be addressed in two ways: either the violation of necessary conditions should lead to a direction of improvement from a local solution, or a sufficient condition for global optimality has to be fulfilled. The way to either goal is paved with subdifferentials and their generalizations on a per problem basis.

In the article [29], the counting function on binary values is extended to the signed case in order to count the number of transitions between contiguous locations. A generalized subdifferential for the sign change counting function is given where classical subdifferentials remain intractable. An attempt to prove global optimality at some point, for the 4-dimensional first non trivial example, is made by using a sufficient condition specially tailored among all the cases for this subdifferential.

HIPERCOM2 Team

6. New Results

6.1. Wireless Sensor Networks

6.1.1. Node activity scheduling and routing in Wireless Sensor Networks

Participants: Cédric Adjih, Ichrak Amdouni, Pascale Minet.

The need to maximize network lifetime in wireless ad hoc networks and especially in wireless sensor networks requires the use of energy efficient algorithms and protocols. Motivated by the fact that a node consumes the least energy when its radio is in sleep state, we achieve energy efficiency by scheduling nodes activity. Nodes are assigned time slots during which they can transmit and they can turn off their radio when they are neither transmitting nor receiving. Compared to classical TDMA-based medium access scheme, spatial bandwidth use is optimized: non interfering nodes are able to share the same time slots, collisions are avoided and overhearing and interferences are reduced. In our work about time slots assignment, two cases are studied. First, when nodes require equal channel access, we use node coloring. Second, when nodes have heterogeneous traffic demands, we designed the traffic aware time slot assignment algorithm TRASA. Unlike the majority of previous works, we generalize the definition of node coloring and slot allocation problems. Indeed, we set the maximum distance between two interfering nodes as a parameter of these problems. We prove that they are NP-complete, making heuristic approaches inevitable in practice. A central directive of this thesis is to design self-adaptive solutions. This adaptivity concerns many aspects such as the mission given by the application, the heterogeneity of node traffic demands, the network density, the regularity of network topology, and the failure of wireless links.

In the GETRF project, we target the energy efficiency in wireless sensor networks. We proposed node activity scheduling approaches that determine active and inactive slots for sensor nodes as to enable them to turn off their radio and save energy in the inactive slots.

1. First, we proposed a scheduling algorithm based on node coloring of grid sensor networks called VCM. This proposal was strengthen with mathematical analysis of the optimal number of colors needed to color an infinite grid. VCM produced an optimal number of colors when the transmission range tends to infinity. Also, this algorithm does not require message exchange between sensors to determine colors.

2. Second, this work was extended to adapt it to general graphs: the graph is divided into cells and the color of the cell is the color of the node on the left bottom of the cell. Nodes inside the cell are scheduled successively.

In addition to the energy efficiency, we targeted the delay optimization for data collection applications in grid wireless sensor networks. We profit from the previous work VCM and integrate it with a new hierarchical routing method to minimize data collection delays.

6.1.2. Time slot and channel assignment in multichannel Wireless Sensor Networks

Participants: Pascale Minet, Ridha Soua, Erwan Livolant.

Applying WSNs in industrial environment requires fast and reliable data gathering (or data convergecast). If packets are forwarded individually to the sink, it is called raw data convergecast. We resort to the multichannel paradigm to enhance the data gathering delay, the robustness against interferences and the throughput. Since some appli- cations require deterministic and bounded convergecast delays, we target conflict free joint time slot and channel assignment solutions that minimize the schedule length. Such solutions allow nodes to save energy by sleeping in any slot where they are not involved in transmissions. We extend existing multichannel results to take into account a sink equipped with multiple radio interfaces and heterogeneous traffic demands. Indeed, we compute the theoretical bounds, that is the minimum number of time slots needed to complete convergecast, in various topologies with different traffic demands. These bounds are provided for different acknowledgment policies. For each of them, we provide a graph-based interference model. We also give optimal schedules that achieve these optimal bounds. We formalize the problem of multichannel slot assignment using integer linear programming and solve with GLPK tool for small configurations.

We propose MODESA, a centralized joint time slot and channel assignment algorithm. We prove the optimality of MODESA in lines, multilines and balanced trees topologies. By simulations, we show that MODESA outperforms TMCP, a well known subtree-based scheduling. We improve MODESA with different channel allocation strategies depending on the channel selection criteria (channels load balancing or preference of channels with the best qualities). Moreover, we show that resorting to multipath routing minimizes the convergecast delay. This work is extended in MUSIKA to take into account multi-sinks WSNs and traffic differentiation: the problem is formalized using integer linear programming and solved with GLPK. Simulations results show that the schedule length is minimized and the buffer size is reduced. We then address the adaptivity challenge. The slot assignment should be more flexible and able to adapt to application and environment variability (e.g., alarms, temporary additional demands). Theoretical bounds on the number of additional slots intro- duced to cope with traffic changes, are given. AMSA, an incremental solution, is proposed. Its performances are evaluated in two cases: retransmissions or temporary changes in appli- cation needs.

6.1.3. WSN Redeployment

Participants: Pascale Minet, Saoucene Ridene, Ines Khoufi.

This is a joint work with Telecom SudParis: Anis Laouiti.
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. DVFA, is our 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 DVFA gives very good coverage rate (between 98% and 100%) and ensures the connectivity between sensors.

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.

6.1.4. Opportunistic routing cross-layer schemes for low duty-cycle wireless sensor networks

Participants: Mohamed Zayani, Paul Muhlethaler.

This is a joint work with Nadjib Aitsaadi from University of Paris 12.
The opportunistic aspect of routing is suitable with such networks where the topology is dynamic and protocols based on topological information become inefficient. Previous work initiated by Paul Muhlethaler and Nadjib Aitsaadi consisted in a geographical receiver-oriented scheme based on RI-MAC protocol (Receiver-Initiated MAC). This scheme is revised and a new contribution proposes to address the same problem with a sender-oriented approach. After scrutinising different protocols belonging to this classification, the B-MAC protocol is chosen to build a new opportunistic cross-layer scheme. Our choice is motivated by the ability of this protocol to provide to a sender the closest neighbor to the destination (typically a sink). In other words, such a scheme enables us to obtain shorter paths in terms of hops which would increase the efficiency of information delivery. In counterparts, as it relies on long preambles (property of B-MAC) to solicit all the neighborhood, it needs larger delays and energy consumption (1% of active time). Nevertheless, this proposal remains interesting as the studied networks are dedicated to infrequent event detection and are not real time-oriented.

Starting from a simulator coded by Nadjib Aitsaadi for the receiver-oriented scheme, the new scheme has been coded under many variants. On top of ideal techniques, a realistic variant has been considered and modelled. Its particularity can be summarized in the election process of the next hop. Indeed, it is based on sending bursts by the potential candidates to receive a packet from a sender. These bursts express the closeness of each candidate to the destination and correspond to the binary complement of the distance to this destination.

The opportunistic cross-layer scheme, when designed with RI-MAC, has shown solid performances in carrying the information about a rare event detection to a sink. This is verified for an event detected by several nodes. Nevertheless, the efficiency of such a design becomes less obvious when the detection is performed by a very small number of nodes. the opportunistic routing using RI-MAC relies on a minor set of potential candidates to forward a packet. In other words, a sender can only select an awake neighbor (typically closer to the sink) as the next hop. To overcome this limitation, we initially proposed to limit the number of hops to reach the sink.The principle of B-MAC perfectly matches with this idea. It is also important to highlight the ability of an opportunistic cross-layer built over B-MAC to avoid collisions. B-MAC- and RI-MAC-based proposals are suitable to convey emergency packets in dense and large WSNs when the event is reported by a significant set of nodes. When this set is limited, the sake of efficiency rather suggests a scheme based on B-MAC. It should be remembered that the proposed schemes extremely limit the energy consumption compared to classical networks.

6.1.5. Data dissemination in Urban Environment

Participants: Belhaoua Asma, Nadjib Achir, Paul Muhlethaler.

Over the last decade, wireless sensor networks have brought valid solutions to real-world monitoring problems. Sensors are now incorporated in all our modern life facilities, such as mobile phones, vehicles, buses, bus stations, bikes, etc. For example, mobile phones, with their increasing capabilities are used as voice communication device but also as a sensing device able to collect data such as image, audio, GPS position, speed, etc. All these sensors could play an important role in the provisioning of a multitude of dynamic information about their environmental trends. Considering that, WSN could be considered as a valid solution to urban monitoring problems by bringing new services for the city or for the citizens. According to the last requirement, the main question that we need to answer is how the data could be collected and/or transmitted? Several algorithms were developed recently for sensor data gathering in WSN. However, the majority of existing works on WSN has focused only on specific areas applications, such as environmental monitoring, military target tracking, weather forecast, home automation, intrusion detection, etc. In this training we studied the existing strategies of dissemination in Delay/ Disruption Tolerant Networks (DTN). The main objective is to identify those that can be applied to urban environments. We implemented and tested several strategies in the WSNet network simulator on a dense network.

6.2. Cognitive Radio Networks

6.2.1. Multichannel time slot assignment in Cognitive Radio Sensor Networks

Participants: Ons Mabrouk, Pascale Minet, Ridha Soua, Ichrak Amdouni.

This is a joint work with Hanen Idoudi and Leila Saidane from ENSI, Tunisia.
Current Wireless Sensor Networks (WSNs) are deployed over unlicensed frequency bands that face an increased level of interference from various wireless systems. Cognitive Radio Sensor Networks (CRSNs) overcome this problem by allowing sensor nodes to access new spectrum bands to minimize interferences. In this paper, we focus on the MultiChannel Time Slot Assignment problem (MC-TSA) in CRSN. Each sensor node is assigned the number of time slots it needs to transfer its own data as well as the data received from its children in the rooting tree rooted at the sink without interfering with other secondary users. Besides, sensor nodes cannot transmit on a channel occupied by a primary user. Our objective is to increase the network throughput offered to sensor nodes. We start by formulating the MC-TSA problem as an Integer Linear Program where the goal is to minimize the number of slots in the schedule.We then propose an Opportunistic centralized TIme slot assignment in COgnitive Radio sensor networks (OTICOR). We evaluate its performance in terms of number of slots and throughput.

6.2.2. Leader election in Cognitive Radio Networks

Participants: Paul Muhlethaler, Dimitrios Milioris.

This is a joint work with Philippe Jacquet from Alcatel-Lucent Bell Labs.
In this sudy we have introduced a new algorithm (green election) to achieve a dis- tributed leader election in a broadcast channel that is more efficient than the classic Part-and-Try algorithm. The algorithm has the adavantage of having a reduced overhead $log(log(N))$ rather than $log(N)$. More importantly the algorithm has the a greatly reduced energy consumption since it requires $O(N^{1/k})$ burst transmissions instead of $O(N/k)$, per election, k being a parameter depending on the physical properties of the medium of communication.

One of the applications of green election is for wireless col- lision algorithms in particular in cognitive wireless networks where the secondary network is WiFi IEEE 802.11. Since the green election is low energy consuming, it can be used as a systematic and repetitive medium access control that will naturally prevail over the WiFi CSMA scheme.

6.3. Development, implementation and distribution of the Ey-Wifi module for the NS3 simulation tool

Participants: Hana Baccouch, Cédric Adjih, Paul Muhlethaler.

Ey-Wifi module is an ns-3 module developed within the Mobsim project. Ey-Wifi stands for Elimination-Yield for WiFi networks. The main goal of Ey-Wifi is to integrate the features of the EY-NPMA channel access scheme in the ns-3 Wifi module. EY-NPMA (Elimination-Yield Non-Pre-emptive Priority Multiple Access) is a contention based protocol that has been used as the medium access scheme in HIPERLAN type 1. The main advantages of EY-NPMA are: low collision rate, more determinism and priority support. EY-NPMA is based on active signaling (black burst): a node requests access to the medium by transmitting a burst signal. More precisely, the channel access cycle comprises three phases : priority phase, elimination phase and yield phase. Compared to Wifi, EY-NPMA adds the transmission of a burst in the elimination phase: it reduces the number of nodes, that will compete in next "yield" phase (equivalent to the contention window based access of WiFi).

Furthermore, the performances of Ey-Wifi have been evaluated and compared with those of Wifi with ns-3. Distribution of Ey-Wifi module: The module and a tutorial explaining how to use it, are available at: http://hipercom.inria.fr/Ey-Wifi

6.4. Mobile ad hoc and mesh networks

6.4.1. Geographic routing and location services

Participants: Selma Boumerdassi, Pascale Minet, Paul Muhlethaler.

Thanks to its scalable nature, geographic routing is an interesting alternative to topological routing for ad-hoc networks. In fact, in order to set up such a network, each node needs to know the location of the others and location services are in charge to provide such an information.

Two kinds of location services have been provided using either a flooding or a rendez-vous, a node in the network being chosen as a server for the rendez-vous. In the scope of our research, we have proposed different mechanisms based on social groups and/or communities and studied their impact on the control traffic of various protocols. For example, based on the simulations of SLS and SFLS using NS-2, we have demonstrated that the social behaviour of nodes has a strong impact on location services and therefore that next-generation location services should take the relationships between the network users into account.

6.4.2. Optimized Broadcast Scheme for Mobile Ad hoc Networks

Participants: Ahmed Amari, Nadjib Achir, Paul Muhlethaler.

In this training we propose an optimized broadcasting mechanism, which uses very limited signaling overhead. The main objective is to select the most appropriate relay nodes according to a given cost function. Basically, after receiving a broadcast packet each potential relay node computes a binary code according to a given cost function. Then, each node starts a sequence of transmit/listen intervals following this code. In other words, each 0 corresponds to a listening interval and each 1 to a transmit interval. During this active acknowledgment signaling period, each receiver applies the following rule: if it detects a signal during any of its listening intervals, it quits the selection process, since a better relay has also captured the packet. Finally, we split the transmission range into several sectors and we propose that all the nodes within the same sector use the same CDMA orthogonal spreading codes to transmit their signals. The CDMA codes used in two different sectors are orthogonal, which guarantees that the packet is broadcast in all possible directions.

6.5. Learning for an efficient and dynamic management of network resources and services

Participants: Dana Marinca, Pascale Minet.

To guarantee an efficient and dynamic management of network resources and services we intend to use a powerful mathematical tool: prediction and learning from prediction. Prediction will be concerned with guessing the short-term, average-term and long-term evolution of network or network components state, based on knowledge about the past elements and/or other available information. Basically, the prediction problem could be formulated as follows: a forecaster observes the values of one or several metrics giving indications about the network state (generally speaking the network represents the environment). At each time t, before the environment reveals the new metric values, the forecaster predicts the new values based on previous observations. Contrary to classical methods where the environment evolution is characterized by stochastic process, we suppose that the environment evolution follows an unspecified mechanism, which could be deterministic, stochastic, or even adaptive to a given behavior. The prediction process should adapt to unpredictable network state changes due to its non-stationary nature. To properly address the adaptivity challenge, a special type of forecasters is used: the experts. These experts analyse the previous environment values, apply their own computation and make their own prediction. The experts predictions are given to the forecaster before the next environment values are revealed. The forecaster can then make its own prediction depending on the experts' "advice". The risk of a prediction may be defined as the value of a loss function measuring the discrepancy between the predicted value and the real environment value. The principal notion to optimize the behavior of the forecasters is the regret, seen as a difference between the forecaster's accumulated loss and that of each expert. To optimize the prediction process means to construct a forecasting strategy that guarantees a small loss with respect to defined experts. Adaptability of the forecaster is reflected in the manner in which it is able to follow the better expert according to the context. We intend to use and improve this prediction technique to design dynamically adaptive regret matching algorithms that will be applied to dynamically manage the resources in wireless networks, especially in sensor networks. These algorithms will allow the network to choose an optimal behavior, otherwise called a correlated equilibrium, from a defined behaviors' set. This behavior will be able to evolve in time to adapt to the network context evolution. We will focus mainly but not exclusively on applications like: the choice of communication channels depending on the predicted quality of transmission, energy efficiency, network nodes deployment, efficient routing, and intelligent switching between available technologies in a multi-technology context.

6.6. Vehicular Ad hoc NETworks (VANETs) for car merging

Participant: Paul Muhlethaler.

This is a joint work with Oyunchimeg Shagdar from the IMARA team.
Cooperative Adaptive Cruise Control (CACC) systems are intended to make driving safer and more efficient by utilizing information exchange between vehicles (V2V) and/or between vehicles and infrastructures (V2I). An important application of CACC is safe vehicle merging when vehicles join a main road, achieved by compiling information on the movement of individual main road vehicles. To support such road safety applications, the IEEE standardized the 802.11p amendment dedicated to V2V and V2I communications.

In this study, we have seek answers to the questions as to whether the IEEE 802.11p can support merging control and how the communications performance is translated into the CACC performance. We have built an analytical model of the IEEE 802.11p medium access control (MAC) for transmissions of the ETSI-standardized Cooperative Awareness Messages (CAM) and Decentralized Environmental Notification Messages (DENM) to support merging control. We have also developed a highway merging decision algorithm. Using computer simulations, packet delivery ratio (PDR), and packet inter-reception (PIR) time of IEEE 802.11p-based V2V and V2I communications and their impact on the CACC performance have been investigated. Our study has disclosed several useful insights including that PIR and throughput provide a good indication of the CACC performance, while improving PDR does not necessarily enhance the CACC performance. Moreover, thanks to its ability to reliably provide information at constant time intervals, the V2I structure offers a better support for CACC than V2V.

MADYNES Project-Team

6. New Results

6.1. Android Security

Participants: Olivier Festor, Abdelkader Lahmadi [contact], Eric Finickel.

Android-based devices include smart phones 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 2013, we have designed and extended a monitoring framework integrating observed network and system activities of running Android applications. We extended and enhanced our modular NetFlow probe [48] running on android devices to export observed network flow records to a collection point for their processing and analysis. Our embedded probe includes a new set of IPFIX information elements that we have designed [41] to encapsulate geographic location information within exported flows. This work was done in collaboration with the Univerisity of Twente, where they developed the flow collector and the analysis application.

We have also developed an embedded logging probe that exports available logs generated by an Android device to a big data enabled store [25]. We have analyzed the collected logs using TreeMapping visualization technique [46] to display behavioral graphs of Android applications. The generated graphs are able to provide an aggregated view of the different components of a running application. This view is useful to improve the understanding of the behaviour of an application.

6.2. Sensor networks monitoring

Participants: Rémi Badonnel, Alexandre Boeglin, Isabelle Chrisment, Olivier Festor, Abdelkader Lahmadi [contact], Anthea Mayzaud, 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.

We developed and designed a novel algorithm and a supporting framework [16] that improves a distributed poller-pollee based 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. Results demonstrate that our approach is less aggressive and less resource consuming than its competitors.

In a previous work, we developed a first fully operational content centric networking protocol stack (CCNx) dedicated to a wireless sensor network. During this year, we have extended this implementation and designed a novel monitoring service [32] to efficiently aggregate data in a WSN. The developed solution has been implemented in the Contiki operating system and evaluated using the Cooja simulator. We have compared the performance of our proposed solution with the SPIN protocol in terms of the number of exchanged messages and response times. Our results show that our solution provides better performance for collecting and aggregating data inside the network using operators such as maximum or average.

This year, we also analyzed security attacks against LLN networks, and more specifically those targeting the RPL routing protocol. In that context, we introduced a taxonomy in order to classify these attacks into three main categories. The attacks against resources, such as DIS flooding attacks and increased rank attacks, permit to reduce the network lifetime through the generation of fake control messages or the building of RPL loops. The attacks against the topology, such as wormhole attacks or DAO inconsistency attacks, permit the network to converge to a sub-optimal configuration or to isolate one or several nodes. Finally, attacks against network traffic, such as eaves-dropping attacks and decreased rank attacks, permit to capture and analyse large part of the RPL traffic.

Based on this taxonomy, we compared the properties of attacks and discussed methods and techniques for monitoring them. In particular, we are investigating efficient solutions for supporting security monitoring in these resource-constrained environments [17]. We considered DODAG inconsistency attacks as a first case study. Scenarios were constructed to evaluate the performance of the RPL network when such attacks are carried out. Via an implementation in Contiki, it was identified that the internal mechanism proposed by RPL, which involves ignoring packets with the appropriate IPv6 header after a fixed threshold is reached, uses an arbitrary value for the threshold. A new function that dynamically scales this threshold was developed to improve performance of the network while under attack. In addition, a comparative study between the (1) no threshold, (2) fixed threshold and (3) dynamic threshold scenarios has been performed.

6.3. Monitoring of anonymous networks

Participants: Isabelle Chrisment [contact], Olivier Festor, Juan Pablo Timpanaro.

Anonymous networks have emerged to protect the privacy of network users and to secure the data exchange over the Internet. Nevertheless, the monitoring of these networks has not been investigated very much and only few networks have been studied. Large scale monitoring on these systems allows us to understand how they behave and which type of data which is shared among users.

In 2013, we continued our research about anonymous systems, with a special focus on the I2P network [3]. The I2P network provides an abstraction layer to permit two parties to communicate in an anonymous and secure manner. This network is optimized for anonymous web hosting and anonymous file-sharing. I2P's file-sharing community is highly active, where users deploy their file-sharing applications on top of the network. I2P uses a variation of Onion routing, thus assuring the unlinkability between a user and its file-sharing application.

Current statistics service for the I2P network do not provide values about the type of applications deployed in the network nor the geographical localization of users. We conducted the first large-sale monitoring on the I2P anonymous system, characterizing users and services running on top of the network. We first designed and implemented a distributed monitoring architecture based on probes placed in the I2P's distributed hash table (I2P's netDB), which allows us to collect a vast amount of network metadata. So, our distributed monitoring architecture provides us with different insights about the I2P network.

We were able to detect the behavior of particular applications, notably their period of activity. By considering the behavior of a particular anonymous service along with a particular set of I2P users, we determined in which measure this set of users was responsible for the activity of the anonymous service. We thus conducted a correlation analysis between the behavior of I2P users from two top cities along with the behavior of anonymous file-sharing clients (I2PSnark clients) throughout a particular period of time. By applying

[3] http://i2p2.de

Pearson's correlation coefficient, we achieved a group-based characterization and we determined that the activity of users from those cities explained 38% of all detected file-sharing activity [22], [2].

Starting from our limitations to de-anonymise a particular I2P user, we studied I2P's unidirectional tunnels and the mechanism used to create these tunnels. We discovered a vulnerability in this mechanism, vulnerability which allows an attacker to detect whether a user is the last participant in an inbound tunnel. With this knowledge, we showed that it would be possible to attack an I2P's eepsite in order to de-anonymise the eepsite's operator [39].

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 2013 has been dedicated to the issue of past hidden vulnerable states [8]. Even though a known vulnerability may not be present on a current system, it could have been unknowingly active in the past providing an entry point for attacks that may still constitute a potential security threat in the present. Indeed, vulnerabilities can survive within active systems for a long period of time without being known. During this period, attackers may perform well-planned and clean attacks (e.g., stealing information) without being noticed by security entities (e.g., system administrators, intrusion detection systems, self-protection modules). Changes on the system or even its normal activity can alter or erase the remaining evidence on the current configuration. In that context, we have defined a new strategy for assessing past hidden vulnerable states. This solution is based on a mathematical model for describing and detecting unknown past security exposures and on an OVAL-based framework able to autonomously build and monitor the evolution of network devices and to outsource the assessment of their exposure in an automatic manner. We also have developed an implementation prototype that efficiently performs assessment activities over an SVN repository of IOS system images. Experimental results have confirmed the feasibility and scalability of our solution.

A second part aimed at light-weighting the vulnerability assessment process in the context of mobile devices [9]. Security activities imply a consumption of resources that should be taken to a minimum in order to maximize the performance and responsiveness of such critical environments. Sometimes users may prefer to deactivate security processes such as antivirus software instead of having a short battery lifetime. The proposed approach centralizes main logistic vulnerability assessment aspects as a service while mobile clients only need to provide the server with required data to analyze known vulnerabilities described with the OVAL language. By configuring the analysis frequency as well as the percentage of vulnerabilities to evaluate at each security assessment, our probabilistic solution permits to bound client resource allocation and also to outsource the assessment process. The strategy consists in distributing evaluation activities across time thus alleviating the workload on mobile devices, and simultaneously ensuring a complete and accurate coverage of the vulnerability dataset. This technique results in a faster assessment process, typically done in the cloud, and considerably reduces the resource allocation on the client side. A prototype of our vulnerability assessment framework for Android has been selected and presented during the demonstration session of the IEEE/IFIP IM'2013 international conference [10].

We are currently investigating new methods for remediating known vulnerabilities, formalizing the change decision problem as a satisfiability or SAT problem [27]. By specifying our vulnerability knowledge source as a logical formula, fixing those system properties we can not change and freeing those variables for which changes are available, our objective is to use a SAT solving engine for determining what changes have to be made so as to secure the system. In order to provide proactive and reactive solutions, we are interested in the concept of future state descriptions to specify how a system will look like after applying a specific change.

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 [4].

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 [12], [11].

Instead of storing all the content at every nodes on the path, MPC strategy caches only popular content. With MPC, each node counts 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.

Besides, Online Social Networks (OSN) have gained tremendous popularity on the Internet. Millions of users interact with each other through OSN such as Facebook or Twitter. New ubiquitous devices (smartphones, tablets) appeared and include functionalities to instantaneously share information through OSN. As a central component of CCN is in-network caching, the content's availability depends on several criteria such as cache strategies and replacement policies, cache size or content popularity. OSN carry extremely valuable information about users and their relationships. This knowledge can help to drastically improve the efficiency of Content Centric Networks. Thus, we propose to include social information in the design of a new caching strategy for Content Centric Networking. We designed *SACS*, a novel caching strategy for CCN based on the social information of users [28]. Our socially-aware caching strategy gives priority to content issued by Influential users and cache it pro-actively into the CCN network. We performed simulations of our caching strategy and show its ability to improve the cache performances of CCN. In addition, we implemented a prototype on PlanetLab and performed large-scale experiments. Our solution improves the caching performances of CCN by 2.5 times on real testbed.

6.6. QoS in Wireless Sensor Networks

Participants: François Despaux, Abdelkader Lahmadi, Evangelia Tsiontsiou, Kévin Roussel, Moutie Chehaider, Ye-Qiong Song [contact].

[4] http://www.ccnx.org

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. Our research on WSN QoS is thoroughly organized in four topics:

- self-adaptive MAC protocol for both QoS and energy efficiency

  By combining our two previous MAC protocols called Queue-MAC and CoSenS, we extended Queue-MAC to iQueue-MAC to support multi-hop transmission [23], [6]. iQueue-MAC provides immediate yet energy-efficient throughput enhancement for dealing with burst or heavy traffic. Combined with CSMA/CA, iQueue-MAC makes use of queue length of each sensor node and allocates suitable TDMA slots to them for packets transmission. During light traffic period, no extra slots will be allocated; iQueue-MAC acts like other low duty-cycle MACs to conserve power. While in burst or heavy traffic period, iQueue-MAC senses the build up of packet queues and dynamically schedules adequate number of slots for packet transmission. Within ANR QUASIMODO project, we have implemented iQueue-MAC on STM32W108 chips that offer IEEE 802.15.4 standard communication. We set up several real-world experimental scenarios, including a 46 nodes multi-hop test-bed for simulating a general application, and conducted numerous experiments to evaluate iQueue-MAC, in comparison with other traffic adaptive duty-cycle protocols, such as multi- channel version RI-MAC and CoSenS. Results clearly show that iQueue-MAC outperforms multi-channel version of RI-MAC and CoSenS in terms of packet delay and throughput.

- 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 lower complexity and can deal with dynamic topology changes (which is the case in duty-cycled WSN). Through intensive numerical experiments, we have shown that OC has much less complexity compared with SAMCRA, known as one of the best existing algorithms. Sub-optimal paths can be obtained with a distributed version of OC, and following this principle, a first OC-based routing protocol is implemented over Contiki rime stack on TelosB motes. Its improvement and performance evaluation, as well as its integration to uIP/RPL stack is our ongoing work.

- Systems and middleware for supporting QoS in wireless sensor networks

  For supporting new protocols implementation which require to interact with low level services (MAC, Radio drivers, hardware timers) and integration to the Internet of Things approach, we focused on the OS for WSN. Several contributions have been made available for both ContikiOS (https://github.com/contiki-os/contiki/pull/519) and RiotOS (https://github.com/RIOT-OS/RIOT/pull/408, https://github.com/RIOT-OS/RIOT/pull/459). This allows to preparing for the next step towards the implementation of iQueue-MAC on both ContikiOS and RiotOS and compare experimentally with other protocols. In parallel and as part of LAR project, we also investigated the integration of different types of WSN using a gateway to make the data access transparent following RESTful webservice through CoAP/UPD/6loWPAN [24].

- End-to-end performance in multi-hop networks

  Probabilistic end-to-end performance guarantee may be required when dealing with real-time applications. As part of ANR QUASIMODO project, we are dealing with Markov modeling of multi-hop networks running slotted CSMA/CA (beacon enabled mode of IEEE 802.15.4). One of the problem of the existing models resides in their strong assumptions that may not be directly used to assess the end-to-end delay in practice. In particular, realistic radio channel, capture effect and OS-related implementation factors are not taken into account [15], [14]. We proposed to explore a new

approach which is based on process mining to extract the Markov chain model from the execution of the protocol code.

6.7. Routing in Wireless Sensor Networks

Participants: Emmanuel Nataf [contact], Patrick-Olivier Kamgueu.

Our work on the estimation of the remaining energy inside a sensor is published in [18]. We have integrated this model in the standard routing protocol for wireless sensors networks (RPL) and compared our energy based routing against a routing plane based on the quality of transmission between sensors [30].

We have built a new model to combine together several criteria, as the remaining energy, the expected transmission rate and the hop count into one quality indicator. To achieve this, we propose to use fuzzy logic either because it is a recognized mathematical tool for combining heterogeneous data and because it can be implemented with a small memory footprint. Our work is fully integrated in the standard protocol and does not need additional messages or new protocol states.

We bought 35 sensors and deployed them in the Loria building. The goal of this deployment is manyfold :

- to build and observe a real network in a real environment;
- to provide the team with a demonstrative tool to help the understanding of our work;
- to provide the team with a testbed for other works on IoT, like the security monitoring or the QoS.

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 2013, these efforts on VoIP risk management have led the PhD defense of Oussema Dabbebi. This work has been structured into three axes [1]. The first axis concerns the automation of the risk management process in VoIP enterprise network. In this context, we have developed a mathematical model for assessing risk, a set of progressive countermeasures to counter attackers and mitigation algorithms that evaluate the risk level and takes the decision to activate a subset of countermeasures [4]. To improve our strategy, we have coupled it with an anomaly detection system based on SVM and a self-configuration mechanism which provides feedback about countermeasure efficiency. The second axis deals with the extension of our adaptive risk strategy to P2PSIP infrastructures. We have implemented a specific risk model and a dedicated set of countermeasures with respect to its peer-to-peer nature. For that, we have identified attack sources and established different threat scenarios. We have analysed the RELOAD framework and proposed trust mechanisms to address its residual attacks. Finally, the third axis focuses on VoIP services in the cloud where we have proposed a risk strategy and several strategies to deploy and apply countermeasures [5].

6.9. Pervasive Computing

Participants: Laurent Ciarletta [contact], Olivier Festor, Ye-Qiong Song, Yannick Presse, Emmanuel Nataf.

Vincent Chevrier, Thomas Navarrete Gutierrez and Julien Vaubourg (MAIA team) 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.

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're applying this work on UAvs, dynamic networks (adhoc, mesh, P2P, wireless sensors and actuators), energy-constrained / location aware services, smart grids etc.

Such systems can be seen as complex and are present everywhere in our environment: internet, electricity distribution networks, transport networks. This systems have as characteristics: a large number of autonomous entities, dynamic structures, different time and space scales and emergent phenomena.

Application domains such as Smart Spaces, Smart Cities, Smart Trasportation Systems and Smart Grid makes us sometimes use Smart* or SmartX as a generic word. Madynes is focusing on the networking aspects of such systems and on the tools to develop and assess them. We cooperate with other teams and most notably the Maia team to be able to encompass issues and research questions that combine both networking and cognitive aspects.

In 2013 we worked on the following research topics :

- Assessment and evaluation of complex systems. Continuing the work on multi-modeling and co-simulation, we have participated with the MAIA team on the development of an architecture for the control of complex systems based on multi-agent simulation, a CPS co-simulation (next item) and a Smart grid simulation tool (last item), and continue working on the AA4MM framework (Agents and artefacts for Multiple heterogeneous Models).

  A control architecture has been proposed by Tomas Navarrete, based on an "equation-free" approach. We use a multi-agent model to evaluate the global impact of local control actions before applying the most pertinent set of actions. Associated to our architecture, an experimental platform has been developed to confront the basic ideas or the architecture within the context of simulated "free-riding" phenomenon in peer to peer file exchange networks. We have demonstrated that our approach allows to drive the system to a state where most peers share files, despite given initial conditions that are supposed to drive the system to a state where no peer shares. We have also executed experiments with different configurations of the architecture to identify the different means to improve the performance of the architecture.

  This work helped us to identify [13] the key issues related to the usage of the multi-agent paradigm in the context of control of complex systems.

- 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 be-

tween 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 [29].

- Smart grids and Smart spaces are another application domain. MS4SG (cf. has given us the opportunity to link multi-simulations tools such as HLA (High Level Architecture) and FMI (Functional Mockup Interface) thanks to our AA4MM framework. We've so far successfully applied our solution to the simulation of smart appartment complex and to combining the electrical and networking part of a Smart Grid (first deliverable and first workshop with EDF R&D, Supélec and SIANI were in september 2013). A paper has also been accepted to Simutools 2014. In 2014, we will continue working on the hybrid protocols and on the UAV platform, and apply our co-simulation work to Smart Grids and other Smart* [13].

6.10. SCADA Systems Security

Participants: Olivier Festor, Abdelkader Lahmadi [contact], Bilel Saadallah.

SCADA is a term used in several industries and it stands for *Supervisory Control and Data Acquisitions*. It refers to a centralized control and monitoring system for a variety of machinery and equipment involved with many industrial activities including: power generation and distribution, transportation, nuclear plants, manufacturing processes, etc. SCADA systems use a family of network protocols (PROFINET, MODBUS, DNP3) to monitor and control these industrial activities or even our homes. SCADA systems are becoming target to different attacks exploiting traditional IT vulnerabilities, e.g. buffer overflows, script crossing, crafted network packets, or specific vulnerabilities related to control and estimation algorithms employed by control processes. Several of them are daily discovered and disclosed or remain still unknown. The most threaten accidents in SCADA networks are caused by targeted attacks, where adversaries exploit those vulnerabilities available in software or network protocols components to disturb and make damage to the physical process. Therefore, it is important to provide new methods and tools for protecting SCADA network from malicious cyber attacks targeting physical processes and infrastructures.

During the year 2013, we have firstly designed and setup a SCADA test bed [31] to be able to analyze and develop security methods for several controlled physical systems. The testbed uses a Profinet based network to control experimental real-time simulated physical processes through hardware programmable logic controllers (PLCs). Secondly, we have developed a novel methodology to automatically discover a pattern of behaviour of a running controlled system through the analysis of communication messages traveling in its control loop network. The method applies process mining techniques on the exchanged communication packets between control and feedback devices to infer a model of the controlled running system. The extracted model will be then used to build a tailored anomaly-based intrusion detection module for the studied system.

6.11. Dynamic resource allocation for network virtualization

Participants: Said Seddiki, Bilel Nefzi, Mounir Frikha, Ye-Qiong Song [contact].

The objective of this research topic is to develop different resource allocation mechanisms in Network Virtualization, for creating multiple virtual networks (VNs) from a single physical network. It is accomplished by logical segmentation of the network nodes and their physical links. Sharing resources and improving utilization are the main idea of virtualization. Finding effective solutions for the needs expressed by users without deteriorating the performance of different VNs is a research challenge. In addition, solutions should meet different performance criteria required by network infrastructure.

We proposed several approaches that aim to select substrate nodes [21] with sufficient CPU, disk, and other resources, as well as substrate links with enough spare bandwidth [19], [20]. These dynamic approaches, where online monitoring of the VN is required, allow adaptively changing the resource allocations. We have shown through simulations that the proposed approaches offer higher utilization of physical network and better managing the satisfaction of virtual networks by minimizing the packet delays inside the physical node. They also provide a fair and efficient allocation of link capacity and avoid bottlenecks. The next step is the implementation of these propositions using OPENFLOW in a software defined network.

6.12. Crowdsourcing Services

Participants: Thomas Silverston [contact], Olivier Festor, Abdelkader Lahmadi, Elian Aubry.

Nowadays cities invest more in their public services, and particularly digital ones, to improve their resident's quality of life and attract more people. Thus, new crowdsourcing services appear and they are based on contributions made by mobile users equipped with smartphones. For example, the respect of the traffic code is essential to ensure citizens' security and welfare in their city. We therefore designed CrowdOut, a new mobile crowdsourcing service for improving road safety in cities. CrowdOut allows users to report traffic offense they witness in real time and to map them on a city plan. CrowdOut has been implemented and experiments and demonstrations have been performed in the urban environment of the Grand Nancy, in France. This service allows users appropriating their urban environment with an active participation regarding the collectivity. This service also represents a tool for city administrators to help for decisions and improve their urbanization policy, or to check the impact of their policy in the city environment.

MAESTRO Project-Team

5. New Results

5.1. Network Science

Participants: Eitan Altman, Konstantin Avrachenkov, Mahmoud El Chamie, Julien Gaillard, Philippe Nain, Giovanni Neglia, Marina Sokol.

5.1.1. Epidemic models of propagation of content

In [15], E. Altman and P. Nain, in collaboration with Y. Xu (MAESTRO member at the time of submission) and A. Shwartz (Technion, Israel), focus on the propagation of content in peer-to-peer (P2P) networks. They first study the transient behavior of some P2P networks whenever information is replicated and disseminated according to epidemic-like dynamics. They then use the insight gained from the previous analysis in order to predict how efficient could measures taken against P2P networks be. They first introduce a stochastic model which extends a classical epidemic model, and characterize the P2P swarm behavior in presence of free riding peers. They then study a second model in which a peer initiates a contact with another peer chosen randomly. In both cases the network is shown to exhibit phase transitions: a small change in the parameters causes a large change in the behavior of the network. The authors show, in particular, how phase transitions affect measures of content providers against P2P networks that distribute non-authorized music, books or articles, and what is the efficiency of counter-measures. In addition, this analytic framework can be generalized to characterize the heterogeneity of cooperative peers.

5.1.2. The design of recommendation systems (RS) for social networks

Recommendation systems take advantage of products and users information in order to propose items to targeted consumers. In [50], J. Gaillard, E. Altman, M. El Bèze and E. Ethis (both from Univ. Avignon) propose a framework to overcome the usual scalability issues of nowadays systems. The system includes a dynamic adaptation to enhance the accuracy of rating predictions by applying a new similarity measure. They perform several experiments on films data from Vodkaster, showing that systems incorporating dynamic adaptation improve significantly the quality of recommendations compared to static ones.

In [51] the same authors propose new modifications of the recommendation algorithm that allow not only to present a recommendation but also to propose a list of words which appeared frequently in recommendations of other people who watched that film and who have been identified to have similar preferences, according to their opinions on common movies.

5.1.3. Network centrality measures

A class of centrality measures called betweenness centralities reflects degree of participation of edges or nodes in communication between different parts of the network. The original shortest-path betweenness centrality is based on counting shortest paths which go through a node or an edge. One of shortcomings of this metric is that it ignores the paths that might be one or two hops longer than the shortest paths, while the edges on such paths can be important for communication processes in the network. To rectify this shortcoming a current flow betweenness centrality has been proposed. Similarly to the shortest-path betweenness, it has prohibitive complexity for large size networks. In [42] K. Avrachenkov, N. Litvak (Univ. of Twente, the Netherlands), V. Medyanikov (St. Petersburg State Univ., Russia) and M. Sokol propose and analyze two regularizations of the current flow betweenness centrality, α-current flow betweenness and truncated α-current flow betweenness, which can be computed fast and correlate well with the original current flow betweenness. In particular, the new centrality measures indicate well vulnerability of a network.

5.1.4. Average consensus protocols

Information can flow in a network through communication links connecting the nodes. Not all the links have the same importance and it is common in complex networks to distinguish "weak" links/ties and "strong" ones. Depending on the specific network, the strength of a link connecting two nodes can be quantified by its transmission capacity, the inter-meeting rate between the two nodes, the level of mutual trust of the two nodes, etc.. The topology of connections and the strength of the links are two factors that affect the speed of spread of information in the network. In [63], M. El Chamie and G. Neglia in collaboration with L. Severini (student at Univ. of Nice Sophia Antipolis, France) have shown that the topology can have stronger effect on the information spread than the strength of the links. In particular, they have considered an iterative belief propagation process as in average consensus protocols where each node in the network has a certain belief (a real number) that is updated iteratively by the weighted average of the nodes' belief and the ones they connected to. They have shown by simulations on random graphs that a topological optimization can have a significant faster spread of beliefs than any weight selection optimization techniques. They have also given a 2-hop message averaging that performs faster convergence than standard algorithms.

The activity on "Reducing communication overhead of average consensus protocols", described in MAESTRO's 2012 activity report has lead to the publication [49].

5.2. Wireless Networks

Participants: Eitan Altman, Philippe Nain, Giovanni Neglia, Oussama Habachi.

5.2.1. Delay Tolerant Networks

We have pursued our study of optimal control in delay tolerant network. We studied the trade-off between delivery delay and energy consumption in a delay tolerant network in which a message (or a file) has to be delivered to each of several destinations by epidemic relaying. In addition to the destinations, there are several other nodes in the network that can assist in relaying the message. The optimal control policy was obtained in the mean-field limit of large number of mobiles by C. Singh, E. Altman, A. Kumar and R. Sundaresan in [33].

Our analysis of DTNs so far was done with mobility models in which all individuals move independently of each other. In [61]. S. Patil, M. Kumar and E. Altman have studied through simulations the multicast time in DTNs where the mobility of individuals follow dependent movement such as the one of flocking birds. This model is typical to cooperative movement and could be useful to describe a rescue team in an area hit by a disaster. We showed the impact of the parameters defining the mobility on the multicast time. If instead of broadcasting packets one first codes them (using network coding) then one can obtain substantial gain in the performance. This is shown in the case that all packets that are to be sent are available for coding before transmission. In [16], E. Altman studies in collaboration with F. de Pellegrini (CREATE-NET) and L. Sassatelli how to optimally decide on the amount of coded packets to create as a function of time in the case that the information to be coded is not available before transmission. This allows to optimize the system performance for the case of real-time traffic.

In [11], A. Ali, M. Panda, T. Chahed and E. Altman design and study a reliable transport protocol for DTNs consisting of both unicast and multicast flows. The improvement in reliability is brought in by a novel Global Selective ACKnowledgment (G-SACK) scheme and random linear network coding (RLC). The motivation for using network coding and G-SACKs comes from the observation that one should take the maximum advantage of the contact opportunities which occur quite infrequently in DTNs. Network coding and G-SACKs perform "mixing" of packet and acknowledgment information, respectively, at the contact opportunities and essentially solve the randomness and finite capacity limitations of DTNs. In contrast to earlier work on network coding in DTNs, we observe and explain the gains due to network coding even under an inter-session setting. Our results from extensive simulations of appropriately chosen "minimal" topologies quantify the gains due to each enhancement feature. In a related publication [67], A. Ali, L. Sassatelli, E. Altman and T. Chahed present an overview of theoretical background that is used for evaluating transport protocols in DTNs.

In [13], E. Altman formulates in collaboration with A. P. Azad, T. Başar (Univ. Illinois at Urbana Champain) and F. De Pellegrini (CREATE-NET) a problem where both transmission and activation of mobile terminals are controlled as a linear optimal control problem. They solve the problem by making use of this linearity in order to obtain explicit expressions for the objective function as a function of the control actions trajectories (rather than as a function of both actions and state trajectories). This allows them to compute the optimal strategies explicitly.

In [26], E. Altman studies in collaboration with D. Fiems (Ghent Univ.) a class of Markov-modulated stochastic recursive equations. This class includes multi-type branching processes with immigration as well as linear stochastic equations. Conditions are established for the existence of a stationary solution and expressions for the first two moments of this solution are found. Furthermore, the transient characteristics of the stochastic recursion are investigated: the first two moments of the transient solution are obtained as well. Finally, to illustrate the approach, the results are applied to the performance evaluation of packet forwarding in delay-tolerant mobile ad-hoc networks.

In [34], G. Neglia in collaboration with X. Zhang, H. Wang (both from Fordham Univ., Bronx, USA), J. Kurose and D. Towsley (both from Univ. of Massachusetts at Amherst, USA) has also investigated the benefits of applying Random Linear Coding (RLC) to unicast application in DTNs. Under RLC, nodes store and forward random linear combinations of packets as they encounter each other. For the case of a single group of packets originating from the same source and destined for the same destination, they have proved a lower bound on the probability that the RLC scheme achieves the minimum time to deliver the group of packets. Although RLC achieves a significant reduction in group delivery delay, it fares worse in terms of average packet delivery delay and network transmissions. When replication control is employed, RLC schemes reduce the group delivery delay without increasing the number of transmissions. In general, the benefit achieved by RLC is more significant under stringent resource (bandwidth and buffer) constraints, limited signaling, highly dynamic networks, and when it is applied to packets from same flow. For more practical settings with multiple continuous flows in the network, the researchers have shown the importance of deploying RLC schemes with a carefully tuned replication control in order to achieve reduction in average delay.

In [60], the same authors investigated the problem of determining the routing that minimizes the maximum/average delivery time or the maximum/average delivery delay for a set of packets in a deterministic Delay Tolerant Network, i.e. in a network for which all the nodes' transmission opportunities are known in advance. While the general problem with multiple sources and multiple destinations is NP-hard, they have presented a polynomial-time algorithm that can efficiently compute the optimal routing in the case of a single destination or of a single packet that needs to be routed to multiple destinations.

In [59], P. Nain in collaboration with D. Towsley (Univ. of Massachusetts at Amherst, USA), A. Bar-Noy and F. Yu (both from City Univ. of New York, USA), P. Basu (Raytheon BBN Technologies, USA), and M. P. Johnson (Univ. of California, Los Angeles, USA) consider the problem of estimating the end-to-end latency of intermittently connected paths in disruption/delay tolerant networks. While computing the time to traverse such a path may be straightforward in fixed, static networks, doing so becomes much more challenging in dynamic networks, in which the state of an edge in one timeslot (i.e., its presence or absence) is random, and may depend on its state in the previous timeslot. The authors compute the expected traversal time (ETT) for a dynamic path in a number of special cases of stochastic edge dynamics models, and for three different edge failure models, culminating in a surprisingly nontrivial yet realistic "hybrid network" setting in which the initial configuration of edge states for the entire path is known. The ETT for this "initial configuration" setting can be computed in quadratic time (as a function of path length), by an algorithm based on probability generating functions. Several linear-time upper and lower bounds on the ETT are provided and evaluated using numerical simulations.

5.2.2. Interference coordination in wireless networks

In [47], R. Combes, E. Altman and Z. Altman (Orange Labs, Issy les Moulineaux) model a LTE wireless network with Inter-Cell Interference Coordination (ICIC) at the flow level where users arrive and depart dynamically, in order to optimize quality of service indicators perceivable by users such as file transfer time

for elastic traffic. They propose an algorithm to tune the parameters of ICIC schemes automatically based on measurements. The convergence of the algorithm to a local optimum is proven, and a heuristic to improve its convergence speed is given. Numerical experiments show that the distance between local optima and the global optimum is very small, and that the algorithm is fast enough to track changes in traffic on the time scale of hours. The proposed algorithm can be implemented in a distributed way with very small signaling load.

In [46], the same authors introduce self-organizing mechanisms as control loops, and study the conditions for stability when running control loops in parallel. Based on control theory, they propose a distributed coordination mechanism to stabilize the system. In certain cases, coordination can be achieved without any exchange of information between control loops. The mechanism remains valid in the presence of noise via stochastic approximation. Instability and coordination in the context of wireless networks are illustrated with two examples. The paper is essentially concerned with linear systems, and the applicability of our results for non-linear systems is discussed.

5.2.3. Streaming over wireless

The Quality of Experience (QoE) of streaming service is often degraded by playback interruptions. To mitigate these, the media player prefetches streaming contents before starting playback, at a cost of delay. In [66], Y. Xu, S. E. Elayoubi, E. Altman and R. El-Azouzi study the QoE of streaming from the perspective of flow dynamics. First, a framework is developed for QoE when streaming users join the network randomly and leave after downloading completion. They compute the distribution of prefetching delay using partial differential equations, and the probability generating function of playout buffer starvation using ordinary differential equations. Second, they extend the framework to characterize the throughput variation caused by opportunistic scheduling at the base station in the presence of fast fading. This study reveals that the flow dynamics is the fundamental reason of playback starvation. The QoE of streaming service is dominated by the average throughput of opportunistic scheduling, while the variance of throughput has very limited impact on starvation behavior.

5.2.4. Dynamic coverage of mobile sensor networks

B. Liu (Univ. of Massachusetts at Lowell, USA), O. Dousse (Nokia Research Center, Switzerland), P. Nain, and D. Towsley (Univ. of Massachusetts at Amherst, USA) study in [30] the dynamic aspects of the coverage of a mobile sensor network resulting from continuous movement of sensors. As sensors move around, initially uncovered locations may be covered at a later time, and intruders that might never be detected in a stationary sensor network can now be detected by moving sensors. However, this improvement in coverage is achieved at the cost that a location is covered only part of the time, alternating between covered and not covered. The authors characterize area coverage at specific time instants and during time intervals, as well as the time durations that a location is covered and uncovered. They further consider the time it takes to detect a randomly located intruder and prove that the detection time is exponentially distributed. For mobile intruders, a game theoretic approach allows to derive optimal mobility strategies for both sensors and intruders. The optimal sensor strategy is to choose the direction uniformly at random between 0 and 2π. The optimal intruder strategy is to remain stationary. This solution represents a mixed strategy which is a Nash equilibrium of the zero-sum game between mobile sensors and intruders.

5.2.5. Wireless network security

The activity on “Fast and secure rendezvous protocols for mitigating control channel DoS attacks” described in MAESTRO’s 2012 activity report has lead to the publication [35].

5.3. Network Engineering Games

Participants: Eitan Altman, Konstantin Avrachenkov, Ilaria Brunetti, Julien Gaillard, Majed Haddad, Manjesh Kumar Hanawal, Alexandre Reiffers.

5.3.1. Association problem

In [32], A. Silva, in collaboration with H. Tembine, E. Altman and M. Debbah, study a non-cooperative association game where mobiles associate to Base Stations. The authors solve the problem using the theory of optimal transportation after incorporating in it the effect of network congestion. They are able to find a closed form expression for its solution. The authors also solve a global optimization problem for minimizing the total power needed by the mobile terminals over the whole network.

5.3.2. Cognitive radio

In [52] O. Habachi considers a non-cooperative Opportunistic Spectrum Access (OSA) where Secondary Users (SUs) access opportunistically the spectrum licensed for Primary Users (PUs) in TV white spaces (TVWS). As sensing licensed channels is time and energy consuming, the author considers a hierarchical Cognitive Radio (CR) architecture, where CR base stations sense a subset of the spectrum in order to locate some free frequencies. Thereafter, a SU that needs to communicate through TVWS sends a request to a CR base station for a free channel. The author models the problem using a Partially Observable Stochastic Game (POSG), and he takes into consideration the energy consumption of CR base stations and the Quality of Service of SUs. Since solving POSG optimally may require a significant amount of time and computational complexity, the author then models the OSA problem using a game theoretical approach, and proposes a symmetric Nash equilibrium solution concept. Finally, the simulations that validate the theoretical findings are provided.

In [24], J. Elias (Univ. Paris Descartes), F. Martignon (Univ. Paris Sud), L. Chen and E. Altman address the joint pricing and network selection problem in cognitive radio networks. The problem is formulated as a Stackelberg game where first the Primary and Secondary operators set the network subscription price to maximize their revenue. Then, users perform the network selection process, deciding whether to pay more for a guaranteed service, or use a cheaper, best-effort secondary network, where congestion and low throughput may be experienced. They use the Nash equilibrium concept to characterize the equilibria for the price setting game. On the other hand, a Wardrop equilibrium is used in the network selection game.

5.3.3. Cooperative games in wireless networks

We have pursued this year our new activity on cooperative games in wireless communications. We have pursued our work on coalition games and started working on the area of matching games. In [56], E. Altman, C. Hasan and J.-M. Gorce (both from Inria project-team SOCRATE) have addressed the problem of association of mobiles to base stations which can be viewed as a coalition game. They formulated the game using a stochastic geometric approach (one Poisson point process representing the base stations and another one representing the mobiles) and studied the impact of switching off base stations (for energy efficient operation).

An important class of games within cooperative games is the matching games. They have been used in stable marriage games (in which a bi-partite graph called matching is to be proposed between a group of men and women based on mutual ranking between this group). A second well-known application of matching games is the college admission problem in which students are assigned to colleges based on their preferences as well as on the preferences of the colleges. We introduced and solved two matching games in wireless communication using the theory of matching games. In [55] the same authors study a game similar to the above ones to match pairs of mobiles where one mobile serves as a relay for the other in the absence of a good direct channel to the base station. The utilities studied here are the outage probabilities. In [65], R. Vaca-Ramirez, E. Altman, J. S. Thompson and V. Ramos-Ramos propose a distributed algorithm for energy efficient virtual Multiple-input/Multiple-output coalition formation. They model cooperation as a game derived from the concept of stable marriage with incomplete lists. Single antenna devices such as mobile and relay stations cooperate in order to improve the user's and system's energy efficiency. In both problems above, the performance of the equilibrium is shown to be close to the social optimum and yet the complexity for achieving the equilibrium is only polynomial (whereas that of computing a global optimal matching is NP hard).

In [40] K. Avrachenkov, L. Cottatellucci (EURECOM) and L. Maggi (CREATE-NET, Italy) study multiple access channels whose channel coefficients follow a quasi-static Markov process on a finite set of states. The authors address the issue of allocating transmission rates to users in each time interval, such that optimality and

fairness of an allocation are preserved throughout a communication, and moreover all the users are consistently satisfied with it. First, it is shown how to allocate the rates in a global optimal fashion. The authors provide a sufficient condition for the optimal rates to fulfill some fairness criteria in a time-consistent way. Then the authors utilize the game-theoretical concepts of time consistent Core and Cooperation Maintenance. It is demonstrated that in the model the sets of rates fulfilling these properties coincide and they also coincide with the set of global optimal rate allocations. The relevance of the presented dynamic rate allocation to LTE systems is also shown.

5.3.4. Bayesian games in networking

K. Veeraruna, E. Altman, R. El-Azouzi and S. Rajesh have studied in [29] a power control problem in which a base station allocates power according to the channel state as reported by the mobiles. The paper addresses the question of how to allocate the power, given that the channel reported by some non-cooperative mobile may be unreliable. They obtain the equilibrium allocation after formulating the problem as a Bayesian game.

In [38], E. Altman and T. Jiménez consider both a cooperative as well as non-cooperative admission into an M/M/1 queue. The only information available is a signal that says whether the queue size is smaller than some value L or not. They first compute the globally optimal and the Nash equilibrium stationary policy as a function of L. They compare the performance to that of full information and of no information on the queue size. They identify the value of L that optimizes the equilibrium performance.

In [58], K. Ibrahimi, E. Altman and M. Haddad introduce a signaling game approach to power control. They consider two players named player I and player II. They assume that player I only knows his channel state without any information about the channel state of player II and vice-versa. Player I moves first and sends a signal to player II which can be accurate or distorted. Player II chooses his power control strategy based on this information and his belief about the nature of the informed player's information. In order to analyze such a model, the proposed scheme game is transformed into an equivalent 4x4 matrix game. The authors establish the existence of Nash equilibria and then derive it numerically and study its properties.

In [53], M. Haddad and E. Altman, in collaboration with P. Wiecek and H. Sidi, present a Bayesian game theoretic framework for determining the decision to which cell a given mobile user should associate in LTE two-tier Heterogeneous Networks. Users are assumed to compete to maximize their throughput by picking the best locally serving cell with respect to their own measurement, their demand and a partial statistical channel state information of other users. In particular, the authors investigate the properties of a hierarchical game, in which the macro-cell BS is a player on its own. They derive analytically the utilities related to the channel quality perceived by users to obtain the equilibria. They show in the Stackelberg formulation, how the operator, by dynamically choosing the offset about the state of the channel, can optimize its global utility while end-users maximize their individual utilities.

5.3.5. Network neutrality and collusion

Representatives of several Internet access providers have expressed their wish to see a substantial change in the pricing policies of the Internet. In particular, they would like to see content providers pay for use of the network, given the large amount of resources they use. This would be in clear violation of the "network neutrality" principle that had characterized the development of the wireline Internet. We proposed and studied possible ways of implementing such payments and of regulating their amount. M. K. Hanawal and E. Altman have pursued in [54] working on network neutrality studying various ways of collusion between an ISP and a content provider and in particular, another form of non-neutrality in which a content provider signals to an ISP information on the popularity of its content and hides this information from other ISPs. They define and compute the price of collusion and study the impact of such signalling on the ISP that is in collusion as well as on the other ones.

In the situation just described, the demand is modelled to be elastic. In contrast, in [62], A. Reiffers and E. Altman study in collaboration with Y. Hayel pricing issues in non-neutral network with non-elastic traffic. A Stackelberg equilibrium is derived and the price of collusion is computed.

Our research on network neutrality started already on 2010 with a research report [83] that has now been published in [14]. We already reported on this publication in 2011 when it became available electronically.

5.3.6. Competition over popularity in social networks

In [39] E. Altman, P. Kumar, S. Venkatramanan and A. Kumar consider a situation where several content producers send their content to some subscriber of a social network. These posts appear on the subscriber's timeline which is assumed to have finite capacity. Whenever a new post arrives to the timeline, an older post leaves it. Therefore to be visible, a source has to keep sending contents from time to time. Each source is modelled as a player in a non-cooperative game in which one trades between the utility for being visible on the timeline and the cost (or effort) for keeping sending content. This game is solved in a Markovian setting the performance measures of interest are computed.

In [37], E. Altman in cooperation with F. De Pellegrini (CREATE-NET), D. Miorandi, T. Jiménez and R. El-Azouzi study situations in which subscribers of a social network take the decision whether to access or not some content, based on the number of views that the content has. Their analysis aims at understanding the way in which information about the quality of a given content can be deduced from view counts when only part of the viewers that access the content are informed about its quality. In this paper they present a game formulation for the behavior of individuals using a mean-field model: the number of individuals is approximated by a continuum of atomless players and for which the Wardrop equilibrium is the solution concept. They derive conditions on the problem's parameters that result in the emergence of threshold equilibria policies. But they also identify some parameters in which other structures are obtained for the equilibrium behavior of individuals.

5.3.7. Evolutionary games

Evolutionary game theory is a relatively young mathematical theory that aims at formalizing in mathematical terms evolution models in biology. In recent years this paradigm has penetrated more and more into other areas such as the linguistics, economics and engineering. The current theory of evolutionary game makes an implicit assumption that the evolution is driven by selfishness of individuals who interact with each other. In mathematical terms this can be stated as "an individual equals a player in a non-cooperative game model". This assumption turns out to be quite restrictive in modeling evolution in biology. It is now more and more accepted among biologist that the evolution is driven by the selfish interests of large groups of individuals; a group may correspond for example to a whole beehive or to an ants' nest. In [43] and [71], I. Brunetti and E. Altman propose an alternative paradigm for modeling evolution where a player does not necessarily represent an interacting individual but a whole class of such individuals. In [71] in particular, they use Markov Decision Evolutionary Games (MDEG) to allow a parent and a child represent the same individual at different states. This is yet another enhancement in what we understand as a player. An important contribution is in the study of the Hawk and Dove game in these new frameworks.

In [27], M. Haddad, J. Gaillard, E. Altman and D. Fiems (Ghent Univ.) study an evolutionary game in the MDEG framework of power control. Aging is taken into account by assuming that as the battery of the mobile becomes empty, high power is not available anymore. The goal of a mobile is to use power that maximizes the amount of traffic it can transmit during its lifetime. We restrict in this work to policies that are state independent and compute the equilibrium.

5.4. Green Networking and Smart Grids

Participants: Sara Alouf, Eitan Altman, Nicaise Choungmo Fofack, Delia Ciullo, Alain Jean-Marie, Giovanni Neglia.

5.4.1. Stochastic geometry methods for wireless design issues

In [64] the issue of energy efficiency in Orthogonal Frequency-Division Multiple Access (OFDMA) wireless networks is discussed by D. Tsilimantos, J.-M. Gorce (Inria project-team SOCRATE) and E. Altman. Their interest is focused on the promising concept of base station (BS) 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 (SINR) distribution, remains the same. The optimal energy efficiency gains are then specified with the help of a simplified but yet realistic BS 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 BS power profile plays.

5.4.2. Analysis of base stations with autonomous energy supply

S. Alouf, A. Jean-Marie and D. Ciullo have started the modeling of wireless communication base stations with autonomous energy supply (solar, wind). One challenge is to account for the random and non-stationary input of energy. A second challenge is to find the correct time and space granularity of the model, so as to ensure both the practical relevance of the model and numerical tractability. The activity will be backed up by a measurement campaign on the Com4Innov platform (http://www.com4innov.com/), that will provide information on energy consumption of different traffic patterns.

5.4.3. Demand-response system

Energy demand aggregators are new actors in the energy scenario: they gather a group of energy consumers and implement a demand-response paradigm. When the energy provider needs to reduce the current energy demand on the grid, it can pay the energy demand aggregator to reduce the load by turning off some of its consumers loads or postponing their activation. Currently this operation involves only greedy energy consumers like industrial plants. In [48], [78] A. Jean-Marie and G. Neglia in collaboration with G. Di Bella, L. Giarré, M. Ippolito and I. Tinnirello (all from Univ. of Palermo, Italy) have studied the potential of aggregating a large number of small energy consumers like home users as it may happen in smart grids. In particular they have addressed the feasibility of such approach by considering which scale the aggregator should reach in order to be able to control a significant power load. The challenge of the study derives from residential users' demand being much less predictable than that of industrial plants. For this reason they have resorted to queuing theory to study analytically the problem and quantify the trade-off between load control and tolerable service delays.

5.5. Content-Oriented Systems

Participants: Sara Alouf, Konstantin Avrachenkov, Nicaise Choungmo Fofack, Delia Ciullo, Alain Jean-Marie, Philippe Nain, Giovanni Neglia, Marina Sokol.

5.5.1. Performance evaluation of hierarchical TTL-based cache networks

N. Choungmo Fofack, P. Nain and G. Neglia, together with D. Towsley (Univ. of Massachusetts at Amherst, USA) have revisited and extended the work that has appeared in [82]. They consider caches that implement an expiration-based eviction policy to manage contents in their memory. These caches are called Time-To-Live (TTL)-based caches. These TTL-based caches can be used to model caches running classical replacement policies such as Least Recently Used (LRU) and Random Replacement (RND). The main characteristic of the latter TTL-based cache models is that they (re)initialize the TTL of a content at both cache hit and cache miss. In a paper that is currently under review, the case of a network of caches where requests for each content are routed as a polytree is analyzed and a framework to evaluate the performance of such general TTL-based cache networks is proposed.

5.5.2. Modeling modern DNS caches

Motivated by the recent behavior of Domain Name System (DNS) caches that do not respect the timeout marked (by Authoritative DNS servers) on resource records, N. Choungmo Fofack and S. Alouf propose in [44] a theoretical model based on renewal arguments to describe this modern behavior. The proposed model for a cache taken in isolation is validated with real traces collected by Inria's IT service at Sophia-Antipolis at one of the Inria's DNS caches. The model of a network of caches is validated by event-driven simulations. This

study suggests that, when inter-request times have a concave cumulative distribution function, client caches (those caches that are fed directly by users requests) should keep each resource record for a constant duration (that may depend on its popularity). However, core caches should draw their timeout values for each record from a distribution which has as high coefficient of variation as possible.

5.5.3. An approximate analysis of general and heterogeneous cache networks

Jointly with M. Dehghan, D. L. Goeckel and D. Towsley (Univ. of Massachusetts at Amherst, USA), N. Choungmo Fofack proposes a simple, accurate, and computationally efficient framework to assess performance of network of caches with arbitrary topology, requests described by renewal processes, and caches running Least Recently Used (LRU), First-In First-Out (FIFO), or Random Replacement (RND) policies. Their framework is based on the characteristic time approximation of LRU, RND and FIFO caches that helps to model the latter as TTL-based caches. Classical results of the theory of (renewal) point processes (e.g. approximation of general point processes by renewal processes, thinning a renewal point process, aggregating/merging independent renewal processes) are used as well as theoretical results established in [82] and [44] on TTL-based caches (e.g. calculation of metrics of interest such hit and occupancy probabilities, characterization of miss streams).

5.5.4. Data placement

Jointly with J.-C. Bermond (Inria project-team COATI), D. Mazauric (Univ. Aix-Marseille) and J. Yu (UFV Vancouver), A. Jean-Marie has pursued the study of combinatorial designs that solve the problem of replicating optimally data over unreliable servers, with the objective of minimizing the variance of the availability of documents. In a forthcoming revision of [81], they use results from Design Theory, particularly the existence of "large triple systems" to solve multiple instances of the problem.

5.5.5. Semi-supervised learning with application to P2P systems

Semi-supervised learning methods constitute a category of machine learning methods which use labelled points together with unlabelled data to tune the classifier. The main idea of the semi-supervised methods is based on an assumption that the classification function should change smoothly over a similarity graph, which represents relations among data points. This idea can be expressed using kernels on graphs such as graph Laplacian. Different semi-supervised learning methods have different kernels which reflect how the underlying similarity graph influences the classification results. In [41] K. Avrachenkov, P. Gonçalves (Inria project-team DANTE) and M. Sokol analyze a general family of semi-supervised methods, provide insights about the differences among the methods and give recommendations for the choice of the kernel parameters and labelled points. In particular, it appears that it is preferable to choose a kernel based on the properties of the labelled points. They illustrate our general theoretical conclusions with an analytically tractable characteristic example, clustered preferential attachment model and classification of content in P2P networks.

5.6. Advances in Methodological Tools

Participants: Konstantin Avrachenkov, Alain Jean-Marie, Philippe Nain.

5.6.1. Perturbation analysis

In [21] K. Avrachenkov and J.-B. Lasserre (LAAS-CNRS) investigate the analytic perturbation of generalized inverses. Firstly the authors analyze the analytic perturbation of the Drazin generalized inverse (also known as reduced resolvent in operator theory). The approach is based on spectral theory of linear operators as well as on a new notion of group reduced resolvent. It allows one to treat regular and singular perturbations in a unified framework. The authors provide an algorithm for computing the coefficients of the Laurent series of the perturbed Drazin generalized inverse. In particular, the regular part coefficients can be efficiently calculated by recursive formulae. Finally, the authors apply the obtained results to the perturbation analysis of the Moore-Penrose generalized inverse in the real domain.

5.6.2. Markov processes

In [20] K. Avrachenkov, L. Cottatellucci (EURECOM), L. Maggi (CREATE-NET, Italy) and Y.-H. Mao (Beijing Normal Univ., China) consider both discrete-time irreducible Markov chains with circulant transition probability matrix P and continuous-time irreducible Markov processes with circulant transition rate matrix Q. In both cases they provide an expression of all the moments of the entropy mixing time. In the discrete case, they prove that all the moments of the mixing time associated with the transition probability matrix $\alpha P + (1-\alpha)P^*$ are maximum in the interval $0 \leq \alpha \leq 1$ when $\alpha = 1/2$, where P^* is the transition probability matrix of the time-reversed chain. Similarly, in the continuous case, they show that all the moments of the mixing time associated with the transition rate matrix $\alpha Q + (1-\alpha)Q^*$ are also maximum in the interval $0 \leq \alpha \leq 1$ when $\alpha = 1/2$, where Q^* is the time-reversed transition rate matrix.

In [23] K. Avrachenkov, in collaboration with A. Piunovskiy and Z. Yi (both from Univ. of Liverpool, UK), study a general homogeneous continuous-time Markov process with restarts. The process is forced to restart from a given distribution at time moments generated by an independent Poisson process. The motivation to study such processes comes from modeling human and animal mobility patterns, restart processes in communication protocols, and from application of restarting random walks in information retrieval. The authors provide a connection between the transition probability functions of the original Markov process and the modified process with restarts. Closed-form expressions for the invariant probability measure of the modified process are derived. When the process evolves on the Euclidean space there is also a closed-form expression for the moments of the modified process. The authors show that the modified process is always positive Harris recurrent and exponentially ergodic with the index equal to (or bigger than) the rate of restarts. Finally, the general results are illustrated by the standard and geometric Brownian motions.

5.6.3. Queueing theory

In [22] K. Avrachenkov, P. Nain and U. Yechiali (Tel Aviv Univ., Israel) consider two independent Poisson streams of jobs flowing into a single-server service system having a limited common buffer that can hold at most one job. If a type-i job ($i = 1, 2$) finds the server busy, it is blocked and routed to a separate type-i retrial (orbit) queue that attempts to re-dispatch its jobs at its specific Poisson rate. This creates a system with three dependent queues. Such a queueing system serves as a model for two competing job streams in a carrier sensing multiple access system. We study the queueing system using multi-dimensional probability generating functions, and derive its necessary and sufficient stability conditions while solving a Riemann-Hilbert boundary value problem. Various performance measures are calculated and numerical results are presented. In particular, numerical results demonstrate that the proposed multiple access system with two types of jobs and constant retrial rates provides incentives for the users to respect their contracts.

5.6.4. Control theory

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. Inspired by the work of A. Ruszczyński, they have considered Markov Decision problems where the metric to be optimized is a risk measure, a metric which generalizes the mathematical expectation and takes risk aversion of agents into account. For infinite-horizon, risk-averse discounted Markov Decision Processes, they have proved approximation bounds which imply the convergence of approximate rolling horizon procedures when the horizon length tends to infinity. They have also analyzed the effects of uncertainties on the transition probabilities, the cost functions and the discount factors [77].

In [17] K. Avrachenkov, U. Ayesta (LAAS-CNRS), J. Doncel (LAAS-CNRS) and P. Jacko (BCAM, Spain) address the problem of fast and fair transmission of flows in a router, which is a fundamental issue in networks like the Internet. They focus on the relaxed version of the problem obtained by relaxing the fixed buffer capacity constraint that must be satisfied at all time epoch. The relaxation allows one to reduce the multi-flow problem into a family of single-flow problems, for which one can analyze both theoretically and numerically the existence of optimal control policies of special structure. In particular, it is shown that the control can be represented by so-called index policies, but not always by threshold policies. The simulation and numerical results show that the index policy achieves a wide range of desirable properties with respect to fairness between

different TCP versions, across users with different round-trip-time and minimum buffer required to achieve full utility of the queue.

5.6.5. Game theory

In [18] K. Avrachenkov, L. Cottatellucci (EURECOM) and L. Maggi (CREATE-NET, Italy) consider simple Markovian games, in which several states succeed each other over time, following an exogenous discrete-time Markov chain. In each state, a different simple static game is played by the same set of players. The authors investigate the approximation of the Shapley-Shubik power index in simple Markovian games (SSM). The authors prove that an exponential number of queries on coalition values is necessary for any deterministic algorithm even to approximate SSM with polynomial accuracy. Motivated by this, the authors propose and study three randomized approaches to compute a confidence interval for SSM. They rest upon two different assumptions, static and dynamic, about the process through which the estimator agent learns the coalition values. Such approaches can also be utilized to compute confidence intervals for the Shapley value in any Markovian game. The proposed methods require a number of queries, which is polynomial in the number of players in order to achieve a polynomial accuracy.

In [19] K. Avrachenkov, L. Cottatellucci (EURECOM) and L. Maggi (CREATE-NET, Italy) study multi-agent Markov decision processes (MDPs) in which cooperation among players is allowed. They find a cooperative payoff distribution procedure (MDP-CPDP) that distributes in the course of the game the payoff that players would earn in the long run game. They show under which conditions such a MDP-CPDP fulfills a time consistency property, contents greedy players, and strengthen the coalition cohesiveness throughout the game. Finally, the authors refine the concept of Core for Cooperative MDPs.

RAP Project-Team

4. New Results

4.1. Algorithms: Bandwidth Allocation in Optical Networks

Participants: Christine Fricker, Jelena Pesic, 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. 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).

Our work on access networks proposed an original dynamic bandwidth allocation (DBA) algorithm and demonstrated its excellent performance. This DBA algorithm was then adapted to a meshed metropolitan network based on TWIN and implementing flow-aware resource sharing. Extensions using a concept called “multipath” were shown to offer an energy efficient solution for wide area networks.

In 2013, we contributed to the Celtic Plus project called SASER/SAVENET. 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 responsibility for the work package (where alternative optical network architectures are also evaluated) is with Alcatel-Lucent Bell Labs.

In 2013, Inria edited the M12 milestone document of Task 6.4 "TWIN implementations and preliminary MAC protocol specifications". A paper on applying the network architecture and MAC/DBA protocols proposed by the team to the domain of data center interconnects has been submitted.

RAP has continued to work on a two-year research contract with Orange Labs on further developing the multi-path architecture (20012-2013). The main contribution in 2013 has been to propose the use of tunable receivers in addition to tunable transmitters. This technological evolution is possible with recent developments in coherent transmission and offers greater flexibility and enhanced efficiency. Work is continuing on evaluating this architecture by simulation (using Onmet++) and by analytical modelling.

4.2. Algorithms: Content-Centric Networking

Participants: Christine Fricker, Philippe Robert, James Roberts, Nada Sbihi.

RAP participated in an ANR project named CONNECT which contributed to the definition and evaluation of a new paradigm for the future Internet: an information-centric network (ICN) where, rather than interconnecting remote hosts like IP, the network directly manages the information objects that users publish, retrieve and exchange. The project ended in December 2012 but we have continued to work on information-centric networking in 2013.

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. 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. The excellent accuracy of this method over a wide range of practically relevant traffic models has been explained mathematically. 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.

This year we completed the analysis of a star network topology with multiple nodes. Several scalings were used to describe the fluid limit behaviour.

4.4. 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.

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.

This year we have investigated a mathematical model of the production of proteins in prokaryotic cells. Up to now most of the mathematical used to study these problems concern the production of *one* fixed class of proteins. When several classes of proteins are considered, each class requires in fact a fraction of the common and limited resources of the cell. One has therefore to understand how the allocation of the resources within the cell is done. Due to the fact that the cytoplasm of the cell is a quite disorganized medium where the components of the cell move, the whole production process has an important stochastic component. A model describing the allocation of the ribosomes of the cell to produce proteins is investigated via a Markovian representation. Asymptotic results for the equilibrium and for the transient behavior have been obtained under a scaling procedure and a reasonable biological assumption of saturation, i.e. when resources of the cell are tight. The equilibrium and the transient behavior have been investigated, it has been shown in particullar that, in the limit, the number of free ribosomes converges in distribution to a Poisson distribution whose parameter satisfies a fixed point equation.

4.5. Stochastic networks: large bike sharing systems

Participants: Christine Fricker, Hanène Mohamed, Nicolas Servel.

This is a collaboration with Nicolas Gast (EPFL). Bike sharing systems were launched by numerous cities to be a urban mode of transportation, for example Velib in Paris. One of the major issues is the availability of the resources: bikes or free slots to return the bikes. These systems became a hot topic in Operation Research and now the importance of stochasticity of such system behavior is commonly admitted. The problem is to understand their behavior and how to manage them in order to provide both resources to users.

Our model is the first one taking into account the finite number of spots at the stations. In a homogeneous model, mean field limit theorems give the dynamic of a large system. Analytical results are obtained and convergence proved in a standard model via Lyapunov functions. It allows to find the best ratio of bikes per station and to measure the improvement of incentive mechanisms, as choosing among two stations for example. We investigate also redistribution of bikes by trucks. Further results deal with heterogeneous system. By mean field techniques, analytical results were recently obtained on systems consisting in several clusters. In a work with Nicolas Servel, we discuss the improvement of choosing between two stations in the same cluster. Our goal is to propose, via a theoretical study and tests, simple algorithms to improve the system behavior.

With Hanene Mohamed, we study the problem of impact of geometry on incentive mechanisms. Our first model under investigation is very close from the Gates-Westcott crystal growth model with its underlying random deposition process.

4.6. Random Graphs

Participants: Nicolas Broutin, Henning Sulzbach.

4.6.1. 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) 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.

4.6.2. 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.

In a stochastic context, the minimum spanning tree is tigthly 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 permits to describe precisely the metric space scaling limit of a random minimum spanning tree, which identifies a novel continuum random tree which is truely different from that of Aldous.

4.6.3. Analysis of recursive partitions

This is joint work with R. Neininger (Frankfurt)

The techniques that we developped in order to estimate the cost of partial match queries in random quad trees have been used to solve an open question about the recursive lamination of the disk. We have proved that the planar dual of the lamination, which is a tree, converges almost surely when suitably rescaled to a compact random tree encoded by a continuous function. We also pinned down the fractal dimension of the limit object.

SOCRATE Project-Team

6. New Results

6.1. Flexible Radio Front-End

The contributions on hardware design are twofold. First, the development of a Full-Duplex architecture for OFDM systems. Second, a proposal of a Wake-Up scheme for home networking with reduced power consumption.

6.1.1. Full-Duplex systems

Zhan et al. [23] focused on the study of active analog self-interference cancellation (AASIC) techniques in full-duplex OFDM systems. This original approach aims at proposing a cancellation technique at RF level for wideband systems. A theoretical study confronted to simulations was proposed with a particular emphasis on the channel estimation of the interfering signal. This study was completed with an analysis on the phase noise and the thermal noise impact.

6.1.2. Wake-Up Architectures

Khoumeri et al. [28] proposed radio architectures for allowing energy savings by letting devices to switch off part of the transmission components when they are not in use. Based on classical WiFi systems, the proposed architecture offers the ability to use a conventional emitter, using only a particular subcarrier fingerprint to identify the node to wake-up, hence avoiding a high level of false wake-up.

6.2. Agile Radio Resource Sharing

The contributions of the axis in *agile radio resource sharing* can be gathered in three groups: (a) green communications; (b) performance analysis; and (c) scheduling and power allocation techniques.

6.2.1. Green Communications

The main contributions in the subject of green communications focus on the problem of increasing the energy efficiency of Orthogonal Frequency-Division Multiple Access (OFDMA) wireless networks. In particular, Tsilimantos et al. in [21] and Hasan et al. in [15] studied different techniques to strategically switch off some of the base stations in cellular systems while guaranteeing a given quality of service (QoS). In [21], the authors use methods from stochastic geometry to determine the number of active cells that can be switched off while the outage probability, or equivalently the signal to interference plus noise ratio (SINR), remains the same. In [15], this problem is studied from a decentralized point of view using methods from coalitional game theory.

6.2.2. Performance Analysis

The contributions in performance analysis are mainly oriented to the field of body area networks (BANs), [17] indoor adaptive OFDM wireless networks and relay channels.

In [17], Lauzier et al. presented the results of a measurement campaign whose primary objective was to characterize the complete mesh of a BAN and analyze the quality of every radio link between the different nodes. In [18], [19], the Multi-Resolution Frequency Domain ParFlow (MRFDPF) model is used to calculate the bit error rate (BER) and study the feasibility of adaptive modulation in OFDM systems.

In the context of relay channels, Ferrand et al. [32] studied the asymptotic *coding gain* of the packet error rate of relay channels in which the radio links are subject to both fading and log-normal shadowing effects simultaneously.

6.2.3. Scheduling and Power Allocation Techniques

Power allocation techniques and scheduling were studied by Ferrand et al. [33] and Wang et al. [9]. More specifically, advances in the study of the achievable rate region of relay channels in the case of global power constraints were reported in [33]. Cooperative scheduling techniques in the context of BAN were proposed in [9] to reduce inter-BAN interference using tools from game theory.

6.3. Software Radio Programming Model

Software defined radio (SDR) technology has evolved rapidly and is now reaching market maturity. Still, a lot of issues have yet to be studied. Mickaël Dardaillon, Kevin Marquet, Tanguy Risset and others highlighted the constraints imposed by recent radio protocols and presented current architectures, solutions, and challenges for programming SDR [31].

6.3.1. Dataflow programming

To enable dynamic adaptation of computation intensive multimedia dataflow applications, Lionel Morel, Kevin Marquet and others have studied language extensions, together with the corresponding run-time support. They show that this approach can be used to monitor and control throughput [20] and offer quality of service [29], with a low impact on the overall performance.

6.3.2. Energy-efficient Localization

Guillaume Salagnac and others address the tradeoff between energy consumption and localization performance in a mobile sensor network application [7]. The focus is on augmenting GPS location with more energy-efficient location sensors to bound position estimate uncertainty while GPS is off. Such combined strategies can cut node energy consumption by one third while still meeting application-specific positioning criteria.

6.3.3. Swap Fairness for Thrashing Mitigation

In the context of shared hosting or virtualization, where multiple users run uncoordinated and selfish workloads, François Goichon, Guillaume Salagnac and Stéphane Frénot introduced an accounting layer that forces swap fairness among processes competing for main memory [13]. It ensures that a process cannot monopolize the swap subsystem by delaying the swap operations of abusive processes, reducing the number of system-wide page faults while maximizing memory utilization.

URBANET Team

6. New Results

6.1. Characterizing and measuring urban networks

Participants: Marco Fiore, Diala Naboulsi, Razvan Stanica, Sandesh Uppoor

6.1.1. Properties of urban vehicular traffic and implications on mobile networking.

The goal of Sandesh Uppoor's PhD thesis [4] was to model and understand the mobility dynamics of high-speed vehicular users and their effect on wireless network architectures in an urban environment. Given the importance of developing the study on a realistic representation of vehicular mobility, we first survey the most popular approaches for the generation of synthetic road traffic and discuss the features of publicly available vehicular mobility datasets. Using original travel demand information of the population of a metropolitan area (Cologne area, Germany), detailed road network data and realistic microscopic driving models, we propose a novel state-of-art vehicular mobility dataset that closely mimics the real-world road traffic dynamics in both time and space [25]. We then study the impact of such mobility dynamics from the perspective of wireless cellular network architecture in presence of a real-world base station deployment. In addition, by discussing the effects of vehicular mobility on autonomous network architecture, we hint at the opportunities for future heterogeneous network paradigms and demonstrate how incomplete representations of vehicular mobility may result in over-optimistic network connectivity and protocol performance [8].

Motivated by the time-evolving mobility dynamics observed in our original dataset, we also propose an on line approach to predict near-future macroscopic traffic flows. We analyze the parameters affecting the mobility prediction in an urban environment and unveil when and where network resource management is more crucial to accommodate the traffic generated by users on-board. Such studies unveil multiple opportunities in transportation management either for building new roads, installing electric charging points, or for designing intelligent traffic light systems, thereby contributing to urban planning.

6.1.2. Feasibility of multi-hop vehicular communications in an urban environment.

Despite the growing interest in a real-world deployment of vehicle- to-vehicle communication, many topological features of the resulting vehicular network remain largely unknown. We still lack a clear understanding of the level of connectivity achievable in large-scale urban scenarios, of the availability and reliability of connected multi-hop paths, and of the evolution of such features over daytime. In [14], we investigate how the instantaneous topology of the vehicular network would look like in the case of a typical middle-sized European city, using the example of the Cologne mobility trace. Through a complex network analysis, we unveil the low connectivity, availability, reliability and navigability of the network, and exploit our findings to derive network design and usage guidelines.

6.1.3. Investigating the accuracy of mobile urban sensing.

Community urban sensing is one of the emerging applications enabled by the growing popularity of mobile user devices, like smartphones and in-vehicle monitoring systems. Such devices feature sensing and wireless communication capabilities, which enable them to sample large-scale phenomena, like air pollution and vehicular traffic congestion, and upload these data to the Internet. In [10], we focus on the above scenario and investigate the level of accuracy that can be achieved in estimating the phenomenon of interest through a mobile crowdsourcing application. Specifically, we take a signal processing-based approach and leverage results on signal reconstruction from sets of irregularly spaced samples. We apply such results to a realistic scenario where samples are collected by vehicular and pedestrian users, and study the accuracy level of the phenomenon estimation as the penetration rate of the sensing application varies.

6.1.4. Analysis of mobile network call detail records.

The growing ubiquity of mobile communications has offered researchers new possibilities to understand human mobility over the last few years. In [22], we analyze Call Detail Records (CDR) made available within the context of the Orange D4D Challenge, focusing on calls of individuals in the city of Abidjan, Ivory Coast, over a period of five months. Our results illustrate how aggregated CDR can be used to tell apart typical and special mobility behaviors, and demonstrate how macroscopic mobility flows extracted from these cellular network data reflect the daily dynamics of a highly populated city. We discuss how these macroscopic mobility flows can help solve problems in developing urban areas.

6.2. Scalable solutions for capillary networks

Participants: Isabelle Augé-Blum, Jin Cui, Marco Fiore, Ochirkhand Erdene-Ochir, Alexandre Mouradian, Hervé Rivano, Razvan Stanica, Fabrice Valois

6.2.1. Real-time wireless sensor networks.

Critical applications for WSNs are emerging, with real-time and reliability requirements. Critical applications are applications on which depend human lives and the environment: a failure of a critical application can thus have dramatic consequences. We are especially interested in anomaly detection applications (forest fire detection, landslide detection, intrusion detection, etc), which require bounded end to end delays and high delivery ratio. Few WSNs protocols of the literature allow to bound end to end delays. Among the proposed solutions, some allow to effectively bound the end to end delays, but do not take into account the characteristics of WSNs (limited energy, large scale, etc). Others take into account those aspects, but do not give strict guaranties on the end to end delays. In this sense, the PhD thesis of Alexandre Mouradian [2] proposes a real-time anomaly detection solution composed of:

- A virtual coordinate system which allows to discriminate nodes in a 2-hop neighborhood and to bound the number of hops between any source and the sink.
- A cross-layer protocol for WSNs (named RTXP) based on the proposed virtual coordinate system. Thanks to these coordinates it is possible to introduce determinism in the accesses to the medium and to bound the hop-count, this allows to bound the end to end delay. RTXP adapts its duty-cycle to the traffic loads and uses an opportunistic routing scheme to increase its delivery ratio. We show, by simulation, that RTXP outperforms real-time protocols of the literature for anomaly detection in WSNs under harsh radio conditions.
- A real-time aggregation scheme to mitigate the alarm storm problem which causes collisions and congestion and thus limit the network lifetime. This scheme is also based on the virtual coordinate system and is used before RTXP in order to reduce the number of similar alarms converging toward the sink.

6.2.2. Formal verification of wireless sensor networks protocols.

WSN protocols used by critical applications must be formally verified in order to provide the strongest possible guaranties: simulations and tests are not sufficient in this context, formal proofs of compliance with the specifications of the application have to be provided.. Unfortunately the radio link is unreliable and it is thus difficult to give hard guarantees on the temporal behavior of the protocols. Indeed, a message may experience a very high number of retransmissions and the temporal guarantee can only be given with a certain probability. This probability must meet the requirements of the application. Network protocols have been successfully verified on a given network topology without taking into account unreliable links. Nevertheless, the probabilistic nature of radio links may change the topology (links which appear and disappear). Thus instead of a single topology we have a set of possible topologies, each topology having a probability to exist. In [12], we propose a method that produces the set of topologies, checks the property on every topology, and gives the probability that the property is verified. This technique is independent from the verification technique, i.e. each topology can be verified using any formal method which can give a “yes” or “no” answer to the question: “Does the model of the protocol respect the property?”. We apply this method on the f-MAC protocol. We use

UPPAAL model checker as verification tool. We implement a tool that automatizes the process and thus show the feasibility of our proposition. We compare the results of the verification with simulation results. It appears that the verification is, as expected, conservative but not overly pessimistic compared to the simulated worst case. Besides we show that f-MAC is a reliable real-time protocol for WSNs (for up to 6 nodes), as we were not able to detect faults.

Moreover, in [2], a verification technique which mixes Network Calculus and Model Checking is proposed, in order to be both scalable and exhaustive. This technique consists in modeling the interaction of each node with the rest of the network with arrival curves and then to verify with UPPAAL that each node is capable of handling these interactions while meeting the deadlines. We apply this methodology in order to formally verify our pervious proposition, RTXP.

6.2.3. Reliability in wireless sensor networks.

WSN critical applications require the respect of time and reliability constraints. In [13], we provide a theoretical study of the reliability in WSNs. We define the reliability as the probability of success of an end-to-end transmission in the WSN. In this work, we use two radio propagation models : a basic model where the nodes have a set of neighbors they can communicate with, with a given probability, and the log-normal shadowing model, where probability of reception depends on the emitter-receiver distance. We determine the reliability of two routing schemes : unicast-based routing (classical routing) and broadcast-based routing (opportunistic routing). We conclude that the broadcast-based routing allows to reach a higher reliability than the unicast case. The main result is that we show the existence of a reliability bottleneck at the sink node in the case of the broadcast-based routing. We show that the addition of another sink improves the reliability of the network in this case.

6.2.4. Resiliency in wireless sensor networks.

Because of their open and unattended deployment, in possibly hostile environments, powerful adversaries can easily launch Denial-of-Service (Dos) attacks on wireless sensor networks, cause physical damage to sensors, or even capture them to extract sensitive information (encryption keys, identities, addresses, etc.). Consequently, the compromised node poses severe security and reliability concerns, since it allows an adversary to be considered as a legitimate node inside the network. To cope with these "insider" attacks, stemming from node compromise, "beyond cryptography" algorithmic solutions must be envisaged to complement the traditional cryptographic solutions. In this sense, in [1], we first propose the resiliency concept. Our goal is to propose a definition of the resiliency in our context (security of WSNs routing protocols) and a new metric to compare routing protocols. The originality of this metric is that we combine the graphical representation (qualitative information) with the aggregation method (quantitative information). We introduce a two dimensional graphical representation with multiple axes forming an equiangular polygon surface. This method allows to aggregate meaningfully several parameters and makes it easier to visually discern various trade-offs, thus greatly simplifying the process of protocol comparison.Secondly, we propose the protocol behaviors enhancing resiliency. Our proposition consists in three elements: (i) introduce random behaviors (ii) limit route length (iii) introduce data replication. Random behaviors increase uncertainty for an adversary, making the protocols unpredictable. Data replication allows route diversification between the sources and the sink, thus improving the delivery success and fairness. Limitation of the route length is necessary to reduce the probability of a data packet to meet a malicious insider along the route. The quantitative metric enables to propose a new resiliency taxonomy of WSNs routing protocols. According to this taxonomy, the gradient based routing is the most resilient when it is combined with the proposed behaviors. Thirdly, several variants of the gradient-based routing (classical and randomized) under more complex and realistic adversary model (several combined attacks) are considered to extend our simulations. Several values of bias are introduced to the randomized variants and two data replication methods (uniform and adaptive) are considered. Without attacks, the most biased variants without replications are the most efficient. However, under moderate attacks, the replication uniform is the most adapted, while under intense attacks, the replication adaptive is the most suitable. Finally, a theoretical study of the resiliency is introduced. We present an analytical study of the biased random walk routing under attacks. The influence of bias is evaluated and two replication methods that previously evaluated by simulations are considered. After presenting the delivery success and the energy consumption of all scenarios, we

evaluate them with our resiliency metric. This study permits to confirm the results obtained with simulations and it shows that the bias is essential to enhance the resiliency of random routing.

6.2.5. Data aggregation in wireless sensor networks.

Data aggregation is a crucial problem in wireless sensor networks due to their constrained-energy and constrained-bandwidth nature. In [26], we highlight the aggregation benefits at the Network layer and MAC layer by modeling the energy consumption for some energy-efficient routing protocols and MAC protocols. Besides, we define two parameters, the aggregation ratio and the packet size coefficient to evaluate the efficiency of an aggregation method, and to discuss the trade-off. Additionally, we investigate the differences between time series and compressive sensing, which are representative state-of-the-art solutions for forecasting aggregation and compressing aggregation respectively.

6.2.6. Routing in delay-tolerant networks.

Delay-Tolerant Networks (DTN) model systems that are characterized by intermittent connectivity and frequent partitioning. Routing in DTNs has drawn much research effort recently. Since very different kinds of networks fall in the DTN category, many routing approaches have been proposed. In particular, the routing layer in some DTNs has information about the schedules of contacts between nodes and about data traffic demand. Such systems can benefit from a previously proposed routing algorithm based on linear programming that minimizes the average message delay. This algorithm, however, is known to have performance issues that limit its applicability to very simple scenarios. In [9], we propose an alternative linear programming approach for routing in Delay-Tolerant Networks. We show that our formulation is equivalent to that presented in a seminal work in this area, but it contains fewer LP constraints and has a structure suitable to the application of Column Generation (CG). Simulation shows that our CG implementation arrives at an optimal solution up to three orders of magnitude faster than the original linear program in the considered DTN examples.

6.2.7. Performance evaluation of vehicular communications.

Wireless vehicular networks face different problems and challenges, especially in a dense urban environment. In [23], we first characterize the different types of loss in vehicular networks: radio propagation problems, expired security messages, collision with one hop neighbor and collisions with hidden terminals. In a second step, we give the architecture of the wireless vehicular network and describe the Medium Access Control (MAC) quality of service mechanisms proposed by vehicular environment standards that aim at meeting the road drivers' expectation and increasing road safety. To complete this image, in [24], we provide a literature survey that covers the solutions proposed in order to enable critical dissemination of urgent messages and surpass the challenging vehicular dynamic topology. More particularly, we detail the following techniques: beaconing frequency reduction, transmit rate control, power control, adaptation of the contention window and adaptation of the carrier sense threshold.

6.2.8. Secure node localization in mobile ad-hoc networks.

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. In [6], we address this open issue by proposing a fully distributed cooperative solution that is robust against independent and colluding adversaries, and can be impaired only by an overwhelming presence of 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.

In a vehicular context, knowledge of the location of vehicles and tracking of the routes they follow are a requirement for a number of applications. However, public disclosure of the identity and position of drivers jeopardizes user privacy, and securing the tracking through asymmetric cryptography may have an exceedingly high computational cost. In [11], we address all of the issues above by introducing A-VIP, a lightweight privacy-preserving framework for tracking of vehicles. A-VIP 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. We assess the effectiveness of A-VIP through testbed implementation results.

6.3. Cellular network solutions

Participants: Marco Fiore, Anis Ouni, Hervé Rivano, Razvan Stanica, Fabrice Valois

6.3.1. Optimizing capacity and energy consumption in wireless mesh networks.

Wireless mesh networks (WMN) are a promising solution to support high data rate and increase the capacity provided to users, e.g. for meeting the requirements of mobile multimedia applications. However, the rapid growth of traffic load generated by the terminals is accompanied by an unsustainable increase of energy consumption, which becomes a hot societal and economical challenges. This thesis relates to the problem of the optimization of network capacity and energy consumption of wireless mesh networks. The network capacity is defined as the maximum achievable total traffic in the network per unit time.

The thesis of Anis Ouni [3] addresses this issue and is divided into four main parts. First, we address the problem of improvement of the capacity of 802.11 wireless mesh networks. We highlight some insensible properties and deterministic factors of the capacity, while it is directly related to a bottleneck problem. Then, we propose a joint TDMA/CSMA scheduling strategy for solving the bottleneck issue in the network.

Second, we focus on broadband wireless mesh networks based on time-frequency resource management. In order to get theoretical bounds on the network performances, we formulate optimization models based on linear programming and column generation algorithm. These models lead to compute an optimal offline configuration which maximizes the network capacity with low energy consumption. A realistic SINR model of the physical layer allows the nodes to perform continuous power control and use a discrete set of data rates.

Third, we use the optimization models to provide practical engineering insights on WMN. We briefly study the tradeoff between network capacity and energy consumption using a realistic physical layer and SINR interference model [27]. In particular, we show that power control and multi-rate functionalities allow to reach optimal throughput with lower energy consumption using a mix of single hop and multi- hop routes.

Finally, we focus on capacity and energy optimization for heterogeneous cellular networks. We develop optimization tools to calculate an optimal configuration of the network that maximizes the network capacity with low energy consumption. We then propose a heuristic algorithm that calculates a scheduling and partial sleeping of base stations in two different strategies, called LAFS and MAFS.

6.3.2. Sleep protocols for heterogeneous LTE networks.

The tremendous increase of the traffic demand in cellular networks imposes a massive densification of the traditional cellular infrastructure. The network architecture becomes heterogeneous, in particular 4G networks where LTE micro-eNodeBs are deployed to strengthen the coverage of macro-eNodeBs. This densification yields major issues related to the energy consumption of the infrastructure. Indeed, there is fixed and significant amount of energy required to run each additional node, whatever the traffic load of the network. Mitigating this fixed energy consumption is therefore a major challenge from a societal and economical viewpoint. Extensive researches about energy-saving highlight that to save energy the better strategy is to switch off the radio part of nodes. This is the heart of wireless sensor networks energy-saving strategies, even though the objective for WSN is to maximize the battery life of each individual nodes. In [18], we develop a parallel between the principles of WSN protocols and the requirements of cellular infrastructures. We then propose a distributed and localized algorithm to dynamically switch off and on the micro-eNodeBs of an LTE heterogeneous network following the traffic demand evolution in time and analyze it in terms of energy savings. We show that one can expect energy savings of approximately 12% when implementing sleep modes whereas the energy cost for sending the traffic decreases by 24%.

6.3.3. Content downloading through a vehicular network.

The focus of the work we present in [7] is twofold: information dissemination from infrastructure nodes deployed along the roads, the so-called Road-Side Units (RSUs), to passing-by vehicles, and content downloading by vehicular users through nearby RSUs. In particular, in order to ensure good performance for both content dissemination and downloading, the presented study addresses the problem of RSU deployment and reviews previous work that has dealt with such an issue. The RSU deployment problem is then formulated as an optimization problem, where the number of vehicles that come in contact with any RSU is maximized, possibly considering a minimum contact time to be guaranteed. Since such optimization problems turn out to be NP-hard, heuristics are proposed to efficiently approximate the optimal solution. The RSU deployment obtained through such heuristics is then used to investigate the performance of content dissemination and downloading through ns2 simulations. Simulation tests are carried out under various real-world vehicular environments, including a realistic mobility model, and considering that the IEEE 802.11p standard is used at the physical and medium access control layers. The performance obtained in realistic conditions is discussed with respect to the results obtained under the same RSU deployment, but in ideal conditions and protocol message exchange. Based on the obtained results, some useful hints on the network system design are provided.

6.3.4. Offloading Floating Car Data.

Floating Car Data (FCD) is currently collected by moving vehicles and uploaded to Internet-based processing centers through the cellular access infrastructure. As FCD is foreseen to rapidly become a pervasive technology, the present network paradigm risks not to scale well in the future, when a vast majority of automobiles will be constantly sensing their operation as well as the external environment and transmitting such information towards the Internet. In order to relieve the cellular network from the additional load that widespread FCD can induce, we study [16] a local gathering and fusion paradigm, based on vehicle-to-vehicle (V2V) communication. We show how this approach can lead to significant gain, especially when and where the cellular network is stressed the most. Moreover, we propose several distributed schemes to FCD offloading based on the principle above that, despite their simplicity, are extremely efficient and can reduce the FCD capacity demand at the access network by up to 95%.

6.3.5. Mobile malware propagation in vehicular networks.

The large-scale adoption of vehicle-to-vehicle (V2V) communication technologies risks to significantly widen the attack surface available to mobile malware targeting critical automobile operations. Given that outbreaks of vehicular computer worms self-propagating through V2V links could pose a significant threat to road traffic safety, it is important to understand the dynamics of such epidemics and to prepare adequate countermeasures. In [17], we perform a comprehensive characterization of the infection process of variously behaving vehicular worms on a road traffic scenario of unprecedented scale and heterogeneity. We then propose a simple yet effective data-driven model of the worm epidemics, and we show how it can be leveraged for smart patching infected vehicles through the cellular network in presence of a vehicular worm outbreak.