
RESEARCH CENTER

FIELD
**Networks, Systems and Services, Distributed Computing**

Activity Report 2016

# Section Scientific Foundations

Edition: 2017-08-25

**ASAP Project-Team**

# 3. Research Program

## 3.1. Theory of distributed systems

Finding models for distributed computations prone to asynchrony and failures has received a lot of attention. A lot of research in this domain focuses on what can be computed in such models, and, when a problem can be solved, what are its best solutions in terms of relevant cost criteria. An important part of that research is focused on distributed computability: what can be computed when failure detectors are combined with conditions on process input values for example. Another part is devoted to model equivalence. What can be computed with a given class of failure detectors? Which synchronization primitives is a given failure class equivalent to? These are among the main topics addressed in the leading distributed computing community. A second fundamental issue related to distributed models is the definition of appropriate models suited to dynamic systems. Up to now, the researchers in that area consider that nodes can enter and leave the system, but do not provide a simple characterization, based on properties of computation instead of description of possible behaviors [58], [51], [53]. This shows that finding dynamic distributed computing models is today a "Holy Grail", whose discovery would allow a better understanding of the essential nature of dynamic systems.

## 3.2. Peer-to-peer overlay networks

A standard distributed system today is related to thousands or even millions of computing entities scattered all over the world and dealing with a huge amount of data. This major shift in scalability requirements has lead to the emergence of novel computing paradigms. In particular, the peer-to-peer communication paradigm imposed itself as the prevalent model to cope with the requirements of large scale distributed systems. Peer-to-peer systems rely on a symmetric communication model where peers are potentially both clients and servers. They are fully decentralized, thus avoiding the bottleneck imposed by the presence of servers in traditional systems. They are highly resilient to peers arrivals and departures. Finally, individual peer behavior is based on a local knowledge of the system and yet the system converges toward global properties.

A peer-to-peer overlay network logically connects peers on top of IP. Two main classes of such overlays dominate, structured and unstructured. The differences relate to the choice of the neighbors in the overlay, and the presence of an underlying naming structure. Overlay networks represent the main approach to build large-scale distributed systems that we retained. An overlay network forms a logical structure connecting participating entities on top of the physical network, be it IP or a wireless network. Such an overlay might form a structured overlay network [59], [60], [61] following a specific topology or an unstructured network [56], [62] where participating entities are connected in a random or pseudo-random fashion. In between, lie weakly structured peer-to-peer overlays where nodes are linked depending on a proximity measure providing more flexibility than structured overlays and better performance than fully unstructured ones. Proximity-aware overlays connect participating entities so that they are connected to close neighbors according to a given proximity metric reflecting some degree of affinity (computation, interest, etc.) between peers. We extensively use this approach to provide algorithmic foundations of large-scale dynamic systems.

## 3.3. Epidemic protocols

Epidemic algorithms, also called gossip-based algorithms [55], [54], constitute a fundamental topic in our research. In the context of distributed systems, epidemic protocols are mainly used to create overlay networks and to ensure a reliable information dissemination in a large-scale distributed system. The principle underlying technique, in analogy with the spread of a rumor among humans via gossiping, is that participating entities continuously exchange information about the system in order to spread it gradually and reliably. Epidemic algorithms have proved efficient to build and maintain large-scale distributed systems in the context of many applications such as broadcasting [54], monitoring, resource management, search, and more generally in building unstructured peer-to-peer networks.

## 3.4. Malicious process behaviors

When assuming that processes fail by simply crashing, bounds on resiliency (maximum number of processes that may crash, number of exchanged messages, number of communication steps, etc.) are known both for synchronous and augmented asynchronous systems (recall that in purely asynchronous systems some problems are impossible to solve). If processes can exhibit malicious behaviors, these bounds are seldom the same. Sometimes, it is even necessary to change the specification of the problem. For example, the consensus problem for correct processes does not make sense if some processes can exhibit a Byzantine behavior and thus propose an arbitrary value. In this case, the validity property of consensus, which is normally "a decided value is a proposed value", must be changed to "if all correct processes propose the same value then only this value can be decided." Moreover, the resilience bound of less than half of faulty processes is at least lowered to "less than a third of Byzantine processes." These are some of the aspects that underlie our studies in the context of the classical model of distributed systems, in peer-to-peer systems and in sensor networks.

## 3.5. Online social networks and recommender systems

Social Networks have rapidly become a fundamental component of today's distributed applications. Web 2.0 applications have dramatically changed the way users interact with the Internet and with each other. The number of users of websites like Flickr, Delicious, Facebook, or MySpace is constantly growing, leading to significant technical challenges. On the one hand, these websites are called to handle enormous amounts of data. On the other hand, news continue to report the emergence of privacy threats to the personal data of social-network users. Our research aims to exploit our expertise in distributed systems to lead to a new generation of scalable, privacy-preserving, social applications.

We also investigate approaches to build implicit social networks, connecting users sharing similar interests. At the heart of the building of such similarity graphs lie k-nearest neighbor (KNN) algorithms. Our research in this area is to design and implement efficient KNN algorithms able to cope with a huge volume of data as well as a high level of dynamism. We investigate the use of such similarity graphs to build highly scalable infrastructures for recommendation systems.

**CIDRE Project-Team**

# 3. Research Program

## 3.1. Our perspective

For many aspects of our everyday life, we heavily rely on information systems, many of which are based on massively networked devices that support a population of interacting and cooperating entities. While these information systems become increasingly open and complex, accidental and intentional failures get considerably more frequent and severe.

Two research communities traditionally address the concern of accidental and intentional failures: the distributed computing community and the security community. While both these communities are interested in the construction of systems that are correct and secure, an ideological gap and a lack of communication exist between them that is often explained by the incompatibility of the assumptions each of them traditionally makes. Furthermore, in terms of objectives, the distributed computing community has favored systems availability while the security community has focused on integrity and confidentiality, and more recently on privacy.

Our long term ambition is to contribute to the building of distributed systems that are trustworthy and respectful of privacy, even when some nodes [0] in the system have been compromised. For that purpose, we are convinced that combining classical security approaches and distributed computing paradigms is an interesting way to enforce the security of large-scale distributed systems. More specifically, since a distributed system is composed of nodes, we assert that the security of large-scale distributed systems has to be addressed at three complementary levels:

- the level of each node: each standalone node has to enforce its own security;
- the level of an *identified* set of *trusted* nodes: the *trusted* nodes can *collaborate* to enforce together their security;
- the level of fully open large-scale distributed and dynamic systems: distributed computing paradigms such as consensus algorithms can be applied to cope with the possible presence of malicious nodes.

Notice that using a distributed architecture can also be an approach allowing the nodes to enforce their security without the need of a trusted third party.

The research activities of the CIDRE project-team focus mainly on the two following research axis:

- **Intrusion Detection System:** the objective is to detect any suspicious events with regard to the security by analyzing some data generated on the monitored system.
- **Privacy-preserving Services:** the objective is to ensure users' privacy even when this property seems incompatible with the provided services, like social networks or location-based services.

In all our studies, we consider a priori that the attacker is omnipotent. He can acts as he wants. Nevertheless, being not a team specialized in cryptography, we consider that we can rely on strong unbroken cryto-systems.

## 3.2. Intrusion Detection / Security Events Monitoring and Management

Today, we are not yet fully entered into a world of "security by design". Security remains often a property that is considered a posteriori, when the system is deployed, which often results in applying patches when vulnerabilities are discovered (also called a "patch and pray" approach). Unfortunately, despite patching, the number of vulnerabilities remains high, as evidenced by the number of vulnerabilities published each year in the Common Vulnerabilities and Exposures (CVE) system. Thus, it is important to be able to early detect cyber-attacks, especially when they exploit vulnerabilities that are unknown. However, the efficiency of

[0] The term node either refers to a device that hosts a network client or service or to the process that runs this client or service.

security events monitoring and management systems (including the IDS - Intrusion Detection Systems) is still an open issue today. Indeed, they are often unable to effectively deal with huge numbers of security events, and they usually produce too many false alarms yet missing some attacks. So one of the main research challenges in IT security remains the definition of efficient security events monitoring systems, i.e., that enable both to process a huge number of security events and to detect any attacks without flooding the security analysts with false alarms.

By exploiting vulnerabilities in operating systems, applications, or network services, an attacker can defeat preventive security mechanisms and violate the security policy of the whole system. The goal of an Intrusion Detection Systems (IDS) is to detect such violations by analyzing some *security events* generated on a monitored system. Ideally, the IDS should produce an alert for any violation (no *false negative*), and only for violations (no *false positive*).

To produce alerts, two detection techniques exist: the misuse based detection and the anomaly based detection. A misuse based detection is actually a signature based detection approach : it allows to detect only the attacks whose signature is available. From our point of view, while useful in practice, misuse detection is intrinsically limited. Indeed, it requires to update in real-time the database of signatures, similarly to what has to be done for antivirus tools. The CIDRE project-team follows the alternative approach, namely the anomaly approach, which consists in detecting a deviation from a referenced behavior. Our contributions on anomaly-based IDS follow three axis:

- **Illegal Information Flow Detection:** our goal is to detect information flows in the monitored system (either a node or a set of trusted nodes) that are allowed by the access control mechanism, but are illegal from the security policy point of view. This approach is particularly appealing to detect intrusions in a standalone node, such as a smartphone.
- **Anomaly-Based Detection in Distributed Applications:** our goal is to specify the normal behavior based on either a formal specification of the distributed application, or previous executions. This approach is particularly appealing to detect intrusions in industrial control systems since these systems exhibit well-defined behaviors at different levels: network level (network communication patterns, protocol specifications, etc.), control level (continue and discrete process control laws), or even the state of the local resources (memory or CPU).
- **Online data analytics:** our goal is to estimate on the fly different statistics or metrics on distributed input streams to detect abnormal behavior with respect to a well-defined criterion such as the distance between different streams, their correlation or their entropy.

Beside the anomaly-based IDS, we have also led research work on alert correlation and visualisation of security events. Indeed, in large systems, multiple (host and network) IDS and many sensors are deployed and they continuously and independently generate notifications (event's observations, warnings and alerts). To cope with this huge amount of collected data, we have studied two different approaches, each with specific goal:

- **Alert Correlation System:** the alerts of *low level* IDSes can be viewed as *security events* of a *high level* IDS whose goal is to correlate these alerts. An alert correlation system aims at exploiting the known relationships between some elements that appear in the flow of low level notifications to generate high semantic meta-alerts. The main goal is to reduce the number of alerts (and especially, false positive) returned to the security analysts and to allow a higher level analysis of the situation (situational awareness).
- **Visualization Tools:** a visualization tools aims at relying on the capacity of human beings to detect patterns and outliers in datasets when these datasets are properly visually represented. Human beings also know pieces of contextual information that are very difficult to formalize so as to make them usable by a computer. Visualization is therefore a very useful complementary tool to detect abnormal events in real time (monitoring), to search for malicious events in log files (data exploration and forensics) and to communicate results (reporting).

## 3.3. Privacy

In a world of ubiquitous technologies, each individual constantly leaves digital traces related to his activities and interests. The current business plan of many web services such as social networks, is based on the sale of these digital traces. Of course, this is usually done in a legal way, the license of use clearly stating that the user gives the right to the service provider for using his personal data. However, on the one hand, users generally do not read these licenses, and on the other hand, these licenses are usually very vague on the use of personal data [0]. In addition these digital traces can potentially be stolen and maliciously used, they must therefore be protected. In this context, users' privacy is now recognized as a fundamental individual right. Any new IT service should thus follow the *privacy-by-design* approach: privacy issues have to be studied from the earliest phase of a project by taking into account the multi-stakeholders and transdisciplinary aspects in order to ensure proper, end-to-end private data protection properties.

In the CIDRE project, we mainly focus on domains in which privacy issues collide with provided services. Here are some concrete examples of such domains:

- **Location-based services:** the challenge is to design services that depend on the user's location while preserving the privacy of his location;
- **Social networks:** the challenge is to demonstrate that it is possible to design social networks respectful of users' privacy;
- **Mobile services:** given that such services are based on user's identity, the challenge is to design mobile services while preserving the users' anonymity;
- **Ad-hoc netwoks:** in ad-hoc networks, any participant can potentially know the relative location of the other participants. Thus, the issue is to allow nodes to forward messages while preserving the privacy of the communications.

For all of these domains, we have proposed new Privacy-Enhancing Techniques (PETs) based on a mix of different foundations such as cryptographic techniques, security policies and access control mechanisms, just to name a few. More generally, we think that a major option to protect users' privacy consists in using a decentralized architecture that enables to transfer control and services from the service providers to the users.

The concept of IDS seems to be in contradiction with the users' privacy. Indeed, an IDS is a monitoring system that needs to collect and analyze information coming from different levels such as network, applications and OS, this information being able to include users' personal data. However, we are confident that IDS and privacy are not completely antagonist. In particular, integrating some privacy features inside an IDS to build a privacy-preserving IDS may allow to limit the amount of information that can leak if one of the nodes within the system is compromised. On the other hand, enabling IDS to detect attacks against privacy as well as security violations can extend the range of their applicability.

[0] Besides, it has been shown that service providers do not necessarily comply with their own license.

**COAST Project-Team**

# 3. Research Program

## 3.1. Introduction

Our scientific foundations are grounded on distributed collaborative systems supported by sophisticated data sharing mechanisms and on service oriented computing with an emphasis on orchestration and on non-functional properties.

Distributed collaborative systems enable distributed group work supported by computer technologies. Designing such systems requires an expertise in Distributed Systems and in Computer-supported collaborative Work research area. Besides theoretical and technical aspects of distributed systems, the design of distributed collaborative systems must take into account the human factor to offer solutions suitable for users and groups. The Coast team vision is to move away from a centralized authority based collaboration towards a decentralized collaboration where users have full control over their data that they can store locally and decide with whom to share them. The Coast team investigates the issues related to the management of distributed shared data and coordination between users and groups.

Service oriented Computing [26] is an established domain on which the ECOO, Score and now the Coast teams have been contributing for a long time. It refers to the general discipline that studies the development of computer applications on the web. A service is an independent software program with a specific functional context and capabilities published as a service contract (or more traditionally an API). A service composition aggregates a set of services and coordinates their interactions. The scale, the autonomy of services, the heterogeneity and some design principles underlying Service Oriented Computing open new research questions that are at the basis of our research. They span the disciplines of **distributed computing**, **software engineering** and **computer supported collaborative work** (CSCW). Our approach to contribute to the general vision of Service Oriented Computing and more generally to the emerging discipline of Service Science has been and is still to focus on the issue of the efficient and flexible construction of reliable and secure high level services through the coordination/orchestration/composition of other services provided by distributed organizations or people.

## 3.2. Consistency Models for Distributed Collaborative Systems

Collaborative systems are distributed systems that allow users to share data. One important issue is to manage consistency of shared data according to concurrent access. Traditional consistency criteria such as serializability, linearizability are not adequate for collaborative systems.

Causality, Convergence and Intention preservation (CCI) [30] are more suitable for developing middleware for collaborative applications.

We develop algorithms for ensuring CCI properties on collaborative distributed systems. Constraints on the algorithms are different according to the kind of distributed system and to the data structure. The distributed system can be centralized, decentralized or peer-to-peer. The type of data can include strings, growable arrays, ordered trees, semantic graphs and multimedia data.

## 3.3. Optimistic Replication

Replication of data among different nodes of a network allows improving reliability, fault-tolerance, and availability. When data are mutable, consistency among the different replicas must be ensured. Pessimistic replication is based on the principle of single-copy consistency while optimistic replication allows the replicas to diverge during a short time period. The consistency model for optimistic replication [28] is called eventual consistency, meaning that replicas are guaranteed to converge to the same value when the system is idle.

Our research focuses on the two most promising families of optimistic replication algorithms for ensuring CCI:

- the operational transformation (OT) algorithms [24]
- the algorithms based on commutative replicated data types (CRDT) [27].

Operational transformation algorithms are based on the application of a transformation function when a remote modification is integrated into the local document. Integration algorithms are generic, being parametrized by operational transformation functions which depend on replicated document types. The advantage of these algorithms is their genericity. These algorithms can be applied to any data type and they can merge heterogeneous data in a uniform manner.

Commutative replicated data types is a new class of algorithms initiated by WOOT [25] a first algorithm designed WithOut Operational Transformations. They ensure consistency of highly dynamic content on peer-to-peer networks. Unlike traditional optimistic replication algorithms, they can ensure consistency without concurrency control. CRDT algorithms rely on natively commutative operations defined on abstract data types such as lists or ordered trees. Thus, they do not require a merge algorithm or an integration procedure.

## 3.4. Process Orchestration and Management

Process Orchestration and Management is considered as a core discipline behind Service Management and Computing. It includes the analysis, the modelling, the execution, the monitoring and the continuous improvement of enterprise processes and is for us a central domain of studies.

Much efforts have been devoted in the past years to establish standard business process models founded on well grounded theories (e.g. Petri Nets) that meet the needs of both business analysts but also of software engineers and software integrators. This has lead to heated debate in the BPM community as the two points of view are very difficult to reconcile. On one side, the business people in general require models that are easy to use and understand and that can be quickly adapted to exceptional situations. On the other side, IT people need models with an operational semantic in order to be able transform them into executable artefacts. Part of our work has been an attempt to reconcile these point of views. It resulted in the development of the Bonita Business process management system and more recently on our work in crisis management where the same people are designing, executing and monitoring the process as it executes. But more generally, and at a larger scale, we have been considering the problem of processes spanning the barriers of organisations and thus more general problem of service composition as a way to coordinate inter organisational construction of applications providing value based on the composition of lower level services [22].

## 3.5. Service Composition

We are considering processes as pieces of software whose execution traverse the boundaries of organisations. This is especially true with service oriented computing where processes compose services produced by many organisations. We tackle this problem from very different perspectives, trying to find the best compromise between the need for privacy of internal processes from organisations and the necessity to publicize large part of them, proposing to distribute the execution and the orchestration of processes among the organisations themselves, and attempting to ensure non-functional properties in this distributed setting [21].

Non-functional aspects of service composition relate to all the properties and service agreements that one wants to ensure and that are orthogonal to the actual business but that are important when a service is selected and integrated in a composition. This includes transactional context, security, privacy, and quality of service in general. Defining and orchestrating services on a large scale while providing the stakeholders with some strong guarantees on their execution is a first class problem for us. For a long time, we have proposed models and solutions to ensure that some properties (e.g. transactional properties) were guaranteed on process execution, either through design or through the definition of some protocols. Our work has also been extended to the problems of security, privacy and service level agreement among partners. These questions are still central in our work. One major problem of current approaches is to monitor the execution of the compositions, integrating the distributed dimension. This problem can be tackled using event-based

algorithms and techniques. Using our event oriented composition framework DISC, we have obtained new results dedicated to the runtime verification of violations in service choreographies.

**CTRL-A Team**

# 3. Research Program

## 3.1. Modeling and control techniques for autonomic computing

### *3.1.1. Continuous control*

Continuous control was used to control computer systems only very recently and in few occasions, despite the promising results that were obtained. This is probably due to many reasons, but the most important seems to be the difficulty by both communities to transform a computer system problem into an automatic control problem. The aim of the team is to explore how to formalize typical autonomic commuting cases into typical control problems. Many new methodological tools will probably be useful for that, e.g., we can cite the hybrid system approach, predictive control or event-based control approach. Computer systems are not usual for the control system community and they often present non-conventional control aspects like saturation control. New methodological tools are required for an efficient use of continuous-time control in computer science.

### *3.1.2. Discrete control*

Discrete control techniques are explored at long-term, to integrate more control in the BZR language, and adress more general control issues, wider than BZR's limitations. Directions are : expressiveness (taking into account in the LTS models value domains of the variables in the program) ; adaptive control (where the controller itself can dynamically switch between differents modes) ; distributed control (for classes of problems where communicating controllers can be designed) ; optimal control (w.r.t. weight functions, on states, transitions, and paths, with multicriteria techniques) ; timed and hybrid control bringing a new dimension for modeling and control, giving solutions where discrete models fail.

## 3.2. Design and programming for autonomic computing

### *3.2.1. Reactive programming*

Autonomic systems are intrinsically reconfigurable. To describe, specify or design these systems, there is a need to take into account this reconfigurability, within the programming languages used. We propose to consider the reconfigurability of systems from the angle of two properties: the notion of time, as we want to describe the state and behavior of the system before, and after its reconfiguration; the notion of dynamicity of the system, i.e., considering that the system's possible behaviors throughout execution are not completely known, neither at design-time nor at initial execution state. To describe and design such reactive systems, we propose to use the synchronous paradigm. It has been successfully used, in industry, for the design of embedded systems. It allows the description of behaviors based on a specific model of time (discrete time scale, synchronous parallel composition), providing properties which are important w.r.t. the safety of the described system: reactivity, determinism, preservation of safety properties by parallel composition (with other parts of the system or with its environment). Models and languages for control, proposed in this framework, provide designers, experts of the application domain, with a user-friendly access to highly technical formal methods of DCS, by encapsulating them in the compilation of concrete programming languages, generating concrete executable code. They are based on discrete models, but also support programming of sampled continuous controllers.

### *3.2.2. Component-based approach and domain-specific languages*

For integration of the previous control kernels into wider frameworks of reconfigurable systems, they have to be integrated in a design flow, and connected on the one side with higher-level specification languages (with help of DSLs), and on the other side with the generated code level target execution machines. This calls for the adoption of a component-based approach with necessary features, available typically in Fractal, for explicitly identifying the control interfaces and mechanisms.

Structuring and instrumentation for controllability will involve encapsulation of computations into components, specification of their local control (activation, reconfiguration, suspension, termination), and exporting appropriate interfaces (including behavior abstraction). Modeling the configurations space requires determining the controlled aspects (e.g., heterogenous CPUs loads, fault-tolerance and variability, memory, energy/power consumption, communication/bandwidth, QoS level) and their control points, as well as APIs for monitors and actions. Compilation and execution will integrate this in a complete design flow involving : extraction of a reactive model from components; instrumentation of execution platforms to be controllable; combination with other controllers; general "glue" and wrapper code.

Integration of reactive languages and control techniques in component-based systems brings interesting questions of co-existence w.r.t. other approaches like Event-Condition-Action (ECA) rules, or Complex Event Processing (CPE).

## 3.3. Infrastructure-level support for autonomic computing

The above general kernel of model-based control techniques can be used in a range of different computing infrastructures, representing complementary targets and abstraction levels, exploring the two axes :

- from hardware, to operating system/virtual machine, to middleware, to applications/service level;
- across different criteria for adaptation: resources and energy, quality of service, dependability.

### *3.3.1. Software and adaptive systems*

Autonomic administration loops at operating systems or middleware level are already very widespread. An open problem remains in design techniques for controllers with predictability and safety, e.g. w.r.t. the reachable states. We want to contribute to the topic of discrete control techniques for these systems, and tackle e.g. problems of coordination of multiple autonomic loops in data-centers, as in the ANR project CtrlGreen. Another target application is the control of clusters in map-reduce applications. The objective is to use continuous time control in order to tune finely the number of required clusters for an application running on a map-reduce server. This will use results of the ANR project MyCloud that enables to simulate clients on a real map-reduce server. On a longer term, we are interested in control problems in administration loops of event-based virtual machines, or in the deployment of massively parallel computation of the Cloud.

### *3.3.2. Hardware and reconfigurable architectures*

Reconfigurable architectures based on Field Programmable Gate Arrays (FPGA) are an active research area, where infrastructures are more and more supportive of reconfiguration, but its correct control remains an important issue. Work has begun in the ANR Famous project on identifying domain-specific control criteria and objectives, monitors and management APIs, and on integrating control techniques in the high-level RecoMARTE environment. On a longer term, we want to work on methods and tools for the programming of **multicore architectures**, exploiting the reconfigurability potentials and issues (because of variability, loss of cores), e.g. in our cooperation with ST Microelectronics, using a Fractal-based programming framework in the P2012 project, and in cooperation with Inria Lille (Adam), or with the CEA and TIMA on integrating control loops in the architecture for a fine control of the energy and of the required nodes for running a given application task.

### *3.3.3. Applications and autonomic systems*

In autonomic systems, control systems remain a lively source of inspiration, partly because the notion of control loop implementation is known and practiced naturally. On a wider scale, we started a cooperation with Orange Labs on "intelligent" building automation and control for the Smart Grid, through modeling and control of appliances w.r.t. their power consumption modes, at home, building, and city levels. Other partners on these topics are CEA LETI/DACLE and Schneider Electric.

We could explore more systems and applications e.g., Human-Machine Interfaces, or the orchestration of services. They can help design more general solutions, and result in a more complete methodology.

**MIMOVE Team**

# 3. Research Program

## 3.1. Introduction

MiMove targets research enabling next-generation mobile distributed systems, from their conception and design to their runtime support. These systems are challenged by their own success and consequent massive growth, as well as by the present and future, fast evolving, global networking and computing environment. This context is well-captured by the Future Internet vision, whose mobile constituents are becoming the norm rather than the exception. MiMove's research topics relate to a number of scientific domains with intensive ongoing research, such as ubiquitous computing, self-adaptive systems, wireless sensor networks, participatory sensing and social networks. In the following, we discuss related state-of-the-art research – in particular work focusing on middleware for mobile systems – and we identify the open research challenges that drive our work.

## 3.2. Emergent mobile distributed systems

Emergent mobile distributed systems promise to provide solutions to the complexity of the current and future computing and networking environments as well as to the ever higher demand for ubiquitous mobile applications, in particular being a response to the volatile and evolving nature of both the former and the latter. Hence, such systems have gained growing interest in the research literature. Notably, research communities have been formed around *self-adaptive systems* and *autonomic systems*, for which various overlapping definitions exist [72]. Self-adaptive systems are systems that are able to adapt themselves to uncertain execution environments, while autonomic systems have been defined as having one or more characteristics known as *self-** properties, including self-configuring, self-healing, self-optimizing and self-protecting [54]. Self-adaptive or autonomic systems typically include an adaptation loop comprising *modeling*, *monitoring*, *analyzing*, *deciding* and *enactment* processes. The adaptation loop provides feedback about changes in the system and its environment to the system itself, which adjusts itself in response. Current research on emergent distributed systems, including mobile ones, addresses all the dimensions of the adaptation loop [31], [25], [61], [83].

In our previous work, we introduced the paradigm of *emergent middleware*, which enables networked systems with heterogeneous behaviors to coordinate through adequate interaction protocols that emerge in an automated way [50], [28], [26]. A key point of that work is the combined study of the application- and middleware-layer behaviors, while current efforts in the literature tend to look only at one layer, either the application [48] or the middleware [19], [49], and take the other for granted (i.e., homogeneous, allowing direct coordination). Furthermore, the uncertainty of the computing and networking environments that is intrinsic to emergent mobile distributed systems [41] calls for taking into account also the underlying network and computational resources in a cross-layer fashion. In another line of work, we studied cross-integration of heterogeneous interaction paradigms at the middleware layer (message passing versus event-based and data sharing), where we investigate functional and QoS semantics of paradigms across their interconnections [43], [53]. Our focus there is to grasp the relation between individual and end-to-end semantics when bridging heterogeneous interaction protocols. In contrast, existing research efforts typically focus on emergent or evolving properties in homogeneous settings [42]. Last but not least, integrating heterogeneous mobile distributed systems into emergent compositions raises the question of dependability. More specifically, the overall correctness of the composition with respect to the individual requirements of the constituent systems can be particularly hard to ensure due to their heterogeneity. Again, current approaches typically deal with homogeneous constraints for dependability [39], [85], [40] with few exceptions [38].

As evident from the above, there is considerable interest and intensive research on emergent mobile distributed systems, while at the same time there are key research questions that remain open despite initial relevant work, including ours, which are summarized in the following:

- How to effectively deal with the combined impact on emergent properties of the different functional layers of mobile distributed systems (e.g., [50], [28], [26], [69])?
- How to perceive and model emergent properties in space and in time across volatile compositions of heterogeneous mobile distributed systems (e.g., [43], [53])?
- How to produce dependable emergent mobile distributed systems, i.e., systems that correctly meet their requirements, despite uncertainty in their emergence and execution exacerbated by heterogeneity (e.g., [38])?

## 3.3. Large-scale mobile sensing and actuation

In the past decade, the increasingly low cost of MEMS [0] devices and low-power microprocessors has led to a significant amount of research into mobile sensing and actuation. The results of this are now reaching the general public, going beyond the largely static use of sensors in scenarios such as agriculture and waste-water management, into increasingly *mobile* systems. These include sensor-equipped smartphones and personal wearable devices focused on the idea of a "quantified self", gathering data about a user's daily habits in order to enable them to improve their well-being. However, in spite of significant advances, the key challenges of these systems arise from largely the same attributes as those of early envisioned mobile systems, introduced in [76] and re-iterated in [75]: relative resource-poverty in terms of computation and communication, variable and unreliable connectivity, and limitations imposed by a finite energy source. These remain true even though modern mobile devices are significantly more powerful compared to their ancestors; the work we expect them to do has increased, and the computation and storage abilities available through fixed infrastructure such as the cloud are larger by order of magnitudes than any single mobile device. The design of algorithms and protocols to efficiently coordinate the sensing, processing, and actuation capabilities of the large number of mobile devices in future systems is a core area of MiMove's research.

Precisely, the focus of MiMove's research interests lies mostly in the systems resulting from the increased popularity of sensor-equipped smart devices that are carried by people, which has led to the promising field of *mobile phone sensing* or *mobile crowd-sensing* [58], [55]. The paradigm is powerful, as it allows overcoming the inherent limitation of traditional sensing techniques that require the deployment of dedicated fixed sensors (e.g., see work on noise mapping using the microphones in users' telephones [70]). Specifically, we are interested in the challenges below, noting that initial work to address them already exists, including that by team members:

- How to efficiently manage the large scale that will come to the fore when millions, even billions of devices will need to be managed and queried simultaneously (e.g., [81], [45])?
- How to efficiently coordinate the available devices, including resource-poor mobile devices and the more-capable cloud infrastructure (e.g., [68], [36], [74], [64])?
- How to guarantee dependability in a mobile computing environment (e.g., [34], [80], [30])?
- How to ensure that the overhead of sensing does not lead to a degraded performance for the user (e.g., [56], [36])?

## 3.4. Mobile social crowd-sensing

Mobile crowd-sensing as introduced in Section 3.3 is further undergoing a transformation due to the widespread adoption of social networking. The resulting mobile *social* crowd-sensing may be qualified as "*people-centric sensing*" and roughly subdivides into two categories [57]: i) *participatory sensing*, and ii) *opportunistic sensing*. Participatory sensing entails direct involvement of humans controlling the mobile devices, while opportunistic sensing requires the mobile device itself to determine whether or not to perform

[0] Micro-Electro-Mechanical Systems.

the sensing task. Orthogonally to the above categorization, mobile sensing can be [55]: i) *personal sensing*, mostly to monitor a person's context and well-being; ii) *social sensing*, where updates are about the social and emotional statuses of individuals; or iii) *urban (public) sensing*, where public data is generated by the public and for the public to exploit. Personal sensing is aimed towards personal monitoring and involves one or just a few devices in direct relationship with their custodian. For instance, SoundSense [62] is a system that enables each person's mobile device to learn the types of sounds the owner encounters through unsupervised learning. Another application example relates to the sensing-based detection of the users' transportation mode by using their smartphones [47]. In social sensing, the mobile device or its owner decides what social information to share about the owner or the owner's environment, with an individual or group of friends [55], [37], [52], [21], [66]. Social sensing is mostly participatory. Therefore, it is the custodian of the device who determines when and where data should be generated. Social participatory sensing is closely related to social networking [63]. On the other hand, within opportunistic social sensing, the underlying system is in charge of acquiring needed data through relevant probes, as opposed to having the end-user providing them explicitly [24], [51], [22]. In urban sensing, also known as public sensing, data can be generated by everyone (or their devices) and exploited by everyone for public knowledge, including environment monitoring, or traffic updates [55]. In participatory urban sensing, users participate in providing information about the environment by exploiting the sensors/actuators embedded in their devices (which can be smartphones, vehicles, tablets, etc.) [55]. However data is only generated according to the owner's willingness to participate. Participatory urban sensing is especially characterized by scale issues at the data level, where data is generated by numerous individuals and should be processed and aggregated for knowledge to be inferred, involving adequate data scaling approaches [44]. Ikarus [84] is an example of participatory sensing, where data is collected by a large number of paragliders throughout their flights. The focus is on aggregating the data and rendering the results on a thermal map.

As outlined above, mobile social crowd-sensing has been a very active field of research for the last few years with various applications being targeted. However, effectively enabling mobile social crowd-sensing still raises a number of challenges, for which some early work may be identified:

- How to ensure that the system delivers the right quality of service, e.g., in terms of user-perceived delay, in spite of the resource constraints of mobile systems (e.g., [71])?
- How to guarantee the right level of privacy (e.g., [33], [73])?
- How to ensure the right level of participation from end-users so that mobile sensing indeed becomes a relevant source of accurate knowledge, which relates to eliciting adequate incentive mechanisms [86], in particular based on the understanding of mobile application usage [78], [77]?
- How to enrich sensor-generated content that is quantitative with user-generated one, thereby raising the issue of leveraging highly unstructured data while benefiting from a rich source of knowledge (e.g., sensing the crowdedness of a place combined with the feeling of people about the crowdedness, which may hint on the place's popularity as much as on discomfort)?

**MYRIADS Project-Team**

# 3. Research Program

## 3.1. Introduction

In this section, we present our research challenges along four work directions: resource and application management in distributed cloud architectures for scaling clouds in Section 3.2 , energy management strategies for greening clouds in Section 3.3 , security and data protection aspects for securing cloud-based information systems and applications in Section 3.4 , and methods for experimenting with clouds in Section 3.5 .

## 3.2. Scaling clouds

### *3.2.1. Resource management in hierarchical clouds*

The next generation of utility computing appears to be an evolution from highly centralized clouds towards more decentralized platforms. Today, cloud computing platforms mostly rely on large data centers servicing a multitude of clients from the edge of the Internet. Servicing cloud clients in this manner suggests that locality patterns are ignored: wherever the client issues his/her request from, the request will have to go through the backbone of the Internet provider to the other side of the network where the data center relies. Besides this extra network traffic and this latency overhead that could be avoided, other common centralization drawbacks in this context stand in limitations in terms of security/legal issues and resilience.

At the same time, it appears that network backbones are over-provisioned for most of their usage. This advocates for placing computing resources directly within the backbone network. The general challenge of resource management for such clouds stands in trying to be locality-aware: for the needs of an application, several virtual machines may exchange data. Placing them *close* to each others can significantly improve the performance of the application they compose. More generally, building an overlay network which takes the hierarchical aspects of the platform without being a hierarchical overlay – which comes with load balancing and resilience issues is a challenge by itself.

The results of these works are planned to be integrated into the Discovery initiative [52] which aims at revisiting OpenStack to offer a cloud stack able to manage utility computing platforms where computing resources are located in small computing centers in the backbone's PoPs (Point of Presence) and interconnected through the backbone's internal links.

### *3.2.2. Resource management in mobile edge clouds*

Mobile edge cloud (MEC) infrastructures are composed of compute, storage and networking resources located at the edge of wide-area networks, in immediate proximity to the end users. Instead of treating the mobile operator's network as a high-latency dumb pipe between the end users and the external service providers, MEC platforms aim at deploying cloud functionalities *within* the mobile phone network, inside or close to the mobile access points. Doing so is expected to deliver added value to the content providers and the end users by enabling new types of applications ranging from Internet-of-Things applications to extremely interactive systems (e.g., augmented reality). Simultaneously, it will generate extra revenue streams for the mobile network operators, by allowing them to position themselves as cloud computing operators and to rent their already-deployed infrastructure to content and application providers.

Mobile edge clouds have very different geographical distribution compared to traditional clouds. While traditional clouds are composed of many reliable and powerful machines located in a very small number of data centers and interconnected by very high-speed networks, mobile edge cloud are composed of a very large number of points-of-presence with a couple of weak and potentially unreliable servers, interconnected with each other by commodity long-distance networks. This creates new demands for the organization of a scalable mobile edge computing infrastructure, and opens new directions for research.

The main challenges that we plan to address are:

- How should an edge cloud infrastructure be designed such that it remains scalable, fault-tolerant, controllable, energy-efficient, etc.?
- How should applications making use of edge clouds be organized? One promising direction is to explore the extent to which stream-data processing platforms such as Apache Spark and Apache Flink can be adapted to become one of the main application programming paradigms in such environments.

### *3.2.3. Self-optimizing applications in multi-cloud environments*

As the use of cloud computing becomes pervasive, the ability to deploy an application on a multi-cloud infrastructure becomes increasingly important. Potential benefits include avoiding dependence on a single vendor, taking advantage of lower resource prices or resource proximity, and enhancing application availability. Supporting multi-cloud application management involves two tasks. First, it involves selecting an initial multi-cloud application deployment that best satisfies application objectives and optimizes performance and cost. Second, it involves dynamically adapting the application deployment in order to react to changes in execution conditions, application objectives, cloud provider offerings, or resource prices. Handling price changes in particular is becoming increasingly complex. The reason is the growing trend of providers offering sophisticated, dynamic pricing models that allow buying and selling resources of finer granularities for shorter time durations with varying prices.

Although multi-cloud platforms are starting to emerge, these platforms impose a considerable amount of effort on developers and operations engineers, provide no support for dynamic pricing, and lack the responsiveness and scalability necessary for handling highly-distributed, dynamic applications with strict quality requirements. The goal of this work is to develop techniques and mechanisms for automating application management, enabling applications to cope with and take advantage of the dynamic, diverse, multi-cloud environment in which they operate.

The main challenges arising in this context are:

- selecting effective decision-making approaches for application adaptation,
- supporting scalable monitoring and adaptation across multiple clouds,
- performing adaptation actions in a cost-efficient and safe manner.

## 3.3. Greening clouds

ICT (Information and Communications Technologies) ecosystem now approaches 5% of world electricity consumption and this ICT energy use will continue grow fast because of the information appetite of Big Data, big networks and big infrastructures as Clouds that unavoidably leads to big power.

### *3.3.1. Smart grids and clouds*

We propose exploiting Smart Grid technologies to come to the rescue of energy-hungry Clouds. Unlike in traditional electrical distribution networks, where power can only be moved and scheduled in very limited ways, Smart Grids dynamically and effectively adapt supply to demand and limit electricity losses (currently 10% of produced energy is lost during transmission and distribution).

For instance, when a user submits a Cloud request (such as a Google search for instance), it is routed to a data center that processes it, computes the answer and sends it back to the user. Google owns several data centers spread across the world and for performance reasons, the center answering the user's request is more likely to be the one closest to the user. However, this data center may be less energy efficient. This request may have consumed less energy, or a different kind of energy (renewable or not), if it had been sent to this further data center. In this case, the response time would have been increased but maybe not noticeably: a different trade-off between quality of service (QoS) and energy-efficiency could have been adopted.

While Clouds come naturally to the rescue of Smart Grids for dealing with this big data issue, little attention has been paid to the benefits that Smart Grids could bring to distributed Clouds. To our knowledge, no previous work has exploited the Smart Grids potential to obtain and control the energy consumption of entire Cloud infrastructures from underlying facilities such as air conditioning equipment (which accounts for 30% to 50% of a data center's electricity bill) to network resources (which are often operated by several actors) and to computing resources (with their heterogeneity and distribution across multiple data centers). We aim at taking advantage of the opportunity brought by the Smart Grids to exploit renewable energy availability and to optimize energy management in distributed Clouds.

### *3.3.2. Energy cost models*

Cloud computing allows users to outsource the computer resources required for their applications instead of using a local installation. It offers on-demand access to the resources through the Internet with a pay-as-you-go pricing model.However, this model hides the electricity cost of running these infrastructures.

The costs of current data centers are mostly driven by their energy consumption (specifically by the air conditioning, computing and networking infrastructure). Yet, current pricing models are usually static and rarely consider the facilities' energy consumption per user. The challenge is to provide a fair and predictable model to attribute the overall energy costs per virtual machine and to increase energy-awareness of users.

Another goal consists in better understanding the energy consumption of computing and networking resources of Clouds in order to provide energy cost models for the entire infrastructure including incentivizing cost models for both Cloud providers and energy suppliers. These models will be based on experimental measurement campaigns on heterogeneous devices. Inferring a cost model from energy measurements is an arduous task since simple models are not convincing, as shown in our previous work. We aim at proposing and validating energy cost models for the heterogeneous Cloud infrastructures in one hand, and the energy distribution grid on the other hand. These models will be integrated into simulation frameworks in order to validate our energy-efficient algorithms at larger scale.

### *3.3.3. Energy-aware users*

In a Cloud moderately loaded, some servers may be turned off when not used for energy saving purpose. Cloud providers can apply resource management strategies to favor idle servers. Some of the existing solutions propose mechanisms to optimize VM scheduling in the Cloud. A common solution is to consolidate the mapping of the VMs in the Cloud by grouping them in a fewer number of servers. The unused servers can then be turned off in order to lower the global electricity consumption.

Indeed, current work focuses on possible levers at the virtual machine suppliers and/or services. However, users are not involved in the choice of using these levers while significant energy savings could be achieved with their help. For example, they might agree to delay slightly the calculation of the response to their applications on the Cloud or accept that it is supported by a remote data center, to save energy or wait for the availability of renewable energy. The VMs are black boxes from the Cloud provider point of view. So, the user is the only one to know the applications running on her VMs.

We plan to explore possible collaborations between virtual machine suppliers, service providers and users of Clouds in order to provide users with ways of participating in the reduction of the Clouds energy consumption. This work will follow two directions: 1) to investigate compromises between power and performance/service quality that cloud providers can offer to their users and to propose them a variety of options adapted to their workload; and 2) to develop mechanisms for each layer of the Cloud software stack to provide users with a quantification of the energy consumed by each of their options as an incentive to become greener.

## 3.4. Securing clouds

### *3.4.1. Security monitoring SLO*

While the trend for companies to outsource their information system in clouds is confirmed, the problem of securing an information system becomes more difficult. Indeed, in the case of infrastructure clouds, physical

resources are shared between companies (also called tenants) but each tenant controls only parts of the shared resources, and, thanks to virtualization, the information system can be dynamically and automatically reconfigured with added or removed resources (for example starting or stopping virtual machines), or even moved between physical resources (for example using virtual machine migration). Partial control of shared resources brings new classes of attacks between tenants, and security monitoring mechanisms to detect such attacks are better placed out of the tenant-controlled virtual information systems, that is under control of the cloud provider. Dynamic and automatic reconfigurations of the information system make it unfeasible for a tenant's security administrator to setup the security monitoring components to detect attacks, and thus an automated self-adaptable security monitoring service is required.

Combining the two previous statements, there is a need for a dependable, automatic security monitoring service provided to tenants by the cloud provider. Our goal is to address the following challenges to design such a security monitoring service:

1. to define relevant Service-Level Objectives (SLOs) of a security monitoring service, that can figure in the Service-Level Agreement (SLA) signed between a cloud provider and a tenant;
2. to design heuristics to automatically configure provider-controlled security monitoring software components and devices so that SLOs are reached, even during automatic reconfigurations of tenants' information systems;
3. to design evaluation methods for tenants to check that SLOs are reached.

Moreover in challenges 2 and 3 the following sub-challenges must be addressed:

- although SLAs are bi-lateral contracts between the provider and each tenant, the implementation of the contracts is based on shared resources, and thus we must study methods to combine the SLOs;
- the designed methods should have a minimal impact on performance.

### *3.4.2. Data Protection in Cloud-based IoT Services*

The Internet of Things is becoming a reality. Individuals have their own swarm of connected devices (e.g. smartphone, wearables, and home connected objects) continually collecting personal data. A novel generation of services is emerging exploiting data streams produced by the devices' sensors. People are deprived of control of their personal data as they don't know precisely what data are collected by service providers operating on Internet (oISP), for which purpose they could be used, for how long they are stored, and to whom they are disclosed. In response to privacy concerns the European Union has introduced, with the Global Data Protection Regulation (GDPR), new rules aimed at enforcing the people's rights to personal data protection. The GDPR also gives strong incentives to oISPs to comply. However, today, oISPs can't make their systems GDPR-compliant since they don't have the required technologies. We argue that a new generation of system is mandatory for enabling oISPs to conform to the GDPR. We plan to to design an open source distributed operating system for native implementation of new GDPR rules and ease the programming of compliant cloud-based IoT services. Among the new rules, transparency, right of erasure, and accountability are the most challenging ones to be implemented in IoT environments but could fundamentally increase people's confidence in oISPs. Deployed on individuals' swarms of devices and oISPs' cloud-hosted servers, it will enforce detailed data protection agreements and accountability of oISPs' data processing activities. Ultimately we will show to what extend the new GDPR rules can be implemented for cloud-based IoT services.

## 3.5. Experimenting with Clouds

Cloud platforms are challenging to evaluate and study with a sound scientific methodology. As with any distributed platform, it is very difficult to gather a global and precise view of the system state. Experiments are not reproducible by default since these systems are shared between several stakeholder. This is even worsen by the fact that microscopic differences in the experimental conditions can lead to drastic changes since typical Cloud applications continuously adapt their behavior to the system conditions.

### 3.5.1. Experimentation methodologies for clouds

We propose to combine two complementary experimental approaches: direct execution on testbeds such as Grid'5000, that are eminently believable but rather labor intensive, and simulations (using *e.g.* SimGrid) that are much more light-weighted, but requires are careful assessment. One specificity of the Myriads team is that we are working on these experimental methodologies *per se*, raising the standards of *good experiments* in our community.

We plan to make SimGrid widely usable beyond research laboratories, in order to evaluate industrial systems and to teach the future generations of cloud practitioners. This requires to frame the specific concepts of Cloud systems and platforms in actionable interfaces. The challenge is to make the framework both easy to use for simple studies in educational settings while modular and extensible to suit the specific needs of every advanced industrial-class users.

We aim at leveraging the convergence opportunities between methodologies by further bridging simulation and real testbeds. The predictions obtained from the simulator should be validated against some real-world experiments obtained on the target production platform, or on a similar platform. This (in)validation of the predicted results often improves the understanding of the modeled system. On the other side, it may even happen that the measured discrepancies are due to some mis-configuration of the real platform that would have been undetected without this (in)validation study. In that sense, the simulator constitutes a precious tool for the quality assurance of real testbeds such as Grid'5000.

Scientists need more help to make there Cloud experiments fully reproducible, in the sprit of Open Science exemplified by the HAL Open Archive, actively backed by Inria. Users still need practical solutions to archive, share and compare the whole experimental settings, including the raw data production (particularly in the case of real testbeds) and their statistical analysis. This is a long lasting task to which we plan to collaborate through the research communities gathered around the Grid'5000 and SimGrid scientific instruments.

Finally, since correction and performance can constitute contradictory goals, it is particularly important to study them jointly. To that extend, we want to bridge the performance studies, that constitute our main scientific heritage, to correction studies leveraging formal techniques. SimGrid already includes to exhaustively explore the possible executions. We plan to continue this work to ease the use of the relevant formal methods to the experimenter studying Cloud systems.

### 3.5.2. Use cases

In system research it is important to work on real-world use cases from which we extract requirements inspiring new research directions and with which we can validate the system services and mechanisms we propose. In the framework of our close collaboration with the Data Science Technology department of the LBNL, we will investigate cloud usage for scientific data management. Next-generation scientific discoveries are at the boundaries of datasets, e.g., across multiple science disciplines, institutions and spatial and temporal scales. Today, data integration processes and methods are largely adhoc or manual. A generalized resource infrastructure that integrates knowledge of the data and the processing tasks being performed by the user in the context of the data and resource lifecycle is needed. Clouds provide an important infrastructure platform that can be leveraged by including knowledge for distributed data integration.

**REGAL Project-Team**

# 3. Research Program

## 3.1. Research rationale

The research of Regal addresses both theoretical and practical issues of *Computer Systems*, i.e., its goal is a dual expertise in theoretical and experimental research. Our approach is a "virtuous cycle" of algorithm design triggered by issues with real systems, which we prove correct and evaluate theoretically, and then eventually implement and test experimentally.

Regal's major challenges comprise communication, sharing of information, and correct execution in large-scale and/or highly dynamic computer systems. While Regal's historically focused in static distributed systems, since some years ago we have covered a larger spectrum of distributed computer systems: multicore computers, clusters, mobile networks, peer-to-peer systems, cloud computing systems, and other communicating entities such as swarms of robots. This holistic approach allows the handling of related problems at different levels. Among such problems we can highlight communication between cores, consensus, fault detection, scalability, search and diffusion of information, allocation resource, replication and consistency of shared data, dynamic content distribution, and multi-core concurrent algorithms.

Computer Systems is a rapidly evolving domain, with strong interactions with industry and modern computer systems, which are increasingly distributed. Ensuring persistence, availability, and consistency of data in a distributed setting is a major requirement: the system must remain correct despite slow networks, disconnection, crashes, failures, churn, and attacks. Easiness of use, performance, and efficiency are equally fundamental. However, these requirements are somewhat conflicting, and there are many algorithmic and engineering trade-offs, which often depend on specific workloads or usage scenarios. At the same time, years of research in distributed systems are now coming to fruition, and are being used by millions of users of web systems, peer-to-peer systems, gaming and social applications, or cloud computing. These new usages bring new challenges of extreme scalability and adaptation to dynamically-changing conditions, where knowledge of the system state might only be partial and incomplete. Therefore, the scientific challenges of the distributed computing systems listed above are subject to additional trade-offs which include scalability, fault tolerance, dynamics, and virtualization of physical infrastructure. Algorithms designed for traditional distributed systems, such as resource allocation, data storage and placement, and concurrent access to shared data, need to be redefined or revisited in order to work properly under the constraints of these new environments.

In in particular, Regal focuses on three key challenges:

- the adaptation of algorithms to the new dynamics of distributed systems;
- data management on extreme large configurations;
- the adaptation of execution support to new multi-core architectures.

We should emphasize that these challenges are complementary: the two first challenges aim at building new distributed algorithms and strategies for large and dynamic distributed configurations whereas the last one focusses on the scalability of internal OS mechanisms.

SPIRALS Project-Team

# 3. Research Program

## 3.1. Introduction

Our research program on self-adaptive software targets two key properties that are detailed in the remainder of this section: *self-healing* and *self-optimization*.

## 3.2. Objective #1: Self-healing - Mining software artifacts to automatically evolve systems

Software systems are under the pressure of changes all along their lifecycle. Agile development blurs the frontier between design and execution and requires constant adaptation. The size of systems (millions of lines of code) multiplies the number of bugs by the same order of magnitude. More and more systems, such as sensor network devices, live in "surviving" mode, in the sense that they are neither rebootable nor upgradable.

Software bugs are hidden in source code and show up at development-time, testing-time or worse, once deployed in production. Except for very specific application domains where formal proofs are achievable, bugs can not be eradicated. As an order of magnitude, on 16 Dec 2011, the Eclipse bug repository contains 366,922 bug reports. Software engineers and developers work on bug fixing on a daily basis. Not all developers spend the same time on bug fixing. In large companies, this is sometimes a full-time role to manage bugs, often referred to as *Quality Assurance* (QA) software engineers. Also, not all bugs are equal, some bugs are analyzed and fixed within minutes, others may take months to be solved [75].

In terms of research, this means that: (i) one needs means to automatically adapt the design of the software system through automated refactoring and API extraction, (ii) one needs approaches to automate the process of adapting source code in order to fix certain bugs, (iii) one needs to revisit the notion of error-handling so that instead of crashing in presence of errors, software adapts itself to continue with its execution, *e.g.*, in degraded mode.

There is no one-size-fits-all solution for each of these points. However, we think that novel solutions can be found by using **data mining and machine learning techniques tailored for software engineering** [76]. This body of research consists of mining some knowledge about a software system by analyzing the source code, the version control systems, the execution traces, documentation and all kinds of software development and execution artifacts in general. This knowledge is then used within recommendation systems for software development, auditing tools, runtime monitors, frameworks for resilient computing, etc.

The novelty of our approach consists of using and tailoring data mining techniques for analyzing software artifacts (source code, execution traces) in order to achieve the **next level of automated adaptation** (*e.g.*, automated bug fixing). Technically, we plan to mix unsupervised statistical learning techniques (*e.g.* frequent item set mining) and supervised ones (*e.g.* training classifiers such as decision trees). This research is currently not being performed by data mining research teams since it requires a high level of domain expertise in software engineering, while software engineering researchers can use off-the-shelf data mining libraries, such as Weka [61].

We now detail the two directions that we propose to follow to achieve this objective.

### *3.2.1. Learning from software history how to design software and fix bugs*

The first direction is about mining techniques in software repositories (*e.g.*, CVS, SVN, Git). Best practices can be extracted by data mining source code and the version control history of existing software systems. The design and code of expert developers significantly vary from the artifacts of novice developers. We will learn to differentiate those design characteristics by comparing different code bases, and by observing the semantic refactoring actions from version control history. Those design rules can then feed the test-develop-refactor constant adaptation cycle of agile development.

**Fault localization of bugs reported in bug repositories.** We will build a solid foundation on empirical knowledge about bugs reported in bug repository. We will perform an empirical study on a set of representative bug repositories to identify classes of bugs and patterns of bug data. For this, we will build a tool to browse and annotate bug reports. Browsing will be helped with two kinds of indexing: first, the tool will index all textual artifacts for each bug report; second it will index the semantic information that is not present by default in bug management software—*i.e.*, "contains a stacktrace"). Both indexes will be used to find particular subsets of bug reports, for instance "all bugs mentioning invariants and containing a stacktrace". Note that queries with this kind of complexity and higher are mostly not possible with the state-of-the-art of bug management software. Then, analysts will use annotation features to annotate bug reports. The main outcome of the empirical study will be the identification of classes of bugs that are appropriate for automated localization. Then, we will run machine learning algorithms to identify the latent links between the bug report content and source code features. Those algorithms would use as training data the existing traceability links between bug reports and source code modifications from version control systems. We will start by using decision trees since they produce a model that is explicit and understandable by expert developers. Depending on the results, other machine learning algorithms will be used. The resulting system will be able to locate elements in source code related to a certain bug report with a certain confidence.

**Automated bug fix generation with search-based techniques.** Once a location in code is identified as being the cause of the bug, we can try to automatically find a potential fix. We envision different techniques: (1) infer fixes from existing contracts and specifications that are violated; (2) infer fixes from the software behavior specified as a test suite; (3) try different fix types one-by-one from a list of identified bug fix patterns; (4) search fixes in a fix space that consists of combinations of atomic bug fixes. Techniques 1 and 2 are explored in [58] and [74]. We will focus on the latter techniques. To identify bug fix patterns and atomic bug fixes, we will perform a large-scale empirical study on software changes (also known as changesets when referring to changes across multiple files). We will develop tools to navigate, query and annotate changesets in a version control system. Then, a grounded theory will be built to master the nature of fixes. Eventually, we will decompose change sets in atomic actions using clustering on changeset actions. We will then use this body of empirical knowledge to feed search-based algorithms (*e.g.* genetic algorithms) that will look for meaningful fixes in a large fix space. To sum up, our research on automated bug fixing will try not only to point to source code locations responsible of a bug, but to search for code patterns and snippets that may constitute the skeleton of a valid patch. Ultimately, a blend of expert heuristics and learned rules will be able to produce valid source code that can be validated by developers and committed to the code base.

### *3.2.2. Run-time self-healing*

The second proposed research direction is about inventing a self-healing capability at run-time. This is complementary to the previous objective that mainly deals with development time issues. We will achieve this in two steps. First, we want to define frameworks for resilient software systems. Those frameworks will help to maintain the execution even in the presence of bugs—*i.e.* to let the system survive. As exposed below, this may mean for example to switch to some degraded modes. Next, we want to go a step further and to define solutions for automated runtime repair, that is, not simply compensating the erroneous behavior, but also determining the correct repair actions and applying them at run-time.

**Mining best effort values.** A well-known principle of software engineering is the "fail-fast" principle. In a nutshell, it states that as soon as something goes wrong, software should stop the execution before entering incorrect states. This is fine when a human user is in the loop, capable of understanding the error or at least rebooting the system. However, the notion of "failure-oblivious computing" [68] shows that in certain domains, software should run in a resilient mode (*i.e.* capable of recovering from errors) and/or best-effort mode—*i.e.* a slightly imprecise computation is better than stopping. Hence, we plan to investigate data mining techniques in order to learn best-effort values from past executions (*i.e.* somehow learning what is a correct state, or the opposite what is not a completely incorrect state). This knowledge will then be used to adapt the software state and flow in order to mitigate the error consequences, the exact opposite of fail-fast for systems with long-running cycles.

**Embedding search based algorithms at runtime.** Harman recently described the field of search-based software engineering [62]. We think that certain search based approaches can be embedded at runtime with the goal of automatically finding solutions that avoid crashing. We will create software infrastructures that allow automatically detecting and repairing faults at run-time. The methodology for achieving this task is based on three points: (1) empirical study of runtime faults; (2) learning approaches to characterize runtime faults; (3) learning algorithms to produce valid changes to the software runtime state. An empirical study will be performed to analyze those bug reports that are associated with runtime information (*e.g.* core dumps or stacktraces). After this empirical study, we will create a system that learns on previous repairs how to produce small changes that solve standard runtime bugs (*e.g.* adding an array bound check to throw a handled domain exception rather than a spurious language exception). To achieve this task, component models will be used to (1) encapsulate the monitoring and reparation meta-programs in appropriate components and (2) support runtime code modification using scripting, reflective or bytecode generation techniques.

## 3.3. Objective #2: Self-optimization - Sharing runtime behaviors to continuously adapt software

Complex distributed systems have to seamlessly adapt to a wide variety of deployment targets. This is due to the fact that developers cannot anticipate all the runtime conditions under which these systems are immersed. A major challenge for these software systems is to develop their capability to continuously reason about themselves and to take appropriate decisions and actions on the optimizations they can apply to improve themselves. This challenge encompasses research contributions in different areas, from environmental monitoring to real-time symptoms diagnosis, to automated decision making. The variety of distributed systems, the number of optimization parameters, and the complexity of decisions often resign the practitioners to design monolithic and static middleware solutions. However, it is now globally acknowledged that the development of dedicated building blocks does not contribute to the adoption of sustainable solutions. This is confirmed by the scale of actual distributed systems, which can—for example—connect several thousands of devices to a set of services hosted in the Cloud. In such a context, the lack of support for smart behaviours at different levels of the systems can inevitably lead to its instability or its unavailability. In June 2012, an outage of Amazon's Elastic Compute Cloud in North Virginia has taken down Netflix, Pinterest, and Instagram services. During hours, all these services failed to satisfy their millions of customers due to the lack of integration of a self-optimization mechanism going beyond the boundaries of Amazon.

The research contributions we envision within this area will therefore be organized as a reference model for engineering **self-optimized distributed systems** autonomously driven by *adaptive feedback control loops*, which will automatically enlarge their scope to cope with the complexity of the decisions to be taken. This solution introduces a multi-scale approach, which first privileges local and fast decisions to ensure the homeostasis [0] property of a single node, and then progressively propagates symptoms in the network in order to reason on a longer term and a larger number of nodes. Ultimately, domain experts and software developers can be automatically involved in the decision process if the system fails to find a satisfying solution. The research program for this objective will therefore focus on the study of mechanisms for **monitoring, taking decisions, and automatically reconfiguring software at runtime and at various scales**. As stated in the self-healing objective, we believe that there is no one-size-fits-all mechanism that can span all the scales of the system. We will therefore study and identify an optimal composition of various adaptation mechanisms in order to produce long-living software systems.

The novelty of this objective is to exploit the wisdom of crowds to define new middleware solutions that are able to continuously adapt software deployed in the wild. We intend to demonstrate the applicability of this approach to distributed systems that are deployed from mobile phones to cloud infrastructures. The key scientific challenges to address can be summarized as follows: *How does software behave once deployed in the wild? Is it possible to automatically infer the quality of experience, as it is perceived by users? Can the*

[0] Homeostasis is the property of a system that regulates its internal environment and tends to maintain a stable, relatively constant condition of properties [Wikipedia].

*runtime optimizations be shared across a wide variety of software? How optimizations can be safely operated on large populations of software instances?*

The remainder of this section further elaborates on the opportunities that can be considered within the frame of this objective.

### 3.3.1. Monitoring software in the wild

Once deployed, developers are generally no longer aware of how their software behave. Even if they heavily use testbeds and benchmarks during the development phase, they mostly rely on the bugs explicitly reported by users to monitor the efficiency of their applications. However, it has been shown that contextual artifacts collected at runtime can help to understand performance leaks and optimize the resilience of software systems [77]. Monitoring and understanding the context of software at runtime therefore represent the first building block of this research challenge. Practically, we intend to investigate crowd-sensing approaches, to smartly collect and process runtime metrics (*e.g.*, request throughput, energy consumption, user context). Crowd-sensing can be seen as a specific kind of crowdsourcing activity, which refers to the capability of lifting a (large) diffuse group of participants to delegate the task of retrieving trustable data from the field. In particular, crowd-sensing covers not only *participatory sensing* to involve the user in the sensing task (*e.g.*, surveys), but also *opportunistic sensing* to exploit mobile sensors carried by the user (*e.g.*, smartphones).

While reported metrics generally enclose raw data, the monitoring layer intends to produce meaningful indicators like the *Quality of Experience* (QoE) perceived by users. This QoE reflects representative symptoms of software requiring to trigger appropriate decisions in order to improve its efficiency. To diagnose these symptoms, the system has to process a huge variety of data including runtime metrics, but also history of logs to explore the sources of the reported problems and identify opportunities for optimizations. The techniques we envision at this level encompass machine learning, principal component analysis, and fuzzy logic [67] to provide enriched information to the decision level.

### 3.3.2. Collaborative decision-making approaches

Beyond the symptoms analysis, decisions should be taken in order to improve the *Quality of Service* (QoS). In our opinion, collaborative approaches represent a promising solution to effectively converge towards the most appropriate optimization to apply for a given symptom. In particular, we believe that exploiting the wisdom of the crowd can help the software to optimize itself by sharing its experience with other software instances exhibiting similar symptoms. The intuition here is that the body of knowledge that supports the optimization process cannot be specific to a single software instance as this would restrain the opportunities for improving the quality and the performance of applications. Rather, we think that any software instance can learn from the experience of others.

With regard to the state-of-the-art, we believe that a multi-levels decision infrastructure, inspired from distributed systems like Spotify [60], can be used to build a decentralized decision-making algorithm involving the surrounding peers before requesting a decision to be taken by more central control entity. In the context of collaborative decision-making, peer-based approaches therefore consist in quickly reaching a consensus on the decision to be adopted by a majority of software instances. Software instances can share their knowledge through a micro-economic model [56], that would weight the recommendations of experienced instances, assuming their age reflects an optimal configuration.

Beyond the peer level, the adoption of algorithms inspired from evolutionary computations, such as genetic programming, at an upper level of decision can offer an opportunity to test and compare several alternative decisions for a given symptom and to observe how does the crowd of applications evolves. By introducing some diversity within this population of applications, some instances will not only provide a satisfying QoS, but will also become naturally resilient to unforeseen situations.

### 3.3.3. Smart reconfigurations in the large

Any decision taken by the crowd requires to propagate back to and then operated by the software instances. While simplest decisions tend to impact software instances located on a single host (*e.g.*, laptop, smartphone),

this process can also exhibit more complex reconfiguration scenarios that require the orchestration of various actions that have to be safely coordinated across a large number of hosts. While it is generally acknowledged that centralized approaches raise scalability issues, we think that self-optimization should investigate different reconfiguration strategies to propagate and apply the appropriate actions. The investigation of such strategies can be addressed in two steps: the consideration of *scalable data propagation protocols* and the identification of *smart reconfiguration mechanisms*.

With regard to the challenge of scalable data propagation protocols, we think that research opportunities encompass not only the exploitation of gossip-based protocols [59], but also the adoption of publish/subscribe abstractions [64] in order to decouple the decision process from the reconfiguration. The fundamental issue here is the definition of a communication substrate that can accommodate the propagation of decisions with relaxed properties, inspired by *Delay Tolerant Networks* (DTN), in order to reach weakly connected software instances. We believe that the adoption of asynchronous communication protocols can provide the sustainable foundations for addressing various execution environments including harsh environments, such as developing countries, which suffer from a partial connectivity to the network. Additionally, we are interested in developing the principle of *social networks of applications* in order to seamlessly group and organize software instances according to their similarities and acquaintances. The underlying idea is that grouping application instances can contribute to the identification of optimization profiles not only contributing to the monitoring layer, but also interested in similar reconfigurations. Social networks of applications can contribute to the anticipation of reconfigurations by exploiting the symptoms of similar applications to improve the performance of others before that problems actually happen.

With regard to the challenge of smart reconfiguration mechanisms, we are interested in building on our established experience of adaptive middleware [72] in order to investigate novel approaches to efficient application reconfigurations. In particular, we are interested in adopting seamless micro-updates and micro-reboot techniques to provide in-situ reconfiguration of pieces of software. Additionally, the provision of safe and secured reconfiguration mechanisms is clearly a key issue that requires to be carefully addressed in order to avoid malicious exploitation of dynamic reconfiguration mechanisms against the software itself. In this area, although some reconfiguration mechanisms integrate transaction models [65], most of them are restricted to local reconfigurations, without providing any support for executing distributed reconfiguration transactions. Additionally, none of the approached published in the literature include security mechanisms to preserve from unauthorized or malicious reconfigurations.

**WHISPER Project-Team**

# 3. Research Program

## 3.1. Scientific Foundations

### *3.1.1. Program analysis*

A fundamental goal of the research in the Whisper team is to elicit and exploit the knowledge found in existing code. To do this in a way that scales to a large code base, systematic methods are needed to infer code properties. We may build on either static [33], [36], [39] or dynamic analysis [57], [61], [67]. Static analysis consists of approximating the behavior of the source code from the source code alone, while dynamic analysis draws conclusions from observations of sample executions, typically of test cases. While dynamic analysis can be more accurate, because it has access to information about actual program behavior, obtaining adequate test cases is difficult. This difficulty is compounded for infrastructure software, where many, often obscure, cases must be handled, and external effects such as timing can have a significant impact. Thus, we expect to primarily use static analyses. Static analyses come in a range of flavors, varying in the extent to which the analysis is *sound*, *i.e.*, the extent to which the results are guaranteed to reflect possible run-time behaviors.

One form of sound static analysis is *abstract interpretation* [36]. In abstract interpretation, atomic terms are interpreted as sound abstractions of their values, and operators are interpreted as functions that soundly manipulate these abstract values. The analysis is then performed by interpreting the program in a compositional manner using these abstracted values and operators. Alternatively, *dataflow analysis* [48] iteratively infers connections between variable definitions and uses, in terms of local transition rules that describe how various kinds of program constructs may impact variable values. Schmidt has explored the relationship between abstract interpretation and dataflow analysis [76]. More recently, more general forms of symbolic execution [33] have emerged as a means of understanding complex code. In symbolic execution, concrete values are used when available, and these are complemented by constraints that are inferred from terms for which only partial information is available. Reasoning about these constraints is then used to prune infeasible paths, and obtain more precise results. A number of works apply symbolic execution to operating systems code [29], [31].

While sound approaches are guaranteed to give correct results, they typically do not scale to the very diverse code bases that are prevalent in infrastructure software. An important insight of Engler et al. [41] was that valuable information could be obtained even when sacrificing soundness, and that sacrificing soundness could make it possible to treat software at the scales of the kernels of the Linux or BSD operating systems. Indeed, for certain types of problems, on certain code bases, that may mostly follow certain coding conventions, it may mostly be safe to *e.g.*, ignore the effects of aliases, assume that variable values are unchanged by calls to unanalyzed functions, etc. Real code has to be understood by developers and thus cannot be too complicated, so such simplifying assumptions are likely to hold in practice. Nevertheless, approaches that sacrifice soundness also require the user to manually validate the results. Still, it is likely to be much more efficient for the user to perform a potentially complex manual analysis in a specific case, rather than to implement all possible required analyses and apply them everywhere in the code base. A refinement of unsound analysis is the CEGAR approach [34], in which a highly approximate analysis is complemented by a sound analysis that checks the individual reports of the approximate analysis, and then any errors in reasoning detected by the sound analysis are used to refine the approximate analysis. The CEGAR approach has been applied effectively on device driver code in tools developed at Microsoft [21]. The environment in which the driver executes, however, is still represented by possibly unsound approximations.

Going further in the direction of sacrificing soundness for scalability, the software engineering community has recently explored a number of approaches to code understanding based on techniques developed in the areas of natural language understanding, data mining, and information retrieval. These approaches view code, as well as other software-reated artifacts, such as documentation and postings on mailing lists, as bags of words structured in various ways. Statistical methods are then used to collect words or phrases that seem to be highly correlated, independently of the semantics of the program constructs that connect them. The obliviousness to program semantics can lead to many false positives (invalid conclusions) [53], but can also highlight trends that are not apparent at the low level of individual program statements. We have previously explored combining such statistical methods with more traditional static analysis in identifying faults in the usage of constants in Linux kernel code [52].

### *3.1.2. Domain Specific Languages*

Writing low-level infrastructure code is tedious and difficult, and verifying it is even more so. To produce non-trivial programs, we could benefit from moving up the abstraction stack to enable both programming and proving as quickly as possible. Domain-specific languages (DSLs), also known as *little languages*, are a means to that end [5] [62].

#### *3.1.2.1. Traditional approach.*

Using little languages to aid in software development is a tried-and-trusted technique [79] by which programmers can express high-level ideas about the system at hand and avoid writing large quantities of formulaic C boilerplate.

This approach is typified by the Devil language for hardware access [7]. An OS programmer describes the register set of a hardware device in the high-level Devil language, which is then compiled into a library providing C functions to read and write values from the device registers. In doing so, Devil frees the programmer from having to write extensive bit-manipulation macros or inline functions to map between the values the OS code deals with, and the bit-representation used by the hardware: Devil generates code to do this automatically.

However, DSLs are not restricted to being "stub" compilers from declarative specifications. The Bossa language [6] is a prime example of a DSL involving imperative code (syntactically close to C) while offering a high-level of abstraction. This design of Bossa enables the developer to implement new process scheduling policies at a level of abstraction tailored to the application domain.

Conceptually, a DSL both abstracts away low-level details and justifies the abstraction by its semantics. In principle, it reduces development time by allowing the programmer to focus on high-level abstractions. The programmer needs to write less code, in a language with syntax and type checks adapted to the problem at hand, thus reducing the likelihood of errors.

#### *3.1.2.2. Embedding DSLs.*

The idea of a DSL has yet to realize its full potential in the OS community. Indeed, with the notable exception of interface definition languages for remote procedure call (RPC) stubs, most OS code is still written in a low-level language, such as C. Where DSL code generators are used in an OS, they tend to be extremely simple in both syntax and semantics. We conjecture that the effort to implement a given DSL usually outweighs its benefit. We identify several serious obstacles to using DSLs to build a modern OS: specifying what the generated code will look like, evolving the DSL over time, debugging generated code, implementing a bug-free code generator, and testing the DSL compiler.

Filet-o-Fish (FoF) [3] addresses these issues by providing a framework in which to build correct code generators from semantic specifications. This framework is presented as a Haskell library, enabling DSL writers to *embed* their languages within Haskell. DSL compilers built using FoF are quick to write, simple, and compact, but encode rigorous semantics for the generated code. They allow formal proofs of the run-time behavior of generated code, and automated testing of the code generator based on randomized inputs, providing greater test coverage than is usually feasible in a DSL. The use of FoF results in DSL compilers that OS developers can quickly implement and evolve, and that generate provably correct code. FoF has been used

to build a number of domain-specific languages used in Barrelfish, [22] an OS for heterogeneous multicore systems developed at ETH Zurich.

The development of an embedded DSL requires a few supporting abstractions in the host programming language. FoF was developed in the purely functional language Haskell, thus benefiting from the type class mechanism for overloading, a flexible parser offering convenient syntactic sugar, and purity enabling a more algebraic approach based on small, composable combinators. Object-oriented languages – such as Smalltalk [42] and its descendant Pharo [26] – or multi-paradigm languages – such as the Scala programming language [64] – also offer a wide range of mechanisms enabling the development of embedded DSLs. Perhaps suprisingly, a low-level imperative language – such as C – can also be extended so as to enable the development of embedded compilers [23].

*3.1.2.3. Certifying DSLs.*

Whilst automated and interactive software verification tools are progressively being applied to larger and larger programs, we have not yet reached the point where large-scale, legacy software – such as the Linux kernel – could formally be proved "correct". DSLs enable a pragmatic approach, by which one could realistically strengthen a large legacy software by first narrowing down its critical component(s) and then focus our verification efforts onto these components.

Dependently-typed languages, such as Coq or Idris, offer an ideal environment for embedding DSLs [32], [27] in a unified framework enabling verification. Dependent types support the type-safe embedding of object languages and Coq's mixfix notation system enables reasonably idiomatic domain-specific concrete syntax. Coq's powerful abstraction facilities provide a flexible framework in which to not only implement and verify a range of domain-specific compilers [3], but also to combine them, and reason about their combination.

Working with many DSLs optimizes the "horizontal" compositionality of systems, and favors reuse of building blocks, by contrast with the "vertical" composition of the traditional compiler pipeline, involving a stack of comparatively large intermediate languages that are harder to reuse the higher one goes. The idea of building compilers from reusable building blocks is a common one, of course. But the interface contracts of such blocks tend to be complex, so combinations are hard to get right. We believe that being able to write and verify formal specifications for the pieces will make it possible to know when components can be combined, and should help in designing good interfaces.

Furthermore, the fact that Coq is also a system for formalizing mathematics enables one to establish a close, formal connection between embedded DSLs and non-trivial domain-specific models. The possibility of developing software in a truly "model-driven" way is an exciting one. Following this methodology, we have implemented a certified compiler from regular expressions to x86 machine code [4]. Interestingly, our development crucially relied on an existing Coq formalization, due to Braibant and Pous, [28] of the theory of Kleene algebras.

While these individual experiments seem to converge toward embedding domain-specific languages in rich type theories, further experimental validation is required. Indeed, Barrelfish is an extremely small software compared to the Linux kernel. The challenge lies in scaling this methodology up to large software systems. Doing so calls for a unified platform enabling the development of a myriad of DSLs, supporting code reuse across DSLs as well as providing support for mechanically-verified proofs.

## 3.2. Research direction: Tools for improving legacy infrastructure software

A cornerstone of our work on legacy infrastructure software is the Coccinelle program matching and transformation tool for C code. Coccinelle has been in continuous development since 2005. Today, Coccinelle is extensively used in the context of Linux kernel development, as well as in the development of other software, such as wine, python, kvm, and systemd. Currently, Coccinelle is a mature software project, and no research is being conducted on Coccinelle itself. Instead, we leverage Coccinelle in other research projects [24], [25], [65], [68], [72], [74], [78][10], [20], both for code exploration, to better understand at a large scale problems in Linux development, and as an essential component in tools that require program matching and transformation. The continuing development and use of Coccinelle is also a source of visibility in the Linux kernel developer

community. We submitted the first patches to the Linux kernel based on Coccinelle in 2007. Since then, over 4500 patches have been accepted into the Linux kernel based on the use of Coccinelle, including around 3000 by over 500 developers from outside our research group.

Our recent work has focused on driver porting. Specifically, we have considered the problem of porting a Linux device driver across versions, particularly backporting, in which a modern driver needs to be used by a client who, typically for reasons of stability, is not able to update their Linux kernel to the most recent version. When multiple drivers need to be backported, they typically need many common changes, suggesting that Coccinelle could be applicable. Using Coccinelle, however, requires writing backporting transformation rules. In order to more fully automate the backporting (or symmetrically forward porting) process, these rules should be generated automatically. We have carried out a preliminary study in this direction with David Lo of Singapore Management University; this work, published at ICSME 2016 [17], is limited to a port from one version to the next one, in the case where the amount of change required is limited to a single line of code. Whisper has been awarded an ANR PRCI grant, to start in March 2017, to collaborate with the group of David Lo on scaling up the rule inference process and proposing a fully automatic porting solution.

## 3.3. Research direction: developing infrastructure software using Domain Specific Languages

We wish to pursue a *declarative* approach to developing infrastructure software. Indeed, there exists a significant gap between the high-level objectives of these systems and their implementation in low-level, imperative programming languages. To bridge that gap, we propose an approach based on domain-specific languages (DSLs). By abstracting away boilerplate code, DSLs increase the productivity of systems programmers. By providing a more declarative language, DSLs reduce the complexity of code, thus the likelihood of bugs.

Traditionally, systems are built by accretion of several, independent DSLs. For example, one might use Devil [7] to interact with devices, Bossa [6] to implement the scheduling policies. However, much effort is duplicated in implementing the back-ends of the individual DSLs. Our long term goal is to design a unified framework for developing and composing DSLs, following our work on Filet-o-Fish [3]. By providing a single conceptual framework, we hope to amortize the development cost of a myriad of DSLs through a principled approach to reusing and composing them.

Beyond the software engineering aspects, a unified platform brings us closer to the implementation of mechanically-verified DSLs. Dagand's recent work using the Coq proof assistant as an x86 macro-assembler [4] is a step in that direction, which belongs to a larger trend of hosting DSLs in dependent type theories [27], [63], [32]. A key benefit of those approaches is to provide – by construction – a formal, mechanized semantics to the DSLs thus developed. This semantics offers a foundation on which to base further verification efforts, whilst allowing interaction with non-verified code. We advocate a methodology based on incremental, piece-wise verification. Whilst building fully-certified systems from the top-down is a worthwhile endeavor [49], we wish to explore a bottom-up approach by which one focuses first and foremost on crucial subsystems and their associated properties.

Our current work on DSLs has two complementary goals: (i) the design of a unified framework for developing and composing DSLs, following our work on Filet-o-Fish, and (ii) the design of domain-specific languages for domains where there is a critical need for code correctness, and corresponding methodologies for proving properties of the run-time behavior of the system.

**ALPINES Project-Team**

# 3. Research Program

## 3.1. Overview

The research described here is directly relevant to several steps of the numerical simulation chain. Given a numerical simulation that was expressed as a set of differential equations, our research focuses on mesh generation methods for parallel computation, novel numerical algorithms for linear algebra, as well as algorithms and tools for their efficient and scalable implementation on high performance computers. The validation and the exploitation of the results is performed with collaborators from applications and is based on the usage of existing tools. In summary, the topics studied in our group are the following:

- Numerical methods and algorithms
  - Mesh generation for parallel computation
  - Solvers for numerical linear algebra
  - Computational kernels for numerical linear algebra
- Validation on numerical simulations

## 3.2. Domain specific language - parallel FreeFem++

In the engineering, researchers, and teachers communities, there is a strong demand for simulation frameworks that are simple to install and use, efficient, sustainable, and that solve efficiently and accurately complex problems for which there are no dedicated tools or codes available. In our group we develop FreeFem++ (see http://www.freefem.org/ff++), a user dedicated language for solving PDEs. The goal of FreeFem++ is not to be a substitute for complex numerical codes, but rather to provide an efficient and relatively generic tool for:

- getting a quick answer to a specific problem,
- prototyping the resolution of a new complex problem.

The current users of FreeFem++ are mathematicians, engineers, university professors, and students. In general for these users the installation of public libraries as MPI, MUMPS, Ipopt, Blas, lapack, OpenGL, fftw, scotch, is a very difficult problem. For this reason, the authors of FreeFem++ have created a user friendly language, and over years have enriched its capabilities and provided tools for compiling FreeFem++ such that the users do not need to have special knowledge of computer science. This leads to an important work on porting the software on different emerging architectures.

Today, the main components of parallel FreeFem++ are:

1. definition of a coarse grid,
2. splitting of the coarse grid,
3. mesh generation of all subdomains of the coarse grid, and construction of parallel datat structures for vectors and sparse matrices from the mesh of the subdomain,
4. call to a linear solver,
5. analysis of the result.

All these components are parallel, except for point (5) which is not in the focus of our research. However for the moment, the parallel mesh generation algorithm is very simple and not sufficient, for example it addresses only polygonal geometries. Having a better parallel mesh generation algorithm is one of the goals of our project. In addition, in the current version of FreeFem++, the parallelism is not hidden from the user, it is done through direct calls to MPI. Our goal is also to hide all the MPI calls in the specific language part of FreeFem++.

## 3.3. Solvers for numerical linear algebra

Iterative methods are widely used in industrial applications, and preconditioning is the most important research subject here. Our research considers domain decomposition methods and iterative methods and its goal is to develop solvers that are suitable for parallelism and that exploit the fact that the matrices are arising from the discretization of a system of PDEs on unstructured grids.

One of the main challenges that we address is the lack of robustness and scalability of existing methods as incomplete LU factorizations or Schwarz-based approaches, for which the number of iterations increases significantly with the problem size or with the number of processors. This is often due to the presence of several low frequency modes that hinder the convergence of the iterative method. To address this problem, we study direction preserving solvers in the context of multilevel domain decomposition methods with adaptive coarse spaces and multilevel incomplete decompositions. A judicious choice for the directions to be preserved through filtering or low rank approximations allows us to alleviate the effect of low frequency modes on the convergence.

We also focus on developing boundary integral equation methods that would be adapted to the simulation of wave propagation in complex physical situations, and that would lend themselves to the use of parallel architectures, which includes devising adapted domain decomposition approaches. The final objective is to bring the state of the art on boundary integral equations closer to contemporary industrial needs.

## 3.4. Computational kernels for numerical linear algebra

The design of new numerical methods that are robust and that have well proven convergence properties is one of the challenges addressed in Alpines. Another important challenge is the design of parallel algorithms for the novel numerical methods and the underlying building blocks from numerical linear algebra. The goal is to enable their efficient execution on a diverse set of node architectures and their scaling to emerging high-performance clusters with an increasing number of nodes.

Increased communication cost is one of the main challenges in high performance computing that we address in our research by investigating algorithms that minimize communication, as communication avoiding algorithms. We propose to integrate the minimization of communication into the algorithmic design of numerical linear algebra problems. This is different from previous approaches where the communication problem was addressed as a scheduling or as a tuning problem. The communication avoiding algorithmic design is an aproach originally developed in our group since 2007 (initially in collaboration with researchers from UC Berkeley and CU Denver). While at mid term we focus on reducing communication in numerical linear algebra, at long term we aim at considering the communication problem one level higher, during the parallel mesh generation tool described earlier.

**AVALON Project-Team**

# 3. Research Program

## 3.1. Energy Application Profiling and Modelization

International roadmaps schedule to build exascale systems by the 2018 time frame. According to the Top500 list published in November 2013, the most powerful supercomputer is the Tianhe-2 platform, a machine with more than 3,000,000 cores. It consumes more than 17 MW for a maximum performance of 33 PFlops while the Defense Advanced Research Projects Agency (DARPA) has set to 20 MW the maximum energy consumption of an exascale supercomputer [40].

Energy efficiency is therefore a major challenge for building next generation large scale platforms. The targeted platforms will gather hundreds of million cores, low power servers, or CPUs. Besides being very important, their power consumption will be dynamic and irregular.

Thus, to consume energy efficiently, we aim at investigating two research directions. First, we need to improve the measure, the understanding, and the analysis of the large-scale platform energy consumption. Unlike approaches [41] that mix the usage of internal and external wattmeters on a small set of resources, we target high frequency and precise internal and external energy measurements of each physical and virtual resources on large scale distributed systems.

Secondly, we need to find new mechanisms that consume less and better on such platforms. Combined with hardware optimizations, several works based on shutdown or slowdown approaches aim at reducing energy consumption of distributed platforms and applications. To consume less, we first plan to explore the provision of accurate estimation of the energy consumed by applications without pre-executing and knowing them while most of the works try to do it based on in-depth application knowledge (code instrumentation [44], phase detection for specific HPC applications [49], etc.). As a second step, we aim at designing a framework model that allows interactions, dialogues and decisions taken in cooperation between the user/application, the administrator, the resource manager, and the energy supplier. While smart grid is one of the last killer scenarios for networks, electrical provisioning of next generation large IT infrastructures remains a challenge.

## 3.2. Data-intensive Application Profiling, Modeling, and Management

Recently, the term “Big Data” has emerged to design data sets or collections so large that they become intractable for classical tools. This term is most of the time implicitly linked to “analytics” to refer to issues such as curation, storage, search, sharing, analysis, and visualization. However, the Big Data challenge is not limited to data-analytics, a field that is well covered by programming languages and run-time systems such as Map-Reduce. It also encompasses data-intensive applications. These applications can be sorted into two categories. In High Performance Computing (HPC), data-intensive applications leverage post-petascale infrastructures to perform highly parallel computations on large amount of data, while in High Throughput Computing (HTC), a large amount of independent and sequential computations are performed on huge data collections.

These two types of data-intensive applications (HTC and HPC) raise challenges related to profiling and modeling that the Avalon team proposes to address. While the characteristics of data-intensive applications are very different, our work will remain coherent and focused. Indeed, a common goal will be to acquire a better understanding of both the applications and the underlying infrastructures running them to propose the best match between application requirements and infrastructure capacities. To achieve this objective, we will extensively rely on logging and profiling in order to design sound, accurate, and validated models. Then, the proposed models will be integrated and consolidated within a single simulation framework (SIMGRID). This will allow us to explore various potential “what-if?” scenarios and offer objective indicators to select interesting infrastructure configurations that match application specificities.

Another challenge is the ability to mix several heterogeneous infrastructures that scientists have at their disposal (*e.g.,* Grids, Clouds, and Desktop Grids) to execute data-intensive applications. Leveraging the aforementioned results, we will design strategies for efficient data management service for hybrid computing infrastructures.

## 3.3. Resourc-Agnostic Application Description Model

When programming in parallel, users expect to obtain performance improvement, whatever the cost is. For long, parallel machines have been simple enough to let a user program them given a minimal abstraction of their hardware. For example, MPI [43] exposes the number of nodes but hides the complexity of network topology behind a set of collective operations; OpenMP [47] simplifies the management of threads on top of a shared memory machine while OpenACC [46] aims at simplifying the use of GPGPU.

However, machines and applications are getting more and more complex so that the cost of manually handling an application is becoming very high [42]. Hardware complexity also stems from the unclear path towards next generations of hardware coming from the frequency wall: multi-core CPU, many-core CPU, GPGPUs, deep memory hierarchy, etc. have a strong impact on parallel algorithms. Hence, even though an abstract enough parallel language (UPC, Fortress, X10, etc.) succeeds, it will still face the challenge of supporting distinct codes corresponding to different algorithms corresponding to distinct hardware capacities.

Therefore, the challenge we aim to address is to define a model, for describing the structure of parallel and distributed applications that enables code variations but also efficient executions on parallel and distributed infrastructures. Indeed, this issue appears for HPC applications but also for cloud oriented applications. The challenge is to adapt an application to user constraints such as performance, energy, security, etc.

Our approach is to consider component based models [50] as they offer the ability to manipulate the software architecture of an application. To achieve our goal, we consider a "compilation" approach that transforms a resource-agnostic application description into a resource-specific description. The challenge is thus to determine a component based model that enables to efficiently compute application mapping while being tractable. In particular, it has to provide an efficient support with respect to application and resource elasticity, energy consumption and data management. OpenMP runtime is a specific use case that we target.

## 3.4. Application Mapping and Scheduling

This research axis is at the crossroad of the Avalon team. In particular, it gathers results of the three other research axis. We plan to consider application mapping and scheduling through the following three issues.

### *3.4.1. Application Mapping and Software Deployment*

Application mapping and software deployment consist in the process of assigning distributed pieces of software to a set of resources. Resources can be selected according to different criteria such as performance, cost, energy consumption, security management, etc. A first issue is to select resources at application launch time. With the wide adoption of elastic platforms, *i.e.,* platforms that let the number of resources allocated to an application to be increased or decreased during its execution, the issue is also to handle resource selection at runtime.

The challenge in this context corresponds to the mapping of applications onto distributed resources. It will consist in designing algorithms that in particular take into consideration application profiling, modeling, and description.

A particular facet of this challenge is to propose scheduling algorithms for dynamic and elastic platforms. As the amount of elements can vary, some kind of control of the platforms must be used accordingly to the scheduling.

### 3.4.2. Non-Deterministic Workflow Scheduling

Many scientific applications are described through workflow structures. Due to the increasing level of parallelism offered by modern computing infrastructures, workflow applications now have to be composed not only of sequential programs, but also of parallel ones. New applications are now built upon workflows with conditionals and loops (also called non-deterministic workflows).

These workflows can not be scheduled beforehand. Moreover cloud platforms bring on-demand resource provisioning and pay-as-you-go billing models. Therefore, there is a problem of resource allocation for non-deterministic workflows under budget constraints and using such an elastic management of resources.

Another important issue is data management. We need to schedule the data movements and replications while taking job scheduling into account. If possible, data management and job scheduling should be done at the same time in a closely coupled interaction.

### 3.4.3. Security Management in Cloud Infrastructure

Security has been proven to be sometimes difficult to obtain [48] and several issues have been raised in Clouds. Nowadays virtualization is used as the sole mechanism to secure different users sharing resources on Clouds. But, due to improper virtualization of all the components of Clouds (such as micro-architectural components), data leak and modification can occur. Accordingly, next-generation protection mechanisms are required to enforce security on Clouds and provide a way to cope with the current limitation of virtualization mechanisms.

As we are dealing with parallel and distributed applications, security mechanisms must be able to cope with multiple machines. Our approach is to combine a set of existing and novel security mechanisms that are spread in the different layers and components of Clouds in order to provide an in-depth and end-to-end security on Clouds. To do it, our first challenge is to define a generic model to express security policies.

Our second challenge is to work on security-aware resource allocation algorithms. The goal of such algorithms is to find a good trade-off between security and unshared resources. Consequently, they can limit resources sharing to increase security. It leads to complex trade-off between infrastructure consolidation, performance, and security.

**DATAMOVE Team**

# 3. Research Program

## 3.1. Motivation

Today's largest supercomputers [0] are composed of few millions of cores, with performances almost reaching 100 PetaFlops [0] for the largest machine. Moving data in such large supercomputers is becoming a major performance bottleneck, and the situation is expected to worsen even more at exascale and beyond. The data transfer capabilities are growing at a slower rate than processing power ones. The profusion of available flops will very likely be underused due to constrained communication capabilities. It is commonly admitted that data movements account for 50% to 70% of the global power consumption [0]. Thus, data movements are potentially one of the most important source of savings for enabling supercomputers to stay in the commonly adopted energy barrier of 20 MegaWatts. In the mid to long term, non volatile memory (NVRAM) is expected to deeply change the machine I/Os. Data distribution will shift from disk arrays with an access time often considered as uniform, towards permanent storage capabilities at each node of the machine, making data locality an even more prevalent paradigm.

The DataMove team works on **optimizing data movements for large scale computing** mainly at two related levels:

- Resource allocation
- Integration of numerical simulation and data analysis

The resource and job management system (also called batch scheduler or RJMS) is in charge of allocating resources upon user requests for executing their parallel applications. The growing cost of data movements requires adapted scheduling policies able to take into account the influence of intra-application communications, I/Os as well as contention caused by data traffic generated by other concurrent applications. Modelling the application behavior to anticipate its actual resource usage on such architecture is known to be challenging, but it becomes critical for improving performances (execution time, energy, or any other relevant objective). The job management system also needs to handle new types of workloads: high performance platforms now need to execute more and more often data intensive processing tasks like data analysis in addition to traditional computation intensive numerical simulations. In particular, the ever growing amount of data generated by numerical simulation calls for a tighter integration between the simulation and the data analysis. The challenge here is to reduce data traffic and to speed-up result analysis by performing result processing (compression, indexation, analysis, visualization, etc.) as closely as possible to the locus and time of data generation. This emerging trend called *in situ analytics* requires to revisit the traditional workflow (loop of batch processing followed by postmortem analysis). The application becomes a whole including the simulation, in situ processing and I/Os. This motivates the development of new well-adapted resource sharing strategies, data structures and parallel analytics schemes to efficiently interleave the different components of the application and globally improve the performance.

## 3.2. Strategy

DataMove targets HPC (High Performance Computing) at Exascale. But such machines and the associated applications are expected to be available only in 5 to 10 years. Meanwhile, we expect to see a growing number of petaflop machines to answer the needs for advanced numerical simulations. A sustainable exploitation of these petaflop machines is a real and hard challenge that we address. We may also see in the coming years a convergence between HPC and Big Data, HPC platforms becoming more elastic and supporting Big Data

[0] Top500 Ranking, http://www.top500.org
[0] $10^{15}$ floating point operations per second
[0] SciDAC Review, 2010, http://www.scidacreview.org/1001/pdf/hardware.pdf

jobs, or HPC applications being more commonly executed on cloud like architectures. This is the second top objective of the 2015 US Strategic Computing Initiative [0]: *Increasing coherence between the technology base used for modelling and simulation and that used for data analytic computing*. We contribute to that convergence at our level, considering more dynamic and versatile target platforms and types of workloads.

Our approaches should entail minimal modifications on the code of numerical simulations. Often large scale numerical simulations are complex domain specific codes with a long life span. We assume these codes as being sufficiently optimized. We influence the behavior of numerical simulations through resource allocation at the job management system level or when interleaving them with analytics code.

To tackle these issues, we propose to intertwine theoretical research and practical developments in an agile mode. Algorithms with performance guarantees are designed and experimented on large scale platforms with realistic usage scenarios developed with partner scientists or based on logs of the biggest available computing platforms (national supercomputers like Curie, or the BlueWaters machine accessible through our collaboration with Argonne National Lab). Conversely, a strong experimental expertise enables to feed theoretical models with sound hypotheses, to twist proven algorithms with practical heuristics that could be further retro-feeded into adequate theoretical models.

A central scientific question is to make the relevant choices for optimizing performance (in a broad sense) in a reasonable time. HPC architectures and applications are increasingly complex systems (heterogeneity, dynamicity, uncertainties), which leads to consider the **optimization of resource allocation based on multiple objectives**, often contradictory (like energy and run-time for instance). Focusing on the optimization of one particular objective usually leads to worsen the others. The historical positioning of some members of the team who are specialists in multi-objective optimization is to generate a (limited) set of trade-off configurations, called *Pareto points*, and choose when required the most suitable trade-off between all the objectives. This methodology differs from the classical approaches, which simplify the problem into a single objective one (focus on a particular objective, combining the various objectives or agglomerate them). The real challenge is thus to combine algorithmic techniques to account for this diversity while guaranteeing a target efficiency for all the various objectives.

The DataMove team aims to elaborate generic and effective solutions of practical interest. We make our new algorithms accessible through the team flagship software tools, **the OAR batch scheduler and the in situ processing framework FlowVR**. We maintain and enforce strong links with teams closely connected with large architecture design and operation, as well as scientists of other disciplines, in particular computational biologists, with whom we elaborate and validate new usage scenarios.

## 3.3. Research Directions

DataMove targets HPC (High Performance Computing) at Exascale. But such machines and the associated applications are expected to be available only in 5 to 10 years. Meanwhile, we expect to see a growing number of petaflop machines to answer the needs for advanced numerical simulations. A sustainable exploitation of these petaflop machines is a real and hard challenge that we address. We may also see in the coming years a convergence between HPC and Big Data, HPC platforms becoming more elastic and supporting Big Data jobs, or HPC applications being more commonly executed on cloud like architectures. This is the second top objective of the 2015 US Strategic Computing Initiative [0]: *Increasing coherence between the technology base used for modelling and simulation and that used for data analytic computing*. We contribute to that convergence at our level, considering more dynamic and versatile target platforms and types of workloads.

Our approaches should entail minimal modifications on the code of numerical simulations. Often large scale numerical simulations are complex domain specific codes with a long life span. We assume these codes as being sufficiently optimized. We influence the behavior of numerical simulations through resource allocation at the job management system level or when interleaving them with analytics code.

[0] https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

[0] https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

To tackle these issues, we propose to intertwine theoretical research and practical developments in an agile mode. Algorithms with performance guarantees are designed and experimented on large scale platforms with realistic usage scenarios developed with partner scientists or based on logs of the biggest available computing platforms (national supercomputers like Curie, or the BlueWaters machine accessible through our collaboration with Argonne National Lab). Conversely, a strong experimental expertise enables to feed theoretical models with sound hypotheses, to twist proven algorithms with practical heuristics that could be further retro-feeded into adequate theoretical models.

A central scientific question is to make the relevant choices for optimizing performance (in a broad sense) in a reasonable time. HPC architectures and applications are increasingly complex systems (heterogeneity, dynamicity, uncertainties), which leads to consider the **optimization of resource allocation based on multiple objectives**, often contradictory (like energy and run-time for instance). Focusing on the optimization of one particular objective usually leads to worsen the others. The historical positioning of some members of the team who are specialists in multi-objective optimization is to generate a (limited) set of trade-off configurations, called *Pareto points*, and choose when required the most suitable trade-off between all the objectives. This methodology differs from the classical approaches, which simplify the problem into a single objective one (focus on a particular objective, combining the various objectives or agglomerate them). The real challenge is thus to combine algorithmic techniques to account for this diversity while guaranteeing a target efficiency for all the various objectives.

The DataMove team aims to elaborate generic and effective solutions of practical interest. We make our new algorithms accessible through the team flagship software tools, **the OAR batch scheduler and the in situ processing framework FlowVR**. We maintain and enforce strong links with teams closely connected with large architecture design and operation, as well as scientists of other disciplines, in particular computational biologists, with whom we elaborate and validate new usage scenarios.

**HIEPACS Project-Team**

# 3. Research Program

## 3.1. Introduction

The methodological component of HIEPACS concerns the expertise for the design as well as the efficient and scalable implementation of highly parallel numerical algorithms to perform frontier simulations. In order to address these computational challenges a hierarchical organization of the research is considered. In this bottom-up approach, we first consider in Section 3.2 generic topics concerning high performance computational science. The activities described in this section are transversal to the overall project and their outcome will support all the other research activities at various levels in order to ensure the parallel scalability of the algorithms. The aim of this activity is not to study general purpose solution but rather to address these problems in close relation with specialists of the field in order to adapt and tune advanced approaches in our algorithmic designs. The next activity, described in Section 3.3 , is related to the study of parallel linear algebra techniques that currently appear as promising approaches to tackle huge problems on extreme scale platforms. We highlight the linear problems (linear systems or eigenproblems) because they are in many large scale applications the main computational intensive numerical kernels and often the main performance bottleneck. These parallel numerical techniques, which are involved in the IPL C2S@EXA, will be the basis of both academic and industrial collaborations, some are described in Section 4.1 , but will also be closely related to some functionalities developed in the parallel fast multipole activity described in Section 3.4 . Finally, as the accuracy of the physical models increases, there is a real need to go for parallel efficient algorithm implementation for multiphysics and multiscale modeling in particular in the context of code coupling. The challenges associated with this activity will be addressed in the framework of the activity described in Section 3.5 .

Currently, we have one major application (see Section 4.1 ) that is in material physics. We will collaborate to all steps of the design of the parallel simulation tool. More precisely, our applied mathematics skill will contribute to the modelling, our advanced numerical schemes will help in the design and efficient software implementation for very large parallel simulations. We also participate to a few co-design actions in close collaboration with some applicative groups, some of them being involved in the IPL C2S@EXA. The objective of this activity is to instantiate our expertise in fields where they are critical for designing scalable simulation tools. We refer to Section 4.2 for a detailed description of these activities.

## 3.2. High-performance computing on next generation architectures

**Participants:** Emmanuel Agullo, Olivier Coulaud, Mathieu Faverge, Luc Giraud, Abdou Guermouche, Matias Hastaran, Guillaume Latu, Grégoire Pichon, Florent Pruvost, Pierre Ramet, Jean Roman, Emrullah Fatih Yetkin.

The research directions proposed in HIEPACS are strongly influenced by both the applications we are studying and the architectures that we target (i.e., massively parallel heterogeneous many-core architectures, ...). Our main goal is to study the methodology needed to efficiently exploit the new generation of high-performance computers with all the constraints that it induces. To achieve this high-performance with complex applications we have to study both algorithmic problems and the impact of the architectures on the algorithm design.

From the application point of view, the project will be interested in multiresolution, multiscale and hierarchical approaches which lead to multi-level parallelism schemes. This hierarchical parallelism approach is necessary to achieve good performance and high-scalability on modern massively parallel platforms. In this context, more specific algorithmic problems are very important to obtain high performance. Indeed, the kind of applications we are interested in are often based on data redistribution for example (e.g., code coupling applications). This well-known issue becomes very challenging with the increase of both the number of computational nodes and the amount of data. Thus, we have both to study new algorithms and to adapt the

existing ones. In addition, some issues like task scheduling have to be restudied in this new context. It is important to note that the work developed in this area will be applied for example in the context of code coupling (see Section 3.5 ).

Considering the complexity of modern architectures like massively parallel architectures or new generation heterogeneous multicore architectures, task scheduling becomes a challenging problem which is central to obtain a high efficiency. Of course, this work requires the use/design of scheduling algorithms and models specifically to tackle our target problems. This has to be done in collaboration with our colleagues from the scheduling community like for example O. Beaumont (Inria REALOPT Project-Team). It is important to note that this topic is strongly linked to the underlying programming model. Indeed, considering multicore architectures, it has appeared, in the last five years, that the best programming model is an approach mixing multi-threading within computational nodes and message passing between them. In the last five years, a lot of work has been developed in the high-performance computing community to understand what is critic to efficiently exploit massively multicore platforms that will appear in the near future. It appeared that the key for the performance is firstly the granularity of the computations. Indeed, in such platforms the granularity of the parallelism must be small so that we can feed all the computing units with a sufficient amount of work. It is thus very crucial for us to design new high performance tools for scientific computing in this new context. This will be developed in the context of our solvers, for example, to adapt to this new parallel scheme. Secondly, the larger the number of cores inside a node, the more complex the memory hierarchy. This remark impacts the behaviour of the algorithms within the node. Indeed, on this kind of platforms, NUMA effects will be more and more problematic. Thus, it is very important to study and design data-aware algorithms which take into account the affinity between computational threads and the data they access. This is particularly important in the context of our high-performance tools. Note that this work has to be based on an intelligent cooperative underlying run-time (like the tools developed by the Inria STORM Project-Team) which allows a fine management of data distribution within a node.

Another very important issue concerns high-performance computing using “heterogeneous” resources within a computational node. Indeed, with the deployment of the GPU and the use of more specific co-processors, it is important for our algorithms to efficiently exploit these new type of architectures. To adapt our algorithms and tools to these accelerators, we need to identify what can be done on the GPU for example and what cannot. Note that recent results in the field have shown the interest of using both regular cores and GPU to perform computations. Note also that in opposition to the case of the parallelism granularity needed by regular multicore architectures, GPU requires coarser grain parallelism. Thus, making both GPU and regular cores work all together will lead to two types of tasks in terms of granularity. This represents a challenging problem especially in terms of scheduling. From this perspective, we investigate new approaches for composing parallel applications within a runtime system for heterogeneous platforms.

In that framework, the SOLHAR project aims at studying and designing algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computers equipped with accelerators. Several attempts have been made to accomplish the porting of these methods on such architectures; the proposed approaches are mostly based on a simple offloading of some computational tasks (the coarsest grained ones) to the accelerators and rely on fine hand-tuning of the code and accurate performance modeling to achieve efficiency. SOLHAR proposes an innovative approach which relies on the efficiency and portability of runtime systems, such as the StarPU tool developed in the STORM team. Although the SOLHAR project will focus on heterogeneous computers equipped with GPUs due to their wide availability and affordable cost, the research accomplished on algorithms, methods and programming models will be readily applicable to other accelerator devices. Our final goal would be to have high performance solvers and tools which can efficiently run on all these types of complex architectures by exploiting all the resources of the platform (even if they are heterogeneous).

In order to achieve an advanced knowledge concerning the design of efficient computational kernels to be used on our high performance algorithms and codes, we will develop research activities first on regular frameworks before extending them to more irregular and complex situations. In particular, we will work first on optimized dense linear algebra kernels and we will use them in our more complicated direct and hybrid

solvers for sparse linear algebra and in our fast multipole algorithms for interaction computations. In this context, we will participate to the development of those kernels in collaboration with groups specialized in dense linear algebra. In particular, we intend develop a strong collaboration with the group of Jack Dongarra at the University of Tennessee and collaborating research groups. The objectives will be to develop dense linear algebra algorithms and libraries for multicore architectures in the context the PLASMA project and for GPU and hybrid multicore/GPU architectures in the context of the MAGMA project. The framework that hosts all these research activities is the associate team MORSE. A new solver has emerged from the associate team, Chameleon. While PLASMA and MAGMA focus on multicore and GPU architectures, respectively, Chameleon makes the most out of heterogeneous architectures thanks to task-based dynamic runtime systems.

A more prospective objective is to study the resiliency in the context of large-scale scientific applications for massively parallel architectures. Indeed, with the increase of the number of computational cores per node, the probability of a hardware crash on a core or of a memory corruption is dramatically increased. This represents a crucial problem that needs to be addressed. However, we will only study it at the algorithmic/application level even if it needed lower-level mechanisms (at OS level or even hardware level). Of course, this work can be performed at lower levels (at operating system) level for example but we do believe that handling faults at the application level provides more knowledge about what has to be done (at application level we know what is critical and what is not). The approach that we will follow will be based on the use of a combination of fault-tolerant implementations of the run-time environments we use (like for example ULFM) and an adaptation of our algorithms to try to manage this kind of faults. This topic represents a very long range objective which needs to be addressed to guaranty the robustness of our solvers and applications. In that respect, we are involved in the EXA2CT FP7 project.

Finally, it is important to note that the main goal of HIEPACS is to design tools and algorithms that will be used within complex simulation frameworks on next-generation parallel machines. Thus, we intend with our partners to use the proposed approach in complex scientific codes and to validate them within very large scale simulations as well as designing parallel solution in co-design collaborations.

## 3.3. High performance solvers for large linear algebra problems

**Participants:** Emmanuel Agullo, Olivier Coulaud, Mathieu Faverge, Aurélien Falco, Luc Giraud, Abdou Guermouche, Yuval Harness, Matias Hastaran, Matthieu Kuhn, Gilles Marait, Julien Pedron, Cyrille Piacibello, Grégoire Pichon, Louis Poirel, Pierre Ramet, Jean Roman.

Starting with the developments of basic linear algebra kernels tuned for various classes of computers, a significant knowledge on the basic concepts for implementations on high-performance scientific computers has been accumulated. Further knowledge has been acquired through the design of more sophisticated linear algebra algorithms fully exploiting those basic intensive computational kernels. In that context, we still look at the development of new computing platforms and their associated programming tools. This enables us to identify the possible bottlenecks of new computer architectures (memory path, various level of caches, inter processor or node network) and to propose ways to overcome them in algorithmic design. With the goal of designing efficient scalable linear algebra solvers for large scale applications, various tracks will be followed in order to investigate different complementary approaches. Sparse direct solvers have been for years the methods of choice for solving linear systems of equations, it is nowadays admitted that classical approaches are not scalable neither from a computational complexity nor from a memory view point for large problems such as those arising from the discretization of large 3D PDE problems. We will continue to work on sparse direct solvers on the one hand to make sure they fully benefit from most advanced computing platforms and on the other hand to attempt to reduce their memory and computational costs for some classes of problems where data sparse ideas can be considered. Furthermore, sparse direct solvers are a key building boxes for the design of some of our parallel algorithms such as the hybrid solvers described in the sequel of this section. Our activities in that context will mainly address preconditioned Krylov subspace methods; both components, preconditioner and Krylov solvers, will be investigated. In this framework, and possibly in relation with the research activity on fast multipole, we intend to study how emerging $\mathcal{H}$-matrix arithmetic can benefit to our solver research efforts.

### 3.3.1. Parallel sparse direct solver

For the solution of large sparse linear systems, we design numerical schemes and software packages for direct and hybrid parallel solvers. Sparse direct solvers are mandatory when the linear system is very ill-conditioned; such a situation is often encountered in structural mechanics codes, for example. Therefore, to obtain an industrial software tool that must be robust and versatile, high-performance sparse direct solvers are mandatory, and parallelism is then necessary for reasons of memory capability and acceptable solution time. Moreover, in order to solve efficiently 3D problems with more than 50 million unknowns, which is now a reachable challenge with new multicore supercomputers, we must achieve good scalability in time and control memory overhead. Solving a sparse linear system by a direct method is generally a highly irregular problem that induces some challenging algorithmic problems and requires a sophisticated implementation scheme in order to fully exploit the capabilities of modern supercomputers.

New supercomputers incorporate many microprocessors which are composed of one or many computational cores. These new architectures induce strongly hierarchical topologies. These are called NUMA architectures. In the context of distributed NUMA architectures, in collaboration with the Inria STORM team, we study optimization strategies to improve the scheduling of communications, threads and I/O. We have developed dynamic scheduling designed for NUMA architectures in the PaStiX solver. The data structures of the solver, as well as the patterns of communication have been modified to meet the needs of these architectures and dynamic scheduling. We are also interested in the dynamic adaptation of the computation grain to use efficiently multi-core architectures and shared memory. Experiments on several numerical test cases have been performed to prove the efficiency of the approach on different architectures. Sparse direct solvers such as PaStiX are currently limited by their memory requirements and computational cost. They are competitive for small matrices but are often less efficient than iterative methods for large matrices in terms of memory. We are currently accelerating the dense algebra components of direct solvers using hierarchical matrices algebra.

In collaboration with the ICL team from the University of Tennessee, and the STORM team from Inria, we are evaluating the way to replace the embedded scheduling driver of the PaStiX solver by one of the generic frameworks, PaRSEC or StarPU, to execute the task graph corresponding to a sparse factorization. The aim is to design algorithms and parallel programming models for implementing direct methods for the solution of sparse linear systems on emerging computer equipped with GPU accelerators. More generally, this work will be performed in the context of the associate team MORSE and the ANR SOLHAR project which aims at designing high performance sparse direct solvers for modern heterogeneous systems. This ANR project involves several groups working either on the sparse linear solver aspects (HIEPACS and ROMA from Inria and APO from IRIT), on runtime systems (STORM from Inria) or scheduling algorithms (REALOPT and ROMA from Inria). The results of these efforts will be validated in the applications provided by the industrial project members, namely CEA-CESTA and Airbus Group Innovations.

### 3.3.2. Hybrid direct/iterative solvers based on algebraic domain decomposition techniques

One route to the parallel scalable solution of large sparse linear systems in parallel scientific computing is the use of hybrid methods that hierarchically combine direct and iterative methods. These techniques inherit the advantages of each approach, namely the limited amount of memory and natural parallelization for the iterative component and the numerical robustness of the direct part. The general underlying ideas are not new since they have been intensively used to design domain decomposition techniques; those approaches cover a fairly large range of computing techniques for the numerical solution of partial differential equations (PDEs) in time and space. Generally speaking, it refers to the splitting of the computational domain into sub-domains with or without overlap. The splitting strategy is generally governed by various constraints/objectives but the main one is to express parallelism. The numerical properties of the PDEs to be solved are usually intensively exploited at the continuous or discrete levels to design the numerical algorithms so that the resulting specialized technique will only work for the class of linear systems associated with the targeted PDE.

In that context, we intend to continue our effort on the design of algebraic non-overlapping domain decomposition techniques that rely on the solution of a Schur complement system defined on the interface introduced by the partitioning of the adjacency graph of the sparse matrix associated with the linear system. Although it

is better conditioned than the original system the Schur complement needs to be precondition to be amenable to a solution using a Krylov subspace method. Different hierarchical preconditioners will be considered, possibly multilevel, to improve the numerical behaviour of the current approaches implemented in our software libraries HIPS and MaPHyS. This activity will be developed in the context of the ANR DEDALES project. In addition to this numerical studies, advanced parallel implementation will be developed that will involve close collaborations between the hybrid and sparse direct activities.

### 3.3.3. Linear Krylov solvers

Preconditioning is the main focus of the two activities described above. They aim at speeding up the convergence of a Krylov subspace method that is the complementary component involved in the solvers of interest for us. In that framework, we believe that various aspects deserve to be investigated; we will consider the following ones:

- preconditioned block Krylov solvers for multiple right-hand sides. In many large scientific and industrial applications, one has to solve a sequence of linear systems with several right-hand sides given simultaneously or in sequence (radar cross section calculation in electromagnetism, various source locations in seismic, parametric studies in general, ...). For “simultaneous" right-hand sides, the solvers of choice have been for years based on matrix factorizations as the factorization is performed once and simple and cheap block forward/backward substitutions are then performed. In order to effectively propose alternative to such solvers, we need to have efficient preconditioned Krylov subspace solvers. In that framework, block Krylov approaches, where the Krylov spaces associated with each right-hand side are shared to enlarge the search space will be considered. They are not only attractive because of this numerical feature (larger search space), but also from an implementation point of view. Their block-structures exhibit nice features with respect to data locality and re-usability that comply with the memory constraint of multicore architectures. We will continue the numerical study and design of the block GMRES variant that combines inexact breakdown detection, deflation at restart and subspace recycling. Beyond new numerical investigations, a software implementation to be included in our linear solver libray will be developed in the context of the DGA HIBOX project.
- Extension or modification of Krylov subspace algorithms for multicore architectures: finally to match as much as possible to the computer architecture evolution and get as much as possible performance out of the computer, a particular attention will be paid to adapt, extend or develop numerical schemes that comply with the efficiency constraints associated with the available computers. Nowadays, multicore architectures seem to become widely used, where memory latency and bandwidth are the main bottlenecks; investigations on communication avoiding techniques will be undertaken in the framework of preconditioned Krylov subspace solvers as a general guideline for all the items mentioned above.

### 3.3.4. Eigensolvers

Many eigensolvers also rely on Krylov subspace techniques. Naturally some links exist between the Krylov subspace linear solvers and the Krylov subspace eigensolvers. We plan to study the computation of eigenvalue problems with respect to the following two different axes:

- Exploiting the link between Krylov subspace methods for linear system solution and eigensolvers, we intend to develop advanced iterative linear methods based on Krylov subspace methods that use some spectral information to build part of a subspace to be recycled, either though space augmentation or through preconditioner update. This spectral information may correspond to a certain part of the spectrum of the original large matrix or to some approximations of the eigenvalues obtained by solving a reduced eigenproblem. This technique will also be investigated in the framework of block Krylov subspace methods.
- In the context of the calculation of the ground state of an atomistic system, eigenvalue computation is a critical step; more accurate and more efficient parallel and scalable eigensolvers are required.

## 3.4. High performance Fast Multipole Method for N-body problems

**Participants:** Emmanuel Agullo, Olivier Coulaud, Quentin Khan, Cyrille Piacibello, Guillaume Sylvand.

In most scientific computing applications considered nowadays as computational challenges (like biological and material systems, astrophysics or electromagnetism), the introduction of hierarchical methods based on an octree structure has dramatically reduced the amount of computation needed to simulate those systems for a given accuracy. For instance, in the N-body problem arising from these application fields, we must compute all pairwise interactions among N objects (particles, lines, ...) at every timestep. Among these methods, the Fast Multipole Method (FMM) developed for gravitational potentials in astrophysics and for electrostatic (coulombic) potentials in molecular simulations solves this N-body problem for any given precision with $O(N)$ runtime complexity against $O(N^2)$ for the direct computation.

The potential field is decomposed in a near field part, directly computed, and a far field part approximated thanks to multipole and local expansions. We introduced a matrix formulation of the FMM that exploits the cache hierarchy on a processor through the Basic Linear Algebra Subprograms (BLAS). Moreover, we developed a parallel adaptive version of the FMM algorithm for heterogeneous particle distributions, which is very efficient on parallel clusters of SMP nodes. Finally on such computers, we developed the first hybrid MPI-thread algorithm, which enables to reach better parallel efficiency and better memory scalability. We plan to work on the following points in HIEPACS.

### *3.4.1. Improvement of calculation efficiency*

Nowadays, the high performance computing community is examining alternative architectures that address the limitations of modern cache-based designs. GPU (Graphics Processing Units) and the Cell processor have thus already been used in astrophysics and in molecular dynamics. The Fast Mutipole Method has also been implemented on GPU. We intend to examine the potential of using these forthcoming processors as a building block for high-end parallel computing in N-body calculations. More precisely, we want to take advantage of our specific underlying BLAS routines to obtain an efficient and easily portable FMM for these new architectures. Algorithmic issues such as dynamic load balancing among heterogeneous cores will also have to be solved in order to gather all the available computation power. This research action will be conduced on close connection with the activity described in Section 3.2 .

### *3.4.2. Non uniform distributions*

In many applications arising from material physics or astrophysics, the distribution of the data is highly non uniform and the data can grow between two time steps. As mentioned previously, we have proposed a hybrid MPI-thread algorithm to exploit the data locality within each node. We plan to further improve the load balancing for highly non uniform particle distributions with small computation grain thanks to dynamic load balancing at the thread level and thanks to a load balancing correction over several simulation time steps at the process level.

### *3.4.3. Fast multipole method for dislocation operators*

The engine that we develop will be extended to new potentials arising from material physics such as those used in dislocation simulations. The interaction between dislocations is long ranged ($O(1/r)$) and anisotropic, leading to severe computational challenges for large-scale simulations. Several approaches based on the FMM or based on spatial decomposition in boxes are proposed to speed-up the computation. In dislocation codes, the calculation of the interaction forces between dislocations is still the most CPU time consuming. This computation has to be improved to obtain faster and more accurate simulations. Moreover, in such simulations, the number of dislocations grows while the phenomenon occurs and these dislocations are not uniformly distributed in the domain. This means that strategies to dynamically balance the computational load are crucial to achieve high performance.

### 3.4.4. Fast multipole method for boundary element methods

The boundary element method (BEM) is a well known solution of boundary value problems appearing in various fields of physics. With this approach, we only have to solve an integral equation on the boundary. This implies an interaction that decreases in space, but results in the solution of a dense linear system with $O(N^3)$ complexity. The FMM calculation that performs the matrix-vector product enables the use of Krylov subspace methods. Based on the parallel data distribution of the underlying octree implemented to perform the FMM, parallel preconditioners can be designed that exploit the local interaction matrices computed at the finest level of the octree. This research action will be conduced on close connection with the activity described in Section 3.3 . Following our earlier experience, we plan to first consider approximate inverse preconditionners that can efficiently exploit these data structures.

## 3.5. Load balancing algorithms for complex simulations

**Participants:** Astrid Casadei, Olivier Coulaud, Aurélien Esnard, Maria Predari, Pierre Ramet, Jean Roman.

Many important physical phenomena in material physics and climatology are inherently complex applications. They often use multi-physics or multi-scale approaches, which couple different models and codes. The key idea is to reuse available legacy codes through a coupling framework instead of merging them into a stand-alone application. There is typically one model per different scale or physics and each model is implemented by a parallel code.

For instance, to model a crack propagation, one uses a molecular dynamic code to represent the atomistic scale and an elasticity code using a finite element method to represent the continuum scale. Indeed, fully microscopic simulations of most domains of interest are not computationally feasible. Combining such different scales or physics is still a challenge to reach high performance and scalability.

Another prominent example is found in the field of aeronautic propulsion: the conjugate heat transfer simulation in complex geometries (as developed by the CFD team of CERFACS) requires to couple a fluid/convection solver (AVBP) with a solid/conduction solver (AVTP). As the AVBP code is much more CPU consuming than the AVTP code, there is an important computational imbalance between the two solvers.

In this context, one crucial issue is undoubtedly the load balancing of the whole coupled simulation that remains an open question. The goal here is to find the best data distribution for the whole coupled simulation and not only for each stand-alone code, as it is most usually done. Indeed, the naive balancing of each code on its own can lead to an important imbalance and to a communication bottleneck during the coupling phase, which can drastically decrease the overall performance. Therefore, we argue that it is required to model the coupling itself in order to ensure a good scalability, especially when running on massively parallel architectures (tens of thousands of processors/cores). In other words, one must develop new algorithms and software implementation to perform a *coupling-aware* partitioning of the whole application. Another related problem is the problem of resource allocation. This is particularly important for the global coupling efficiency and scalability, because each code involved in the coupling can be more or less computationally intensive, and there is a good trade-off to find between resources assigned to each code to avoid that one of them waits for the other(s). What does furthermore happen if the load of one code dynamically changes relatively to the other one? In such a case, it could be convenient to dynamically adapt the number of resources used during the execution.

There are several open algorithmic problems that we investigate in the HIEPACS project-team. All these problems uses a similar methodology based upon the graph model and are expressed as variants of the classic graph partitioning problem, using additional constraints or different objectives.

### 3.5.1. Dynamic load-balancing with variable number of processors

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 one to preserve the parallel code efficiency or 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 re-balance 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.

### 3.5.2. Load balancing of coupled codes

As stated above, the load balancing of coupled code is a major issue, that determines the performance of the complex simulation, and reaching high performance can be a great challenge. In this context, we develop new graph partitioning techniques, called *co-partitioning*. They address the problem of load balancing for two coupled codes: the key idea is to perform a "coupling-aware" partitioning, instead of partitioning these codes independently, as it is classically done. More precisely, we propose to enrich the classic graph model with *inter-edges*, which represent the coupled code interactions. We describe two new algorithms, and compare them to the naive approach. In the preliminary experiments we perform on synthetically-generated graphs, we notice that our algorithms succeed to balance the computational load in the coupling phase and in some cases they succeed to reduce the coupling communications costs. Surprisingly, we notice that our algorithms do not degrade significantly the global graph edge-cut, despite the additional constraints that they impose.

Besides this, our co-partitioning technique requires to use graph partitioning with *fixed vertices*, that raises serious issues with state-of-the-art software, that are classically based on the well-known recursive bisection paradigm (RB). Indeed, the RB method often fails to produce partitions of good quality. To overcome this issue, we propose a *new* direct $k$-way greedy graph growing algorithm, called KGGGP, that overcomes this issue and succeeds to produce partition with better quality than RB while respecting the constraint of fixed vertices. Experimental results compare KGGGP against state-of-the-art methods, such as Scotch, for real-life graphs available from the popular *DIMACS'10* collection.

### 3.5.3. Load balancing strategies for hybrid sparse linear solvers

Graph handling and partitioning play a central role in the activity described here but also in other numerical techniques detailed in sparse linear algebra Section. The Nested Dissection is now a well-known heuristic for sparse matrix ordering to both reduce the fill-in during numerical factorization and to maximize the number of independent computation tasks. By using the block data structure induced by the partition of separators of the original graph, very efficient parallel block solvers have been designed and implemented according to super-nodal or multi-frontal approaches. Considering hybrid methods mixing both direct and iterative solvers such as HIPS or MaPHyS, obtaining a domain decomposition leading to a good balancing of both the size of domain interiors and the size of interfaces is a key point for load balancing and efficiency in a parallel context.

We intend to revisit some well-known graph partitioning techniques in the light of the hybrid solvers and design new algorithms to be tested in the Scotch package.

**KERDATA Project-Team**

# 3. Research Program

## 3.1. Research axis 1: Convergence of Extreme-Scale Computing and Big Data Infrastructures

The tools and cultures of High Performance Computing and Big Data Analytics have evolved in divergent ways. This is to the detriment of both. However, big computations still generate and are needed to analyze Big Data. As scientific research increasingly depends on both high-speed computing and data analytics, the potential interoperability and scaling convergence of these two eco-systems is crucial to the future. Our objective for the next years is premised on the idea that we must begin to systematically map out and account for the ways in which the major issues associated with Big Data intersect with, impinge upon, and potentially change the plans that are now being laid for achieving Exascale computing.

### *3.1.1. High-performance storage for concurrent Big Data applications*

We argue that storage is a plausible pathway to convergence. In this context, we plan to focus on the needs of concurrent Big Data applications that require high-performance storage, as well as transaction support. Although blobs (binary large objects) are an increasingly popular storage model for such applications, state-of-the-art blob storage systems offer no transaction semantics. This demands users to coordinate data access carefully in order to avoid race conditions, inconsistent writes, overwrites and other problems that cause erratic behavior.

We argue there is a gap between existing storage solutions and application requirements, which limits the design of transaction-oriented applications. In this context, one idea on which we plan to focus our efforts is exploring how blob storage systems could provide built-in, multi-blob transactions, while retaining sequential consistency and high throughput under heavy access concurrency.

The early principles of this research direction have already raised interest from our partners at ANL (Rob Ross) and UPM (María Pérez) for potential collaborations. In this direction, the acceptance of our paper on the Týr transactional blob storage system as a Best Student Paper Award Finalist at the SC16 conference [25] is a very encouraging step.

### *3.1.2. Big Data analytics on Exascale HPC machines*

Big Data analytics is another interesting direction that we plan to explore, building on top of these converged storage architectures. Specifically, we will examine the ways in which Exascale infrastructures can be leveraged not only by HPC-centric, but also by scientific, cloud-centric applications. Many of the current state-of-the-art Big Data processing approaches, including Hadoop and Spark [46] are optimized to run on commodity machines. This impacts the mechanisms used to deal with failures and the limited network bandwidth.

A blind adoption of these systems on extreme-scale platforms would result in high overheads. It would therefore prevent users from fully benefiting from the high performance infrastructure. The objective that we set here is to explore design and implementation options for new data analytics systems that can exploit the features of extreme-scale HPC machines: multi-core nodes, multiple memory and storage technologies including a large memory, NVRAM, SSDs, etc.

**Collaboration.** *This axis is addressed in close collaboration with María Pérez (UPM), Rob Ross (ANL), Toni Cortes (BSC), Bogdan Nicolae (formerly at IBM Research, now at Huawei Research).*

*Relevant groups with similar interests are the following ones.*

- *The group of Jack Dongarra, Innovative Computing Laboratory at University of Tennessee/Oak Ridge National Laboratory, working on joint tools Exascale Computing and Big Data.*
- *The group of Satoshi Matsuoka, Tokyo Institute of Technology, working on system software for Clouds and HPC.*
- *The group of Franck Cappello at Argonne National Laboratory/NCSA working on on-demand data analytics and storage for extreme-scale simulations and experiments.*

## 3.2. Research axis 2: Advanced data processing on Clouds

The recent evolutions in the area of Big Data processing have pointed out some limitations of the initial Map-Reduce model. It is well suited for batch data processing, but less suited for real-time processing of dynamic data streams. New types of data-intensive applications emerge, e.g., for enterprises who need to perform analysis on their stream data in ways that can give fast results (i.e., in real time) at scale (e.g., click-stream analysis and network-monitoring log analysis). Similarly, scientists require fast and accurate data processing techniques in order to analyze their experimental data correctly at scale (e.g., collectively analysis of large data sets distributed in multiple geographically distributed locations).

Our plan is to revisit current data management techniques to cope with the volatile requirements of data-intensive applications on large-scale dynamic clouds in a cost-efficient way.

### *3.2.1. Stream-oriented, Big Data processing on clouds*

The state-of-the-art Hadoop Map-Reduce framework cannot deal with stream data applications, as it requires the data to be initially stored in a distributed file system in order to process them. To better cope with the above-mentioned requirements, several systems have been introduced for stream data processing such as Flink [41], Spark [46], Storm [47], and Google MillWheel [49]. These systems keep computation in memory to decrease latency, and preserve scalability by using data-partitioning or dividing the streams into a set of deterministic batch computations.

However, they are designed to work in dedicated environments and they do not consider the performance variability (i.e., network, I/O, etc.) caused by resource contention in the cloud. This variability may in turn cause high and unpredictable latency when output streams are transmitted to further analysis. Moreover, they overlook the dynamic nature of data streams and the volatility in their computation requirements. Finally, they still address failures in a best-effort manner.

Our objective is to investigate new approaches for reliable, stream Big Data processing on clouds. We will explore new mechanisms that expose resource heterogeneity (observed variability in resource utilization at runtime) when scheduling stream data applications. We will also investigate how to adapt to node failures automatically, and to adapt the failure handling techniques to the characteristics of the running application and to the root cause of failures.

### *3.2.2. Geographically distributed workflows on multi-site clouds*

Many data processing jobs in data-intensive applications are modeled as workflows (i.e., as sets of tasks linked according to their data and computation dependencies) to facilitate the management and analysis of large volumes of data. With the fast growth of volumes of data to be handled at larger and larger scales, geographically distributed workflows are emerging as a natural data processing paradigm. This may bring several benefits: resilience to failures, distribution across partitions (e.g., moving computation close to data or vice versa), elastic scaling to support usage bursts, user proximity, etc.

In this context, sharing, disseminating and analyzing the data sets results in frequent large-scale data movements across widely distributed sites. Studies show that the inter-datacenter traffic is expected to triple in the following years. Our objective is to investigate approaches to data management enabling an efficient execution of such geographically distributed workflows running on multi-site clouds.

While in the past years we have addressed some data management issues in this area, mainly in support to efficient task scheduling of scientific workflows running on multisite clouds, we will now focus on an increasingly common scenario where workflows generate and process a huge number of small files, which is particularly challenging. As such workloads generate a deluge of small and independent I/O operations, efficient data and metadata handling is critical. We will explore specific means to better hide latency for data and metadata access in such scenarios, as a way to improve global performance.

**Collaboration.** *This axis is addressed in close collaboration with María Pérez (UPM), Kate Keahey (ANL) and Toni Cortes (BSC).*

*Relevant groups with similar interests include the following ones.*

- *The AMPLab, UC Berkeley, USA, working on scheduling stream data applications in heterogeneous clouds.*
- *The group of Ewa Deelman, USC Information Sciences Institute, working on resource management for workflows in Clouds.*
- *The XTRA group, Nanyang Technological University, Singapore, working on resource provisioning for workflows in the cloud.*

## 3.3. Research axis 3: I/O management, in situ visualization and analysis on HPC systems at extreme scales

Over the past few years, the increasing amounts of data produced by large-scale simulations have motivated a shift from traditional offline data analysis to in situ analysis and visualization. In situ processing started by coupling a parallel simulation with an analysis or visualization library, to avoid the cost of writing data on storage and reading it back. Going beyond this simple pairwise tight coupling, complex analysis workflows today are graphs with one or more data sources and several interconnected analysis components.

### *3.3.1. Toward a joint optimized architecture for in situ visualization and advanced processing*

From Inria and ANL, four tools at least have emerged to address some challenges of coupling simulations with visualization packages or analysis workflows. Each of them focused on some particular aspect:

Damaris (Inria, [12], [4]) exploits dedicated cores to enable jitter-free I/O and in situ visualization;

Decaf (ANL, [36]) implements a coupling service for workflows;

FlowVR (Inria, [48]) connects workflow components for in situ processing;

Swift (ANL, [51]) focuses on implicitly parallel data flows and was optimized for Big Data processing.

Our plan is to explore how these tools could best leverage their respective strengths in a *joint optimized architecture for in situ visualization and advanced processing* in the HPC area. We published a preliminary study describing the lessons learned from using these tools in production environments with real applications [6]. Such a joint architecture will contribute to address the data volume and velocity challenges raised by data-intensive workflows, including complex data-intensive analytics phases. It may also impact, in a subsequent step, future data analysis pipelines for converged Big Data and HPC architectures.

**Collaboration.** *This axis is worked out in close collaboration with Rob Ross (ANL), Tom Peterka (ANL), Matthieu Dorier (ANL), Toni Cortes (BSC), Bruno Raffin (Inria). Some additional collaborations are in discussion with other members of JLESC, and with CEA and Total.*

*Relevant groups with similar interests include the following ones.*

- *The group of Manish Parashar at Rutgers University, USA (I/O management for HPC systems, in situ processing).*
- *The group of Scott Klasky at Oak Ridge National Lab, USA (I/O management for HPC systems, in situ processing).*
- *The CNRS IPSL laboratory (Sébastien Denvil, Pôle de modélisation du climat) in Paris, France (in situ data analytics).*

**POLARIS Team**

# 3. Research Program

## 3.1. Sound and Reproducible Experimental Methodology

**Participants:** Vincent Danjean, Nicolas Gast, Guillaume Huard, Arnaud Legrand, Jean-Marc Vincent.

Experiments in large scale distributed systems are costly, difficult to control and therefore difficult to reproduce. Although many of these digital systems have been built by men, they have reached such a complexity level that we are no longer able to study them like artificial systems and have to deal with the same kind of experimental issues as natural sciences. The development of a sound experimental methodology for the evaluation of resource management solutions is among the most important ways to cope with the growing complexity of computing environments. Although computing environments come with their own specific challenges, we believe such general observation problems should be addressed by borrowing good practices and techniques developed in many other domains of science.

This research theme builds on a transverse activity on *Open science and reproducible research* and is organized into the following two directions: (1) *Experimental design* (2) *Smart monitoring and tracing*. As we will explain in more detail hereafter, these transverse activity and research directions span several research areas and our goal within the POLARIS project is foremost to transfer original ideas from other domains of science to the distributed and high performance computing community.

## 3.2. Multi-Scale Analysis and Visualization

**Participants:** Vincent Danjean, Guillaume Huard, Arnaud Legrand, Jean-Marc Vincent, Panayotis Mertikopoulos.

As explained in the previous section, the first difficulty encountered when modeling large scale computer systems is to observe these systems and extract information on the behavior of both the architecture, the middleware, the applications, and the users. The second difficulty is to *visualize* and *analyze* such *multi-level traces to understand how the performance* of the application *can be improved*. While a lot of efforts are put into visualizing scientific data, in comparison little effort have gone into to developing techniques specifically tailored for understanding the behavior of distributed systems. Many visualization tools have been developed by renowned HPC groups since decades (e.g., BSC [87], Jülich and TU Dresden [86], [55], UIUC [74], [90], [77] and ANL [104], Inria Bordeaux [61] and Grenoble [106], ...) but most of these tools build on the classical information visualization mantra [95] that consists in always first presenting an overview of the data, possibly by plotting everything if computing power allows, and then to allow users to zoom and filter, providing details on demand. However in our context, the amount of data comprised in such traces is several orders of magnitude larger than the number of pixels on a screen and displaying even a small fraction of the trace leads to harmful visualization artifacts [82]. Such traces are typically made of events that occur at very different time and space scales, which unfortunately hinders classical approaches. Such visualization tools have focused on easing interaction and navigation in the trace (through gantcharts, intuitive filters, pie charts and kiviats) but they are very difficult to maintain and evolve and they require some significant experience to identify performance bottlenecks.

Therefore many groups have more recently proposed in combination to these tools some techniques to help identifying the structure of the application or regions (applicative, spatial or temporal) of interest. For example, researchers from the SDSC [85] propose some segment matching techniques based on clustering (Euclidean or Manhattan distance) of start and end dates of the segments that enables to reduce the amount of information to display. Researchers from the BSC use clustering, linear regression and Kriging techniques [94], [81], [73] to identify and characterize (in term of performance and resource usage) application phases and present aggregated representations of the trace [93]. Researchers from Jülich and TU Darmstadt have proposed techniques to identify specific communication patterns that incur wait states [101], [48]

## 3.3. Fast and Faithful Performance Prediction of Very Large Systems

**Participants:** Vincent Danjean, Bruno Gaujal, Arnaud Legrand, Florence Perronnin, Jean-Marc Vincent.

Evaluating the scalability, robustness, energy consumption and performance of large infrastructures such as exascale platforms and clouds raises severe methodological challenges. The complexity of such platforms mandates empirical evaluation but direct experimentation via an application deployment on a real-world testbed is often limited by the few platforms available at hand and is even sometimes impossible (cost, access, early stages of the infrastructure design, ...). Unlike direct experimentation via an application deployment on a real-world testbed, simulation enables fully repeatable and configurable experiments that can often be conducted quickly for arbitrary hypothetical scenarios. In spite of these promises, current simulation practice is often not conducive to obtaining scientifically sound results. To date, most simulation results in the parallel and distributed computing literature are obtained with simulators that are ad hoc, unavailable, undocumented, and/or no longer maintained. For instance, Naicken et al. [47] point out that out of 125 recent papers they surveyed that study peer-to-peer systems, 52% use simulation and mention a simulator, but 72% of them use a custom simulator. As a result, most published simulation results build on throw-away (short-lived and non validated) simulators that are specifically designed for a particular study, which prevents other researchers from building upon it. There is thus a strong need for recognized simulation frameworks by which simulation results can be reproduced, further analyzed and improved.

The *SimGrid* simulation toolkit [59], whose development is partially supported by POLARIS, is specifically designed for studying large scale distributed computing systems. It has already been successfully used for simulation of grid, volunteer computing, HPC, cloud infrastructures and we have constantly invested on the software quality, the scalability [51] and the validity of the underlying network models [49], [99]. Many simulators of MPI applications have been developed by renowned HPC groups (e.g., at SDSC [97], BSC [45], UIUC [105], Sandia Nat. Lab. [100], ORNL [58] or ETH Zürich [75] for the most prominent ones). Yet, to scale most of them build on restrictive network and application modeling assumptions that make them difficult to extend to more complex architectures and to applications that do not solely build on the MPI API. Furthermore, simplistic modeling assumptions generally prevent to faithfully predict execution times, which limits the use of simulation to indication of gross trends at best. Our goal is to improve the quality of SimGrid to the point where it can be used effectively on a daily basis by practitioners to *reproduce the dynamic of real HPC systems*.

We also develop another simulation software, *PSI* (Perfect SImulator) [63], [56], dedicated to the simulation of very large systems that can be modeled as Markov chains. PSI provides a set of simulation kernels for Markov chains specified by events. It allows one to sample stationary distributions through the Perfect Sampling method (pioneered by Propp and Wilson [88]) or simply to generate trajectories with a forward Monte-Carlo simulation leveraging time parallel simulation (pioneered by Fujimoto [67], Lin and Lazowska [80]). One of the strength of the PSI framework is its expressiveness that allows us to easily study networks with finite and infinite capacity queues [57]. Although PSI already allows to simulate very large and complex systems, our main objective is to push its scalability even further and *improve its capabilities by one or several orders of magnitude*.

## 3.4. Local Interactions and Transient Analysis in Adaptive Dynamic Systems

**Participants:** Nicolas Gast, Bruno Gaujal, Florence Perronnin, Jean-Marc Vincent, Panayotis Mertikopoulos.

Many systems can be effectively described by stochastic population models. These systems are composed of a set of $n$ entities interacting together and the resulting stochastic process can be seen as a continuous-time Markov chain with a finite state space. Many numerical techniques exist to study the behavior of Markov chains, to solve stochastic optimal control problems [89] or to perform model-checking [46]. These techniques, however, are limited in their applicability, as they suffer from the *curse of dimensionality*: the state-space grows exponentially with $n$.

This results in the need for approximation techniques. Mean field analysis offers a viable, and often very accurate, solution for large $n$. The basic idea of the mean field approximation is to count the number of entities that are in a given state. Hence, the fluctuations due to stochasticity become negligible as the number of entities grows. For large $n$, the system becomes essentially deterministic. This approximation has been originally developed in statistical mechanics for vary large systems composed of more than $10^{20}$ particles (called entities here). More recently, it has been claimed that, under some conditions, this approximation can be successfully used for stochastic systems composed of a few tens of entities. The claim is supported by various convergence results [68], [78], [103], and has been successfully applied in various domains: wireless networks [50], computer-based systems [71], [84], [98], epidemic or rumour propagation [60], [76] and bike-sharing systems [64]. It is also used to develop distributed control strategies [102], [83] or to construct approximate solutions of stochastic model checking problems [52], [53], [54].

Within the POLARIS project, we will continue developing both the theory behind these approximation techniques and their applications. Typically, these techniques require a homogeneous population of objects where the dynamics of the entities depend only on their state (the state space of each object must not scale with $n$ the number of objects) but neither on their identity nor on their spatial location. Continuing our work in [68], we would like to be able to handle heterogeneous or uncertain dynamics. Typical applications are caching mechanisms [71] or bike-sharing systems [65]. A second point of interest is the use of mean field or large deviation asymptotics to compute the time between two regimes [92] or to reach an equilibrium state. Last, mean-field methods are mostly descriptive and are used to analyse the performance of a given system. We wish to extend their use to solve optimal control problems. In particular, we would like to implement numerical algorithms that use the framework that we developed in [69] to build distributed control algorithms [62] and optimal pricing mechanisms [70].

## 3.5. Distributed Learning in Games and Online Optimization

**Participants:** Nicolas Gast, Bruno Gaujal, Arnaud Legrand, Panayotis Mertikopoulos.

Game theory is a thriving interdisciplinary field that studies the interactions between competing optimizing agents, be they humans, firms, bacteria, or computers. As such, game-theoretic models have met with remarkable success when applied to complex systems consisting of interdependent components with vastly different (and often conflicting) objectives – ranging from latency minimization in packet-switched networks to throughput maximization and power control in mobile wireless networks.

In the context of large-scale, decentralized systems (the core focus of the POLARIS project), it is more relevant to take an inductive, "bottom-up" approach to game theory, because the components of a large system cannot be assumed to perform the numerical calculations required to solve a very-large-scale optimization problem. In view of this, POLARIS' overarching objective in this area is to *develop novel algorithmic frameworks that offer robust performance guarantees when employed by all interacting decision-makers.*

A key challenge here is that most of the literature on learning in games has focused on *static* games with a *finite number of actions* per player [66], [91]. While relatively tractable, such games are ill-suited to practical applications where players pick an action from a continuous space or when their payoff functions evolve over time – this being typically the case in our target applications (e.g., routing in packet-switched networks or energy-efficient throughput maximization in wireless). On the other hand, the framework of online convex optimization typically provides worst-case performance bounds on the learner's *regret* that the agents can attain irrespectively of how their environment varies over time. However, if the agents' environment is determined chiefly by their interactions these bounds are fairly loose, so more sophisticated convergence criteria should be applied.

From an algorithmic standpoint, a further challenge occurs when players can only observe their own payoffs (or a perturbed version thereof). In this bandit-like setting regret-matching or trial-and-error procedures guarantee convergence to an equilibrium in a weak sense in certain classes of games. However, these results apply exclusively to static, finite games: learning in games with continuous action spaces and/or nonlinear payoff functions cannot be studied within this framework. Furthermore, even in the case of finite games,

the complexity of the algorithms described above is not known, so it is impossible to decide a priori which algorithmic scheme can be applied to which application.

**ROMA Project-Team**

# 3. Research Program

## 3.1. Algorithms for probabilistic environments

There are two main research directions under this research theme. In the first one, we consider the problem of the efficient execution of applications in a failure-prone environment. Here, probability distributions are used to describe the potential behavior of computing platforms, namely when hardware components are subject to faults. In the second research direction, probability distributions are used to describe the characteristics and behavior of applications.

### *3.1.1. Application resilience*

An application is resilient if it can successfully produce a correct result in spite of potential faults in the underlying system. Application resilience can involve a broad range of techniques, including fault prediction, error detection, error containment, error correction, checkpointing, replication, migration, recovery, etc. Faults are quite frequent in the most powerful existing supercomputers. The Jaguar platform, which ranked third in the TOP 500 list in November 2011 [59], had an average of 2.33 faults per day during the period from August 2008 to February 2010 [83]. The mean-time between faults of a platform is inversely proportional to its number of components. Progresses will certainly be made in the coming years with respect to the reliability of individual components. However, designing and building high-reliability hardware components is far more expensive than using lower reliability top-of-the-shelf components. Furthermore, low-power components may not be available with high-reliability. Therefore, it is feared that the progresses in reliability will far from compensate the steady projected increase of the number of components in the largest supercomputers. Already, application failures have a huge computational cost. In 2008, the DARPA white paper on "System resilience at extreme scale" [58] stated that high-end systems wasted 20% of their computing capacity on application failure and recovery.

In such a context, any application using a significant fraction of a supercomputer and running for a significant amount of time will have to use some fault-tolerance solution. It would indeed be unacceptable for an application failure to destroy centuries of CPU-time (some of the simulations run on the Blue Waters platform consumed more than 2,700 years of core computing time [54] and lasted over 60 hours; the most time-consuming simulations of the US Department of Energy (DoE) run for weeks to months on the most powerful existing platforms [57]).

Our research on resilience follows two different directions. On the one hand we design new resilience solutions, either generic fault-tolerance solutions or algorithm-based solutions. On the other hand we model and theoretically analyze the performance of existing and future solutions, in order to tune their usage and help determine which solution to use in which context.

### *3.1.2. Scheduling strategies for applications with a probabilistic behavior*

Static scheduling algorithms are algorithms where all decisions are taken before the start of the application execution. On the contrary, in non-static algorithms, decisions may depend on events that happen during the execution. Static scheduling algorithms are known to be superior to dynamic and system-oriented approaches in stable frameworks [65], [71], [72], [82], that is, when all characteristics of platforms and applications are perfectly known, known a priori, and do not evolve during the application execution. In practice, the prediction of application characteristics may be approximative or completely infeasible. For instance, the amount of computations and of communications required to solve a given problem in parallel may strongly depend on some input data that are hard to analyze (this is for instance the case when solving linear systems using full pivoting).

We plan to consider applications whose characteristics change dynamically and are subject to uncertainties. In order to benefit nonetheless from the power of static approaches, we plan to model application uncertainties and variations through probabilistic models, and to design for these applications scheduling strategies that are either static, or partially static and partially dynamic.

## 3.2. Platform-aware scheduling strategies

In this theme, we study and design scheduling strategies, focusing either on energy consumption or on memory behavior. In other words, when designing and evaluating these strategies, we do not limit our view to the most classical platform characteristics, that is, the computing speed of cores and accelerators, and the bandwidth of communication links.

In most existing studies, a single optimization objective is considered, and the target is some sort of absolute performance. For instance, most optimization problems aim at the minimization of the overall execution time of the application considered. Such an approach can lead to a very significant waste of resources, because it does not take into account any notion of efficiency nor of yield. For instance, it may not be meaningful to use twice as many resources just to decrease by 10% the execution time. In all our work, we plan to look only for algorithmic solutions that make a "clever" usage of resources. However, looking for the solution that optimizes a metric such as the efficiency, the energy consumption, or the memory-peak minimization, is doomed for the type of applications we consider. Indeed, in most cases, any optimal solution for such a metric is a sequential solution, and sequential solutions have prohibitive execution times. Therefore, it becomes mandatory to consider multi-criteria approaches where one looks for trade-offs between some user-oriented metrics that are typically related to notions of Quality of Service—execution time, response time, stretch, throughput, latency, reliability, etc.—and some system-oriented metrics that guarantee that resources are not wasted. In general, we will not look for the Pareto curve, that is, the set of all dominating solutions for the considered metrics. Instead, we will rather look for solutions that minimize some given objective while satisfying some bounds, or "budgets", on all the other objectives.

### *3.2.1. Energy-aware algorithms*

Energy-aware scheduling has proven an important issue in the past decade, both for economical and environmental reasons. Energy issues are obvious for battery-powered systems. They are now also important for traditional computer systems. Indeed, the design specifications of any new computing platform now always include an upper bound on energy consumption. Furthermore, the energy bill of a supercomputer may represent a significant share of its cost over its lifespan.

Technically, a processor running at speed $s$ dissipates $s^\alpha$ watts per unit of time with $2 \leq \alpha \leq 3$ [63], [64], [69]; hence, it consumes $s^\alpha \times d$ joules when operated during $d$ units of time. Therefore, energy consumption can be reduced by using speed scaling techniques. However it was shown in [84] that reducing the speed of a processor increases the rate of transient faults in the system. The probability of faults increases exponentially, and this probability cannot be neglected in large-scale computing [80]. In order to make up for the loss in *reliability* due to the energy efficiency, different models have been proposed for fault tolerance: (i) *re-execution* consists in re-executing a task that does not meet the reliability constraint [84]; (ii) *replication* consists in executing the same task on several processors simultaneously, in order to meet the reliability constraints [62]; and (iii) *checkpointing* consists in "saving" the work done at some certain instants, hence reducing the amount of work lost when a failure occurs [79].

Energy issues must be taken into account at all levels, including the algorithm-design level. We plan to both evaluate the energy consumption of existing algorithms and to design new algorithms that minimize energy consumption using tools such as resource selection, dynamic frequency and voltage scaling, or powering-down of hardware components.

### *3.2.2. Memory-aware algorithms*

For many years, the bandwidth between memories and processors has increased more slowly than the computing power of processors, and the latency of memory accesses has been improved at an even slower

pace. Therefore, in the time needed for a processor to perform a floating point operation, the amount of data transferred between the memory and the processor has been decreasing with each passing year. The risk is for an application to reach a point where the time needed to solve a problem is no longer dictated by the processor computing power but by the memory characteristics, comparable to the *memory wall* that limits CPU performance. In such a case, processors would be greatly under-utilized, and a large part of the computing power of the platform would be wasted. Moreover, with the advent of multicore processors, the amount of memory per core has started to stagnate, if not to decrease. This is especially harmful to memory intensive applications. The problems related to the sizes and the bandwidths of memories are further exacerbated on modern computing platforms because of their deep and highly heterogeneous hierarchies. Such a hierarchy can extend from core private caches to shared memory within a CPU, to disk storage and even tape-based storage systems, like in the Blue Waters supercomputer [55]. It may also be the case that heterogeneous cores are used (such as hybrid CPU and GPU computing), and that each of them has a limited memory.

Because of these trends, it is becoming more and more important to precisely take memory constraints into account when designing algorithms. One must not only take care of the amount of memory required to run an algorithm, but also of the way this memory is accessed. Indeed, in some cases, rather than to minimize the amount of memory required to solve the given problem, one will have to maximize data reuse and, especially, to minimize the amount of data transferred between the different levels of the memory hierarchy (minimization of the volume of memory inputs-outputs). This is, for instance, the case when a problem cannot be solved by just using the in-core memory and that any solution must be out-of-core, that is, must use disks as storage for temporary data.

It is worth noting that the cost of moving data has lead to the development of so called "communication-avoiding algorithms" [76]. Our approach is orthogonal to these efforts: in communication-avoiding algorithms, the application is modified, in particular some redundant work is done, in order to get rid of some communication operations, whereas in our approach, we do not modify the application, which is provided as a task graph, but we minimize the needed memory peak only by carefully scheduling tasks.

## 3.3. High-performance computing and linear algebra

Our work on high-performance computing and linear algebra is organized along three research directions. The first direction is devoted to direct solvers of sparse linear systems. The second direction is devoted to combinatorial scientific computing, that is, the design of combinatorial algorithms and tools that solve problems encountered in some of the other research themes, like the problems faced in the preprocessing phases of sparse direct solvers. The last direction deals with the adaptation of classical dense linear algebra kernels to the architecture of future computing platforms.

### *3.3.1. Direct solvers for sparse linear systems*

The solution of sparse systems of linear equations (symmetric or unsymmetric, often with an irregular structure, from a few hundred thousand to a few hundred million equations) is at the heart of many scientific applications arising in domains such as geophysics, structural mechanics, chemistry, electromagnetism, numerical optimization, or computational fluid dynamics, to cite a few. The importance and diversity of applications are a main motivation to pursue research on sparse linear solvers. Because of this wide range of applications, any significant progress on solvers will have a significant impact in the world of simulation. Research on sparse direct solvers in general is very active for the following main reasons:

- many applications fields require large-scale simulations that are still too big or too complicated with respect to today's solution methods;
- the current evolution of architectures with massive, hierarchical, multicore parallelism imposes to overhaul all existing solutions, which represents a major challenge for algorithm and software development;
- the evolution of numerical needs and types of simulations increase the importance, frequency, and size of certain classes of matrices, which may benefit from a specialized processing (rather than resort to a generic one).

Our research in the field is strongly related to the software package MUMPS (see Section 6.1 ). MUMPS is both an experimental platform for academics in the field of sparse linear algebra, and a software package that is widely used in both academia and industry. The software package MUMPS enables us to (i) confront our research to the real world, (ii) develop contacts and collaborations, and (iii) receive continuous feedback from real-life applications, which is extremely critical to validate our research work. The feedback from a large user community also enables us to direct our long-term objectives towards meaningful directions.

In this context, we aim at designing parallel sparse direct methods that will scale to large modern platforms, and that are able to answer new challenges arising from applications, both efficiently—from a resource consumption point of view—and accurately—from a numerical point of view. For that, and even with increasing parallelism, we do not want to sacrifice in any manner numerical stability, based on threshold partial pivoting, one of the main originalities of our approach (our "trademark") in the context of direct solvers for distributed-memory computers; although this makes the parallelization more complicated, applying the same pivoting strategy as in the serial case ensures numerical robustness of our approach, which we generally measure in terms of sparse backward error. In order to solve the hard problems resulting from the always-increasing demands in simulations, special attention must also necessarily be paid to memory usage (and not only execution time). This requires specific algorithmic choices and scheduling techniques. From a complementary point of view, it is also necessary to be aware of the functionality requirements from the applications and from the users, so that robust solutions can be proposed for a wide range of applications.

Among direct methods, we rely on the multifrontal method [73], [74], [78]. This method usually exhibits a good data locality and hence is efficient in cache-based systems. The task graph associated with the multifrontal method is in the form of a tree whose characteristics should be exploited in a parallel implementation.

Our work is organized along two main research directions. In the first one we aim at efficiently addressing new architectures that include massive, hierarchical parallelism. In the second one, we aim at reducing the running time complexity and the memory requirements of direct solvers, while controlling accuracy.

### 3.3.2. Combinatorial scientific computing

Combinatorial scientific computing (CSC) is a recently coined term (circa 2002) for interdisciplinary research at the intersection of discrete mathematics, computer science, and scientific computing. In particular, it refers to the development, application, and analysis of combinatorial algorithms to enable scientific computing applications. CSC's deepest roots are in the realm of direct methods for solving sparse linear systems of equations where graph theoretical models have been central to the exploitation of sparsity, since the 1960s. The general approach is to identify performance issues in a scientific computing problem, such as memory use, parallel speed up, and/or the rate of convergence of a method, and to develop combinatorial algorithms and models to tackle those issues.

Our target scientific computing applications are (i) the preprocessing phases of direct methods (in particular MUMPS), iterative methods, and hybrid methods for solving linear systems of equations, and tensor decomposition algorithms; and (ii) the mapping of tasks (mostly the sub-tasks of the mentioned solvers) onto modern computing platforms. We focus on the development and use of graph and hypergraph models, and related tools such as hypergraph partitioning algorithms, to solve problems of load balancing and task mapping. We also focus on bipartite graph matching and vertex ordering methods for reducing the memory overhead and computational requirements of solvers. Although we direct our attention on these models and algorithms through the lens of linear system solvers, our solutions are general enough to be applied to some other resource optimization problems.

### 3.3.3. Dense linear algebra on post-petascale multicore platforms

The quest for efficient, yet portable, implementations of dense linear algebra kernels (QR, LU, Cholesky) has never stopped, fueled in part by each new technological evolution. First, the LAPACK library [67] relied on BLAS level 3 kernels (Basic Linear Algebra Subroutines) that enable to fully harness the computing power of a single CPU. Then the SCALAPACK library [66] built upon LAPACK to provide a coarse-grain parallel version, where processors operate on large block-column panels. Inter-processor communications

occur through highly tuned MPI send and receive primitives. The advent of multi-core processors has led to a major modification in these algorithms [68], [81], [77]. Each processor runs several threads in parallel to keep all cores within that processor busy. Tiled versions of the algorithms have thus been designed: dividing large block-column panels into several tiles allows for a decrease in the granularity down to a level where many smaller-size tasks are spawned. In the current panel, the diagonal tile is used to eliminate all the lower tiles in the panel. Because the factorization of the whole panel is now broken into the elimination of several tiles, the update operations can also be partitioned at the tile level, which generates many tasks to feed all cores.

The number of cores per processor will keep increasing in the following years. It is projected that high-end processors will include at least a few hundreds of cores. This evolution will require to design new versions of libraries. Indeed, existing libraries rely on a static distribution of the work: before the beginning of the execution of a kernel, the location and time of the execution of all of its component is decided. In theory, static solutions enable to precisely optimize executions, by taking parameters like data locality into account. At run time, these solutions proceed at the pace of the slowest of the cores, and they thus require a perfect load-balancing. With a few hundreds, if not a thousand, cores per processor, some tiny differences between the computing times on the different cores ("jitter") are unavoidable and irremediably condemn purely static solutions. Moreover, the increase in the number of cores per processor once again mandates to increase the number of tasks that can be executed in parallel.

We study solutions that are part-static part-dynamic, because such solutions have been shown to outperform purely dynamic ones [70]. On the one hand, the distribution of work among the different nodes will still be statically defined. On the other hand, the mapping and the scheduling of tasks inside a processor will be dynamically defined. The main difficulty when building such a solution will be to design lightweight dynamic schedulers that are able to guarantee both an excellent load-balancing and a very efficient use of data locality.

## 3.4. Compilers, code optimization and high-level synthesis for FPGA

*Christophe Alias and Laure Gonnord asked to join the ROMA team temporarily, starting from September 2015. This was accepted by the team and by Inria. The text below describes their research domain. The results that they have achieved in 2016 are included in this report.*

The advent of parallelism in supercomputers, in embedded systems (smartphones, plane controllers), and in more classical end-user computers increases the need for high-level code optimization and improved compilers. Being able to deal with the complexity of the upcoming software and hardware while keeping energy consumption at a reasonnable level is one of the main challenges cited in the Hipeac Roadmap which among others cites the two major issues :

- Enhance the efficiency of the design of embedded systems, and especially the design of optimized specialized hardware.
- Invent techniques to "expose data movement in applications and optimize them at runtime and compile time and to investigate communication-optimized algorithms".

In particular, the rise of embedded systems and high performance computers in the last decade has generated new problems in code optimization, with strong consequences on the research area. The main challenge is to take advantage of the characteristics of the specific hardware (generic hardware, or hardware accelerators). The long-term objective is to provide solutions for the end-user developers to use at their best the huge opportunities of these emerging platforms.

### *3.4.1. Compiler algorithms for irregular applications*

In the last decades, several frameworks has emerged to design efficient compiler algorithms. The efficiency of all the optimizations performed in compilers strongly relies on performant *static analyses* and *intermediate representations*. Among these representations, the polyhedral model [75] focus on regular programs, whose execution trace is predictable statically. The program and the data accessed are represented with a single mathematical object endowed with powerful algorithmic techniques for reasoning about it. Unfortunately, most of the algorithms used in scientific computing do not fit totally in this category.

We plan to explore the extensions of these techniques to handle irregular programs with while loops and complex data structures (such as trees, and lists). This raises many issues. We cannot represent finitely all the possible executions traces. Which approximation/representation to choose? Then, how to adapt existing techniques on approximated traces while preserving the correctness? To address these issues, we plan to incorporate new ideas coming from the abstract interpretation community: control flow, approximations, and also shape analysis; and from the termination community: rewriting is one of the major techniques that are able to handle complex data structures and also recursive programs.

### *3.4.2. High-level synthesis for FPGA*

Energy consumption bounds the performance of supercomputers since the end of Dennard scaling. Hence, reducing the electrical energy spent in a computation is the major challenge raised by Exaflop computing. Novel hardware, software, compilers and operating systems must be designed to increase the energy efficiency (in flops/watt) of data manipulation and computation itself. In the last decade, many specialized hardware accelerators (Xeon Phi, GPGPU) has emerged to overcome the limitations of mainstream processors, by trading the genericity for energy efficiency. However, the best supercomputers can only reach 8 Gflops/watt [61], which is far less than the 50 Gflops/watt required by an Exaflop supercomputer. An extreme solution would be to trade all the genericity by using specialized circuits. However such circuits (application specific integrated circuits, ASIC) are usually too expensive for the HPC market and lacks of flexibility. Once printed, an ASIC cannot be modified. Any algorithm update (or bug fix) would be impossible, which clearly not realistic.

Recently, reconfigurable circuits (Field Programmable Gate Arrays, FPGA) has appeared as a credible alternative for Exaflop computing. Major companies (including Intel, Google, Facebook and Microsoft) show a growing interest to FPGA and promising results has been obtained. For instance, in 2015, Microsoft reaches 40 Gflop/watts on a data-center deep learning algorithm mapped on Intel/Altera Arria 10 FPGAs. We believe that FPGA will become the new building block for HPC and Big Data systems. Unfortunately, programming an FPGA is still a big challenge: the application must be defined at circuit level and use properly the logic cells. Hence, there is a strong need for a compiler technology able to *map complex applications specified in a high-level language*. This compiler technology is usually refered as high-level synthesis (HLS).

We plan to investigate how to extend the models and the algorithms developed by the HPC community to map automatically a complex application to an FPGA. This raises many issues. How to schedule/allocate the computations and the data on the FPGA in order to reduce the data transfers while keeping a high throughput? How to use optimally the resources of the FPGA while keeping a low critical path? To address these issues, we plan to develop novel execution models based on process networks and to extend/cross-fertilize the algorithms developed in both HPC and high-level synthesis communities. The purpose of the XtremLogic start-up company, co-founded by Christophe Alias and Alexandru Plesco is to transfer the results of this research to an industrial level compiler.

**STORM Team**

# 3. Research Program

## 3.1. Parallel Computing and Architectures

Following the current trends of the evolution of HPC systems architectures, it is expected that future Exascale systems (i.e. Sustaining $10^{18}$ flops) will have millions of cores. Although the exact architectural details and trade-offs of such systems are still unclear, it is anticipated that an overall concurrency level of $O(10^9)$ threads/tasks will probably be required to feed all computing units while hiding memory latencies. It will obviously be a challenge for many applications to scale to that level, making the underlying system sound like "embarrassingly parallel hardware."

From the programming point of view, it becomes a matter of being able to expose extreme parallelism within applications to feed the underlying computing units. However, this increase in the number of cores also comes with architectural constraints that actual hardware evolution prefigures: computing units will feature extra-wide SIMD and SIMT units that will require aggressive code vectorization or "SIMDization", systems will become hybrid by mixing traditional CPUs and accelerators units, possibly on the same chip as the AMD APU solution, the amount of memory per computing unit is constantly decreasing, new levels of memory will appear, with explicit or implicit consistency management, etc. As a result, upcoming extreme-scale system will not only require unprecedented amount of parallelism to be efficiently exploited, but they will also require that applications generate adaptive parallelism capable to map tasks over heterogeneous computing units.

The current situation is already alarming, since European HPC end-users are forced to invest in a difficult and time-consuming process of tuning and optimizing their applications to reach most of current supercomputers' performance. It will go even worse at horizon 2020 with the emergence of new parallel architectures (tightly integrated accelerators and cores, high vectorization capabilities, etc.) featuring unprecedented degree of parallelism that only too few experts will be able to exploit efficiently. As highlighted by the ETP4HPC initiative, existing programming models and tools won't be able to cope with such a level of heterogeneity, complexity and number of computing units, which may prevent many new application opportunities and new science advances to emerge.

The same conclusion arises from a non-HPC perspective, for single node embedded parallel architectures, combining heterogeneous multicores, such as the ARM big.LITTLE processor and accelerators such as GPUs or DSPs. The need and difficulty to write programs able to run on various parallel heterogeneous architectures has led to initiatives such as HSA, focusing on making it easier to program heterogeneous computing devices. The growing complexity of hardware is a limiting factor to the emergence of new usages relying on new technology.

## 3.2. Scientific and Societal Stakes

In the HPC context, simulation is already considered as a third pillar of science with experiments and theory. Additional computing power means more scientific results, and the possibility to open new fields of simulation requiring more performance, such as multi-scale, multi-physics simulations. Many scientific domains able to take advantage of Exascale computers, these "Grand Challenges" cover large panels of science, from seismic, climate, molecular dynamics, theoretical and astrophysics physics... Besides, embedded applications are also able to take advantage of these performance increase. There is still an on-going trend where dedicated hardware is progressively replaced by off-the-shelf components, adding more adaptability and lowering the cost of devices. For instance, Error Correcting Codes in cell phones are still hardware chips, but with the forthcoming 5G protocol, new software and adaptative solutions relying on low power multicores are also explored. New usages are also appearing, relying on the fact that large computing capacities are becoming more affordable and widespread. This is the case for instance with Deep Neural Networks where the training phase can be done

on supercomputers and then used in embedded mobile systems. The same consideration applies for big data problems, of internet of things, where small sensors provide large amount of data that need to be processed in short amount of time. Even though the computing capacities required for such applications are in general a different scale from HPC infrastructures, there is still a need in the future for high performance computing applications.

However, the outcome of new scientific results and the development of new usages for mobile, embedded systems will be hindered by the complexity and high level of expertise required to tap the performance offered by future parallel heterogeneous architectures.

## 3.3. Towards More Abstraction

As emphasized by initiatives such as the European Exascale Software Initiative (EESI), the European Technology Platform for High Performance Computing (ETP4HPC), or the International Exascale Software Initiative (IESP), the HPC community needs new programming APIs and languages for expressing heterogeneous massive parallelism in a way that provides an abstraction of the system architecture and promotes high performance and efficiency. The same conclusion holds for mobile, embedded applications that require performance on heterogeneous systems.

This crucial challenge given by the evolution of parallel architectures therefore comes from this need to make high performance accessible to the largest number of developpers, abstracting away architectural details providing some kind of performance portability. Disruptive uses of the new technology and groundbreaking new scientific results will not come from code optimization or task scheduling, but they require the design of new algorithms that require the technology to be tamed in order to reach unprecedented levels of performance.

Runtime systems and numerical libraries are part of the answer, since they may be seen as building blocks optimized by experts and used as-is by application developers. The first purpose of runtime systems is indeed to provide *abstraction*. Runtime systems offer a uniform programming interface for a specific subset of hardware (e.g., OpenGL or DirectX are well-established examples of runtime systems dedicated to hardware-accelerated graphics) or low-level software entities (e.g., POSIX-thread implementations). They are designed as thin user-level software layers that complement the basic, general purpose functions provided by the operating system calls. Applications then target these uniform programming interfaces in a portable manner. Low-level, hardware dependent details are hidden inside runtime systems. The adaptation of runtime systems is commonly handled through drivers. The abstraction provided by runtime systems thus enables portability. Abstraction alone is however not enough to provide portability of performance, as it does nothing to leverage low-level-specific features to get increased performance. Consequently, the second role of runtime systems is to *optimize* abstract application requests by dynamically mapping them onto low-level requests and resources as efficiently as possible. This mapping process makes use of scheduling algorithms and heuristics to decide the best actions to take for a given metric and the application state at a given point in its execution time. This allows applications to readily benefit from available underlying low-level capabilities to their full extent without breaking their portability. Thus, optimization together with abstraction allows runtime systems to offer portability of performance. Numerical libraries provide sets of highly optimized kernels for a given field (dense or sparse linear algebra, FFT, etc.) either in an autonomous fashion or using an underlying runtime system.

Application domains cannot resort to libraries for all codes however, computation patterns such as stencils are a representative example of such difficulty. The compiler technology plays here a central role, in managing high level semantics, either through templates, domain specific languages or annotations. Compiler optimizations, and the same applies for runtime optimizations, are limited by the level of semantics they manage. Providing part of the algorithmic knowledge of an application, for instance knowing that it computes a 5-point stencil and then performs a dot product, would lead to more opportunities to adapt parallelism, memory structures, and is a way to leverage the evolving hardware.

Compilers and runtime play a crucial role in the future of high performance applications, by defining the input language for users, and optimizing/transforming it into high performance code. The objective of STORM is to propose better interactions between compiler and runtime and more semantics for both approaches. We recall in the following section the expertise of the team.

**TADAAM Team**

# 3. Research Program

## 3.1. Need for System-Scale Optimization

Firstly, in order for applications to make the best possible use of the available resources, it is impossible to expose all the low-level details of the hardware to the program, as it would make impossible to achieve portability. Hence, the standard approach is to add intermediate layers (programming models, libraries, compilers, runtime systems, etc.) to the software stack so as to bridge the gap between the application and the hardware. With this approach, optimizing the application requires to express its parallelism (within the imposed programming model), organize the code, schedule and load-balance the computations, etc. In other words, in this approach, the way the code is written and the way it is executed and interpreted by the lower layers drives the optimization. In any case, this approach is centered on how computations are performed. Such an approach is therefore no longer sufficient, as the way an application is executing does depend less and less on the organization of computation and more and more on the way its data is managed.

Secondly, modern large-scale parallel platforms comprise tens to hundreds of thousand nodes [0]. However, very few applications use the whole machine. In general, an application runs only on a subset of the nodes [0]. Therefore, most of the time, an application shares the network, the storage and other resources with other applications running concurrently during its execution. Depending on the allocated resources, it is not uncommon that the execution of one application interferes with the execution of a neighboring one.

Lastly, even if an application is running alone, each element of the software stack often performs its own optimization independently. For instance, when considering an hybrid MPI/OpenMP application, one may realize that threads are concurrently used within the OpenMP runtime system, within the MPI library for communication progression, and possibly within the computation library (BLAS) and even within the application itself (pthreads). However, none of these different classes of threads are aware of the existence of the others. Consequently, the way they are executed, scheduled, prioritized does not depend on their relative roles, their locations in the software stack nor on the state of the application.

The above remarks show that in order to go beyond the state-of-the-art, it is necessary to design a new set of mechanisms allowing cross-layer and system-wide optimizations so as to optimize the way data is allocated, accessed and transferred by the application.

## 3.2. Scientific Challenges and Research Issues

In TADAAM, we will tackle the problem of efficiently executing an application, at system-scale, on an HPC machine. We assume that the application is already optimized (efficient data layout, use of effective libraries, usage of state-of-the-art compilation techniques, etc.). Nevertheless, even a statically optimized application will not be able to be executed at scale without considering the following dynamic constraints: machine topology, allocated resources, data movement and contention, other running applications, access to storage, etc. Thanks to the proposed layer, we will provide a simple and efficient way for already existing applications, as well as new ones, to express their needs in terms of resource usage, locality and topology, using a high-level semantic.

It is important to note that we target the optimization of each application independently but also several applications at the same time and at system-scale, taking into account their resource requirement, their network usage or their storage access. Furthermore, dealing with code-coupling application is an intermediate use-case that will also be considered.

[0] More than 22,500 XE6 compute node for the BlueWaters system; 5040 B510 Bullx Nodes for the Curie machine; more than 49,000 BGQ nodes for the MIRA machine.

[0] In 2014, the median case was 2048 nodes for the BlueWaters system and, for the first year of the Curie machine, the median case was 256 nodes

Several issues have to be considered. The first one consists in providing relevant **abstractions and models to describe the topology** of the available resources **and the application behavior**.

Therefore, the first question we want to answer is: **"How to build scalable models and efficient abstractions enabling to understand the impact of data movement, topology and locality on performance?"** These models must be sufficiently precise to grasp the reality, tractable enough to enable efficient solutions and algorithms, and simple enough to remain usable by non-hardware experts. We will work on (1) better describing the memory hierarchy, considering new memory technologies; (2) providing an integrated view of the nodes, the network and the storage; (3) exhibiting qualitative knowledge; (4) providing ways to express the multi-scale properties of the machine. Concerning abstractions, we will work on providing general concepts to be integrated at the application or programming model layers. The goal is to offer means, for the application, to express its high-level requirements in terms of data access, locality and communication, by providing abstractions on the notion of hierarchy, mesh, affinity, traffic metrics, etc.

In addition to the abstractions and the aforementioned models we need to **define a clean and expressive API in a scalable way**, in order for applications to express their needs (memory usage, affinity, network, storage access, model refinement, etc.).

Therefore, the second question we need to answer is: "**how to build a system-scale, stateful, shared layer that can gather applications needs expressed with a high-level semantic?**". This work will require not only to define a clean API where applications will express their needs, but also to define how such a layer will be shared across applications and will scale on future systems. The API will provide a simple yet effective way to express different needs such as: memory usage of a given portion of the code; start of a compute intensive part; phase where the network is accessed intensively; topology-aware affinity management; usage of storage (in read and/or write mode); change of the data layout after mesh refinement, etc. From an engineering point of view, the layer will have a hierarchical design matching the hardware hierarchy, so as to achieve scalability.

Once this has been done, the service layer, will have all the information about the environment characteristics and application requirements. We therefore need to design a set of **mechanisms to optimize applications execution**: communication, mapping, thread scheduling, data partitioning/mapping/movement, etc.

Hence, the last scientific question we will address is: "**How to design fast and efficient algorithms, mechanisms and tools to enable execution of applications at system-scale, in full a HPC ecosystem, taking into account topology and locality?**" A first set of research is related to thread and process placement according to the topology and the affinity. Another large field of study is related to data placement, allocation and partitioning: optimizing the way data is accessed and processed especially for mesh-based applications. The issues of transferring data across the network will also be tackled, thanks to the global knowledge we have on the application behavior and the data layout. Concerning the interaction with other applications, several directions will be tackled. Among these directions we will deal with matching process placement with resource allocation given by the batch scheduler or with the storage management: switching from a best-effort application centric strategy to global optimization scheme.

**ASCOLA Project-Team**

# 3. Research Program

## 3.1. Overview

Since we mainly work on new concepts for the language-based definition and implementation of complex software systems, we first briefly introduce some basic notions and problems of software components (understood in a broad sense, that is, including modules, objects, architecture description languages and services), aspects, and domain-specific languages. We conclude by presenting the main issues related to distribution and concurrency, in particular related to capacity planning issues that are relevant to our work.

## 3.2. Software Composition

**Modules and services.** The idea that building *software components*, i.e., composable prefabricated and parameterized software parts, was key to create an effective software industry was realized very early [72]. At that time, the scope of a component was limited to a single procedure. In the seventies, the growing complexity of software made it necessary to consider a new level of structuring and programming and led to the notions of information hiding, *modules*, and module interconnection languages [79], [55]. Information hiding promotes a black-box model of program development whereby a module implementation, basically a collection of procedures, is strongly encapsulated behind an interface. This makes it possible to guarantee logical invariant *properties* of the data managed by the procedures and, more generally, makes *modular reasoning* possible.

In the context of today's Internet-based information society, components and modules have given rise to *software services* whose compositions are governed by explicit *orchestration or choreography* specifications that support notions of global properties of a service-oriented architecture. These horizontal compositions have, however, to be frequently adapted dynamically. Dynamic adaptations, in particular in the context of software evolution processes, often conflict with a black-box composition model either because of the need for invasive modifications, for instance, in order to optimize resource utilization or modifications to the vertical compositions implementing the high-level services.

**Object-Oriented Programming.** Classes and objects provide another kind of software component, which makes it necessary to distinguish between *component types* (classes) and *component instances* (objects). Indeed, unlike modules, objects can be created dynamically. Although it is also possible to talk about classes in terms of interfaces and implementations, the encapsulation provided by classes is not as strong as the one provided by modules. This is because, through the use of inheritance, object-oriented languages put the emphasis on *incremental programming* to the detriment of modular programming. This introduces a white-box model of software development and more flexibility is traded for safety as demonstrated by the *fragile base class* issue [75].

**Architecture Description Languages.** The advent of distributed applications made it necessary to consider more sophisticated connections between the various building blocks of a system. The *software architecture* [84] of a software system describes the system as a composition of *components* and *connectors*, where the connectors capture the *interaction protocols* between the components [43]. It also describes the rationale behind such a given architecture, linking the properties required from the system to its implementation. *Architecture Description Languages* (ADLs) are languages that support architecture-based development [73]. A number of these languages make it possible to generate executable systems from architectural descriptions, provided implementations for the primitive components are available. However, guaranteeing that the implementation conforms to the architecture is an issue.

**Protocols.** Today, protocols constitute a frequently used means to precisely define, implement, and analyze contracts, notably concerning communication and security properties, between two or more hardware or software entities. They have been used to define interactions between communication layers, security properties of distributed communications, interactions between objects and components, and business processes.

Object interactions [77], component interactions [90], [81] and service orchestrations [56] are most frequently expressed in terms of *regular interaction protocols* that enable basic properties, such as compatibility, substitutability, and deadlocks between components to be defined in terms of basic operations and closure properties of finite-state automata. Furthermore, such properties may be analyzed automatically using, e.g., model checking techniques [53], [62].

However, the limited expressive power of regular languages has led to a number of approaches using more expressive *non-regular* interaction protocols that often provide distribution-specific abstractions, e.g., session types [66], or context-free or turing-complete expressiveness [82], [50]. While these protocol types allow conformance between components to be defined (e.g., using unbounded counters), property verification can only be performed manually or semi-automatically.

## 3.3. Programming languages for advanced modularization

The main driving force for the structuring means, such as components and modules, is the quest for clean *separation of concerns* [57] on the architectural and programming levels. It has, however, early been noted that concern separation in the presence of crosscutting functionalities requires specific language and implementation level support. Techniques of so-called *computational reflection*, for instance, Smith's 3-Lisp or Kiczales's CLOS meta-object protocol [85], [69] as well as metaprogramming techniques have been developed to cope with this problem but proven unwieldy to use and not amenable to formalization and property analysis due to their generality. Methods and techniques from two fields have been particularly useful in addressing such advanced modularization problems: Aspect-Oriented Software Development as the field concerned with the systematic handling of modularization issues and domain-specific languages that provide declarative and efficient means for the definition of crosscutting functionalities.

**Aspect-Oriented Software Development** [68], [41] has emerged over the previous decade as the domain of systematic exploration of crosscutting concerns and corresponding support throughout the software development process. The corresponding research efforts have resulted, in particular, in the recognition of *crosscutting* as a fundamental problem of virtually any large-scale application, and the definition and implementation of a large number of aspect-oriented models and languages.

However, most current aspect-oriented models, notably AspectJ [67], rely on pointcuts and advice defined in terms of individual execution events. These models are subject to serious limitations concerning the modularization of crosscutting functionalities in distributed applications, the integration of aspects with other modularization mechanisms such as components, and the provision of correctness guarantees of the resulting AO applications. They do, in particular, only permit the manipulation of distributed applications on a per-host basis, that is, without direct expression of coordination properties relating different distributed entities [86]. Similarly, current approaches for the integration of aspects and (distributed) components do not directly express interaction properties between sets of components but rather seemingly unrelated modifications to individual components [54]. Finally, current formalizations of such aspect models are formulated in terms of low-level semantic abstractions (see, e.g., Wand's et al semantics for AspectJ [89]) and provide only limited support for the analysis of fundamental aspect properties.

Different approaches have been put forward to tackle these problems, in particular, in the context of so-called *stateful* or *history-based aspect languages* [58], [59], which provide pointcut and advice languages that directly express rich relationships between execution events. Such languages have been proposed to directly express coordination and synchronization issues of distributed and concurrent applications [78], [48], [61], provide more concise formal semantics for aspects and enable analysis of their properties [44], [60], [58], [42]. Furthermore, first approaches for the definition of *aspects over protocols* have been proposed, as well as over regular structures [58] and non-regular ones [88], [76], which are helpful for the modular definition and verification of protocols over crosscutting functionalities.

They represent, however, only first results and many important questions concerning these fundamental issues remain open, in particular, concerning the semantics foundations of AOP and the analysis and enforcement of correctness properties governing its, potentially highly invasive, modifications.

**Domain-specific languages (DSLs)** represent domain knowledge in terms of suitable basic language constructs and their compositions at the language level. By trading generality for abstraction, they enable complex relationships among domain concepts to be expressed concisely and their properties to be expressed and formally analyzed. DSLs have been applied to a large number of domains; they have been particularly popular in the domain of software generation and maintenance [74], [92].

Many modularization techniques and tasks can be naturally expressed by DSLs that are either specialized with respect to the type of modularization constructs, such as a specific brand of software component, or to the compositions that are admissible in the context of an application domain that is targeted by a modular implementation. Moreover, software development and evolution processes can frequently be expressed by transformations between applications implemented using different DSLs that represent an implementation at different abstraction levels or different parts of one application.

Functionalities that crosscut a component-based application, however, complicate such a DSL-based transformational software development process. Since such functionalities belong to another domain than that captured by the components, different DSLs should be composed. Such compositions (including their syntactic expression, semantics and property analysis) have only very partially been explored until now. Furthermore, restricted composition languages and many aspect languages that only match execution events of a specific domain (e.g., specific file accesses in the case of security functionality) and trigger only domain-specific actions clearly are quite similar to DSLs but remain to be explored.

## 3.4. Distribution and Concurrency

While ASCOLA does not investigate distribution and concurrency as research domains per se (but rather from a software engineering and modularization viewpoint), there are several specific problems and corresponding approaches in these domains that are directly related to its core interests that include the structuring and modularization of large-scale distributed infrastructures and applications. These problems include crosscutting functionalities of distributed and concurrent systems, support for the evolution of distributed software systems, and correctness guarantees for the resulting software systems.

Underlying our interest in these domains is the well-known observation that large-scale distributed applications are subject to *numerous crosscutting functionalities* (such as the transactional behavior in enterprise information systems, the implementation of security policies, and fault recovery strategies). These functionalities are typically partially encapsulated in distributed infrastructures and partially handled in an ad hoc manner by using infrastructure services at the application level. Support for a more principled approach to the development and evolution of distributed software systems in the presence of crosscutting functionalities has been investigated in the field of *open adaptable middleware* [49], [71]. Open middleware design exploits the concept of reflection to provide the desired level of configurability and openness. However, these approaches are subject to several fundamental problems. One important problem is their insufficient, framework-based support that only allows partial modularization of crosscutting functionalities.

There has been some *criticism* on the use of *AspectJ-like aspect models* (which middleware aspect models like that of JBoss AOP are an instance of) for the modularization of distribution and concurrency related concerns, in particular, for transaction concerns [70] and the modularization of the distribution concern itself [86]. Both criticisms are essentially grounded in AspectJ's inability to explicitly represent sophisticated relationships between execution events in a distributed system: such aspects therefore cannot capture the semantic relationships that are essential for the corresponding concerns. History-based aspects, as those proposed by the ASCOLA project-team provide a starting point that is not subject to this problem.

From a point of view of language design and implementation, aspect languages, as well as domain specific languages for distributed and concurrent environments share many characteristics with existing distributed languages: for instance, event monitoring is fundamental for pointcut matching, different synchronization strategies and strategies for code mobility [64] may be used in actions triggered by pointcuts. However, these relationships have only been explored to a small degree. Similarly, the formal semantics and formal properties of aspect languages have not been studied yet for the distributed case and only rudimentarily for the concurrent one [44], [61].

## 3.5. Security

Security properties and policies over complex service-oriented and standalone applications become ever more important in the context of asynchronous and decentralized communicating systems. Furthermore, they constitute prime examples of crosscutting functionalities that can only be modularized in highly insufficient ways with existing programming language and service models. Security properties and related properties, such as accountability properties, are therefore very frequently awkward to express and difficult to analyze and enforce (provided they can be made explicit in the first place).

Two main issues in this space are particularly problematic from a compositional point of view. First, information flow properties of programming languages, such as flow properties of Javascript [46], and service-based systems [52] are typically specially-tailored to specific properties, as well as difficult to express and analyze. Second, the enforcement of security properties and security policies, especially accountability-related properties [80], [87], is only supported using ad hoc means with rudimentary support for property verification.

The ASCOLA team has recently started to work on providing formal methods, language support and implementation techniques for the modular definition and implementation of information flow properties as well as policy enforcement in service-oriented systems as well as, mostly object-oriented, programming languages.

## 3.6. Green IT

With the emergence of the Future Internet and the dawn of new IT architecture and computation models such as cloud computing, the usage of data centers (DC) as well as their power consumption increase dramatically [51]. Besides the ecological impact [65], energy consumption is a predominant criterion for DC providers since it determines the daily cost of their infrastructure. As a consequence, power management becomes one of the main challenges for DC infrastructures and more generally for large-scale distributed systems.

To address this problem, we study two approaches: a workload-driven [47] and power-driven one [83]. As part of the workload-driven solution, we adapt the power consumption of the DC depending on the application workload, and evaluate whether this workload to be more reactive. We develop a distributed system from the system to the service-oriented level mainly based on hardware and virtualization capabilities that is managed in a user-transparent fashion. As part of the power-driven approach, we address energy consumption issues through a strong synergy inside the infrastructure software stack and more precisely between applications and resource management systems. This approach is characterized by adapting QoS properties aiming at the best trade-off between cost of energy (typically from the regular electric grid), its availability (for instance, from renewable energy), and service degradation caused, for instance, by application reconfigurations to jobs suspensions.

## 3.7. Capacity Planning for Large Scale Distributed System

Since the last decade, cloud computing has emerged as both a new economic model for software (provision) and as flexible tools for the management of computing capacity [45]. Nowadays, the major cloud features have become part of the mainstream (virtualization, storage and software image management) and the big market players offer effective cloud-based solutions for resource pooling. It is now possible to deploy virtual infrastructures that involve virtual machines (VMs), middleware, applications, and networks in such a simple manner that a new problem has emerged since 2010: VM sprawl (virtual machine proliferation) that consumes valuable computing, memory, storage and energy resources, thus menacing serious resource shortages. Scientific approaches that address VM sprawl are both based on classical administration techniques like the lifecycle management of a large number of VMs as well as the arbitration and the careful management of all resources consumed and provided by the hosting infrastructure (energy, power, computing, memory, network etc.) [63], [91].

The ASCOLA team investigates fundamental techniques for cloud computing and capacity planning, from infrastructures to the application level. Capacity planning is the process of planning for, analyzing, sizing, managing and optimizing capacity to satisfy demand in a timely manner and at a reasonable cost. Applied to distributed systems like clouds, a capacity planning solution must mainly provide the minimal set of resources necessary for the proper execution of the applications (i.e., to ensure SLA). The main challenges in this context are: scalability, fault tolerance and reactivity of the solution in a large-scale distributed system, the analysis and optimization of resources to minimize the cost (mainly costs related to the energy consumption of datacenters), as well as the profiling and adaptation of applications to ensure useful levels of quality of service (throughput, response time, availability etc.).

Our solutions are mainly based on virtualized infrastructures that we apply from the IaaS to the SaaS levels. We are mainly concerned by the management and the execution of the applications by harnessing virtualization capabilities, the investigation of alternative solutions that aim at optimizing the trade-off between performance and energy costs of both applications and cloud resources, as well as arbitration policies in the cloud in the presence of energy-constrained resources.

**DIVERSE Project-Team**

# 3. Research Program

## 3.1. Scientific background

### *3.1.1. Model-driven engineering*

Model-Driven Engineering (MDE) aims at reducing the accidental complexity associated with developing complex software-intensive systems (e.g., use of abstractions of the problem space rather than abstractions of the solution space) [117]. It provides DIVERSE with solid foundations to specify, analyze and reason about the different forms of diversity that occur through the development lifecycle. A primary source of accidental complexity is the wide gap between the concepts used by domain experts and the low-level abstractions provided by general-purpose programming languages [88]. MDE approaches address this problem through modeling techniques that support separation of concerns and automated generation of major system artifacts from models (*e.g.,* test cases, implementations, deployment and configuration scripts). In MDE, a model describes an aspect of a system and is typically created or derived for specific development purposes [70]. Separation of concerns is supported through the use of different modeling languages, each providing constructs based on abstractions that are specific to an aspect of a system. MDE technologies also provide support for manipulating models, for example, support for querying, slicing, transforming, merging, and analyzing (including executing) models. Modeling languages are thus at the core of MDE, which participates to the development of a sound *Software Language Engineering*[0], including an unified typing theory that integrate models as first class entities [120].

Incorporating domain-specific concepts and high-quality development experience into MDE technologies can significantly improve developer productivity and system quality. Since the late nineties, this realization has led to work on MDE language workbenches that support the development of domain-specific modeling languages (DSMLs) and associated tools (*e.g.,* model editors and code generators). A DSML provides a bridge between the field in which domain experts work and the implementation (programming) field. Domains in which DSMLs have been developed and used include, among others, automotive, avionics, and the emerging cyber-physical systems. A study performed by Hutchinson et al. [94] provides some indications that DSMLs can pave the way for wider industrial adoption of MDE.

More recently, the emergence of new classes of systems that are complex and operate in heterogeneous and rapidly changing environments raises new challenges for the software engineering community. These systems must be adaptable, flexible, reconfigurable and, increasingly, self-managing. Such characteristics make systems more prone to failure when running and thus the development and study of appropriate mechanisms for continuous design and run-time validation and monitoring are needed. In the MDE community, research is focused primarily on using models at design, implementation, and deployment stages of development. This work has been highly productive, with several techniques now entering a commercialization phase. As software systems are becoming more and more dynamic, the use of model-driven techniques for validating and monitoring run-time behavior is extremely promising [102].

### *3.1.2. Variability modeling*

While the basic vision underlying *Software Product Lines* (SPL) can probably be traced back to David Parnas seminal article [110] on the Design and Development of Program Families, it is only quite recently that SPLs are emerging as a paradigm shift towards modeling and developing software system families rather than individual systems [108]. SPL engineering embraces the ideas of mass customization and software reuse. It focuses on the means of efficiently producing and maintaining multiple related software products, exploiting what they have in common and managing what varies among them.

[0] See http://planet-sl.org

Several definitions of the *software product line* concept can be found in the research literature. Clements *et al.* define it as *a set of software-intensive systems sharing a common, managed set of features that satisfy the specific needs of a particular market segment or mission and are developed from a common set of core assets in a prescribed way* [107]. Bosch provides a different definition [76]: *A SPL consists of a product line architecture and a set of reusable components designed for incorporation into the product line architecture. In addition, the PL consists of the software products developed using the mentioned reusable assets*. In spite of the similarities, these definitions provide different perspectives of the concept: *market-driven*, as seen by Clements *et al.*, and *technology-oriented* for Bosch.

SPL engineering is a process focusing on capturing the *commonalities* (assumptions true for each family member) and *variability* (assumptions about how individual family members differ) between several software products [82]. Instead of describing a single software system, a SPL model describes a set of products in the same domain. This is accomplished by distinguishing between elements common to all SPL members, and those that may vary from one product to another. Reuse of core assets, which form the basis of the product line, is key to productivity and quality gains. These core assets extend beyond simple code reuse and may include the architecture, software components, domain models, requirements statements, documentation, test plans or test cases.

The SPL engineering process consists of two major steps:

1. **Domain Engineering**, or *development for reuse*, focuses on core assets development.
2. **Application Engineering**, or *development with reuse*, addresses the development of the final products using core assets and following customer requirements.

Central to both processes is the management of **variability** across the product line [90]. In common language use, the term *variability* refers to *the ability or the tendency to change*. Variability management is thus seen as the key feature that distinguishes SPL engineering from other software development approaches [77]. Variability management is thus growingly seen as the cornerstone of SPL development, covering the entire development life cycle, from requirements elicitation [122] to product derivation [127] to product testing [106], [105].

Halmans *et al.* [90] distinguish between *essential* and *technical* variability, especially at requirements level. Essential variability corresponds to the customer's viewpoint, defining what to implement, while technical variability relates to product family engineering, defining how to implement it. A classification based on the dimensions of variability is proposed by Pohl *et al.* [112]: beyond **variability in time** (existence of different versions of an artifact that are valid at different times) and **variability in space** (existence of an artifact in different shapes at the same time) Pohl *et al.* claim that variability is important to different stakeholders and thus has different levels of visibility: **external variability** is visible to the customers while **internal variability**, that of domain artifacts, is hidden from them. Other classification proposals come from Meekel *et al.* [100] (feature, hardware platform, performances and attributes variability) or Bass *et al.* [68] who discuss about variability at the architectural level.

Central to the modeling of variability is the notion of *feature*, originally defined by Kang *et al.* as: *a prominent or distinctive user-visible aspect, quality or characteristic of a software system or systems* [96]. Based on this notion of *feature*, they proposed to use a *feature model* to model the variability in a SPL. A feature model consists of a *feature diagram* and other associated information: *constraints* and *dependency rules*. Feature diagrams provide a *graphical tree-like notation depicting the hierarchical organization of high level product functionalities* represented as features. The root of the tree refers to the complete system and is progressively decomposed into more refined features (tree nodes). Relations between nodes (features) are materialized by *decomposition edges* and *textual constraints*. Variability can be expressed in several ways. Presence or absence of a feature from a product is modeled using *mandatory* or *optional features*. Features are graphically represented as rectangles while some graphical elements (e.g., unfilled circle) are used to describe the variability (e.g., a feature may be optional).

Features can be organized into *feature groups*. Boolean operators *exclusive alternative (XOR)*, *inclusive alternative (OR)* or *inclusive (AND)* are used to select one, several or all the features from a feature group.

Dependencies between features can be modeled using *textual constraints*: *requires* (presence of a feature requires the presence of another), *mutex* (presence of a feature automatically excludes another). Feature attributes can be also used for modeling quantitative (e.g., numerical) information. Constraints over attributes and features can be specified as well.

Modeling variability allows an organization to capture and select which version of which variant of any particular aspect is wanted in the system [77]. To implement it cheaply, quickly and safely, redoing by hand the tedious weaving of every aspect is not an option: some form of automation is needed to leverage the modeling of variability [72], [84]. Model Driven Engineering (MDE) makes it possible to automate this weaving process [95]. This requires that models are no longer informal, and that the weaving process is itself described as a program (which is as a matter of facts an executable meta-model [103]) manipulating these models to produce for instance a detailed design that can ultimately be transformed to code, or to test suites [111], or other software artifacts.

### 3.1.3. Component-based software development

Component-based software development [121] aims at providing reliable software architectures with a low cost of design. Components are now used routinely in many domains of software system designs: distributed systems, user interaction, product lines, embedded systems, etc. With respect to more traditional software artifacts (e.g., object oriented architectures), modern component models have the following distinctive features [83]: description of requirements on services required from the other components; indirect connections between components thanks to ports and connectors constructs [98]; hierarchical definition of components (assemblies of components can define new component types); connectors supporting various communication semantics [80]; quantitative properties on the services [75].

In recent years component-based architectures have evolved from static designs to dynamic, adaptive designs (e.g., SOFA [80], Palladio [73], Frascati [104]). Processes for building a system using a statically designed architecture are made of the following sequential lifecycle stages: requirements, modeling, implementation, packaging, deployment, system launch, system execution, system shutdown and system removal. If for any reason after design time architectural changes are needed after system launch (e.g., because requirements changed, or the implementation platform has evolved, etc) then the design process must be reexecuted from scratch (unless the changes are limited to parameter adjustment in the components deployed).

Dynamic designs allow for *on the fly* redesign of a component based system. A process for dynamic adaptation is able to reapply the design phases while the system is up and running, without stopping it (this is different from stop/redeploy/start). This kind of process supports *chosen adaptation*, when changes are planned and realized to maintain a good fit between the needs that the system must support and the way it supports them [97]. Dynamic component-based designs rely on a component meta-model that supports complex life cycles for components, connectors, service specification, etc. Advanced dynamic designs can also take platform changes into account at run-time, without human intervention, by adapting themselves [81], [124]. Platform changes and more generally environmental changes trigger *imposed adaptation*, when the system can no longer use its design to provide the services it must support. In order to support an eternal system [74], dynamic component based systems must separate architectural design and platform compatibility. This requires support for heterogeneity, since platform evolutions can be partial.

The Models@runtime paradigm denotes a model-driven approach aiming at taming the complexity of dynamic software systems. It basically pushes the idea of reflection one step further by considering the reflection layer as a real model "something simpler, safer or cheaper than reality to avoid the complexity, danger and irreversibility of reality [115]". In practice, component-based (and/or service-based) platforms offer reflection APIs that make it possible to introspect the system (which components and bindings are currently in place in the system) and dynamic adaptation (by applying CRUD operations on these components and bindings). While some of these platforms offer rollback mechanisms to recover after an erroneous adaptation, the idea of Models@runtime is to prevent the system from actually enacting an erroneous adaptation. In other words, the "model at run-time" is a reflection model that can be uncoupled (for reasoning, validation, simulation purposes) and automatically resynchronized.

Heterogeneity is a key challenge for modern component based system. Until recently, component based techniques were designed to address a specific domain, such as embedded software for command and control, or distributed Web based service oriented architectures. The emergence of the Internet of Things paradigm calls for a unified approach in component based design techniques. By implementing an efficient separation of concern between platform independent architecture management and platform dependent implementations, *Models@runtime* is now established as a key technique to support dynamic component based designs. It provides DIVERSE with an essential foundation to explore an adaptation envelop at run-time.

Search Based Software Engineering [92] has been applied to various software engineering problems in order to support software developers in their daily work. The goal is to automatically explore a set of alternatives and assess their relevance with respect to the considered problem. These techniques have been applied to craft software architecture exhibiting high quality of services properties [89]. Multi Objectives Search based techniques [86] deal with optimization problem containing several (possibly conflicting) dimensions to optimize. These techniques provide DIVERSE with the scientific foundations for reasoning and efficiently exploring an envelope of software configurations at run-time.

### 3.1.4. Validation and verification

Validation and verification (V&V) theories and techniques provide the means to assess the validity of a software system with respect to a specific correctness envelop. As such, they form an essential element of DIVERSE's scientific background. In particular, we focus on model-based V&V in order to leverage the different models that specify the envelop at different moments of the software development lifecycle.

Model-based testing consists in analyzing a formal model of a system (*e.g.*, activity diagrams, which capture high-level requirements about the system, statecharts, which capture the expected behavior of a software module, or a feature model, which describes all possible variants of the system) in order to generate test cases that will be executed against the system. Model-based testing [123] mainly relies on model analysis, constraint solving [85] and search-based reasoning [99]. DIVERSE leverages in particular the applications of model-based testing in the context of highly-configurable systems and [125] interactive systems [101] as well as recent advances based on diversity for test cases selection [93].

Nowadays, it is possible to simulate various kinds of models. Existing tools range from industrial tools such as Simulink, Rhapsody or Telelogic to academic approaches like Omega [109], or Xholon [0]. All these simulation environments operate on homogeneous environment models. However, to handle diversity in software systems, we also leverage recent advances in heterogeneous simulation. Ptolemy [79] proposes a common abstract syntax, which represents the description of the model structure. These elements can be decorated using different directors that reflect the application of a specific model of computation on the model element. Metropolis [69] provides modeling elements amenable to semantically equivalent mathematical models. Metropolis offers a precise semantics flexible enough to support different models of computation. ModHel'X [91] studies the composition of multi-paradigm models relying on different models of computation.

Model-based testing and simulation are complemented by runtime fault-tolerance through the automatic generation of software variants that can run in parallel, to tackle the open nature of software-intensive systems. The foundations in this case are the seminal work about N-version programming [67], recovery blocks [113] and code randomization [71], which demonstrated the central role of diversity in software to ensure runtime resilience of complex systems. Such techniques rely on truly diverse software solutions in order to provide systems with the ability to react to events, which could not be predicted at design time and checked through testing or simulation.

### 3.1.5. Empirical software engineering

The rigorous, scientific evaluation of DIVERSE's contributions is an essential aspect of our research methodology. In addition to theoretical validation through formal analysis or complexity estimation, we also aim at applying state-of-the-art methodologies and principles of empirical software engineering. This approach encompasses a set of techniques for the sound validation contributions in the field of software engineering,

[0] http://www.primordion.com/Xholon/

ranging from statistically sound comparisons of techniques and large-scale data analysis to interviews and systematic literature reviews [118], [116]. Such methods have been used for example to understand the impact of new software development paradigms [78]. Experimental design and statistical tests represent another major aspect of empirical software engineering. Addressing large-scale software engineering problems often requires the application of heuristics, and it is important to understand their effects through sound statistical analyses [66].

## 3.2. Research axis

Figure 1 illustrates the four dimensions of software diversity, which form the core research axis of DIVERSE: the **diversity of languages** used by the stakeholders involved in the construction of these systems; the **diversity of features** required by the different customers; the **diversity of runtime environments** in which software has to run and adapt; the **diversity of implementations** that are necessary for resilience through redundancy. These four axis share and leverage the scientific and technological results developed in the area of model-driven engineering in the last decade. This means that all our research activities are founded on sound abstractions to reason about specific aspects of software systems, compose different perspectives and automatically generate parts of the system.

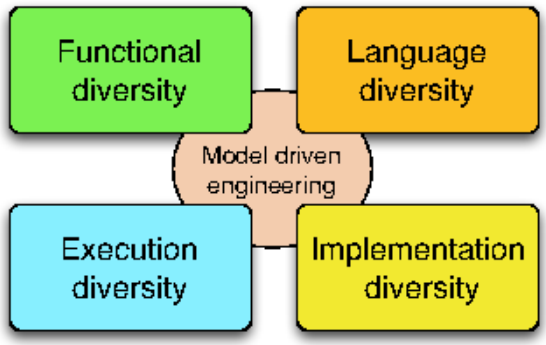


*Figure 1. The four research axis of DIVERSE, which rely on a MDE scientific background*

### *3.2.1. Software Language Engineering*

The engineering of systems involves many different stakeholders, each with their own domain of expertise. Hence more and more organizations are adopting Domain Specific Modeling Languages (DSMLs) to allow domain experts to express solutions directly in terms of relevant domain concepts [117], [88]. This new trend raises new challenges about designing DSMLs, evolving a set of DSMLs and coordinating the use of multiple DSLs for both DSL designers and DSL users.

*3.2.1.1. Challenges*

**Reusability** of software artifacts is a central notion that has been thoroughly studied and used by both academics and industrials since the early days of software construction. Essentially, designing reusable artifacts allows the construction of large systems from smaller parts that have been separately developed and validated, thus reducing the development costs by capitalizing on previous engineering efforts. However, it is still hardly possible for language designers to design typical language artifacts (e.g. language constructs, grammars, editors or compilers) in a reusable way. The current state of the practice usually prevents the reusability of language artifacts from one language to another, consequently hindering the emergence of real engineering techniques around software languages. Conversely, concepts and mechanisms that enable artifacts reusability abound in the software engineering community.

**Variability** in modeling languages occur in the definition of the abstract and concrete syntax as well as in the specification of the language's semantics. The major challenges met when addressing the need for variability are: (i) set principles for modeling language units that support the modular specification of a modeling language; and (ii) design mechanisms to assemble these units in a complete language, according to the set of authorized variation points for the modeling language family.

A new generation of complex software-intensive systems (for example smart health support, smart grid, building energy management, and intelligent transportation systems) presents new opportunities for leveraging modeling languages. The development of these systems requires expertise in diverse domains. Consequently, different types of stakeholders (e.g., scientists, engineers and end-users) must work in a coordinated manner on various aspects of the system across multiple development phases. DSMLs can be used to support the work of domain experts who focus on a specific system aspect, but they can also provide the means for coordinating work across teams specializing in different aspects and across development phases. The support and integration of DSMLs leads to what we call **the globalization of modeling languages**, *i.e.* the use of multiple languages for the coordinated development of diverse aspects of a system. One can make an analogy with world globalization in which relationships are established between sovereign countries to regulate interactions (e.g., travel and commerce related interactions) while preserving each country's independent existence.

*3.2.1.2. Scientific objectives*

We address reuse and variability challenges through the investigation of the time-honored concepts of substitutability, inheritance and components, evaluate their relevance for language designers and provide tools and methods for their inclusion in software language engineering. We will develop novel techniques for the modular construction of language extensions with the support of model syntactical variability. From the semantics perspective, we investigate extension mechanisms for the specification of variability in operational semantics, focusing on static introduction and heterogeneous models of computation. The definition of variation points for the three aspects of the language definition provides the foundations for the novel concept Language Unit (LU) as well as suitable mechanisms to compose such units.

We explore the necessary breakthrough in software languages to support modeling and simulation of heterogeneous and open systems. This work relies on the specification of executable domain specific modeling languages (DSMLs) to formalize the various concerns of a software-intensive system, and of models of computation (MoCs) to explicitly model the concurrency, time and communication of such DSMLs. We develop a framework that integrates the necessary foundations and facilities for designing and implementing executable and concurrent domain-specific modeling languages. It also provides unique features to specify composition operators between (possibly heterogeneous) DSMLs. Such specifications are amenable to support the edition, execution, graphical animation and analysis of heterogeneous models. The objective is to provide both a significant improvement of MoCs and DSMLs design and implementation; and the simulation based validation and verification of complex systems.

We see an opportunity for the automatic diversification of programs' computation semantics, for example through the diversification of compilers or virtual machines. The main impact of this artificial diversity is to provide flexible computation and thus ease adaptation to different execution conditions. A combination of static and dynamic analysis could support the identification of what we call *plastic computation zones* in the code. We identify different categories of such zones: (i) areas in the code in which the order of computation can vary (e.g., the order in which a block of sequential statements is executed); (ii) areas that can be removed, keeping the essential functionality [119] (e.g., skip some loop iterations); (iii) areas that can replaced by alternative code (e.g., replace a try-catch by a return statement). Once we know which zones in the code can be randomized, it is necessary to modify the model of computation to leverage the computation plasticity. This consists in introducing variation points in the interpreter to reflect the diversity of models of computation. Then, the choice of a given variation is performed randomly at run-time.

### *3.2.2. Variability Modeling and Engineering*

The systematic modeling of variability in software systems has emerged as an effective approach to document and reason about software evolutions and heterogeneity (*cf.* Section 3.1.2 ). Variability modeling character-

izes an "envelope" of possible software variations. The industrial use of variability models and their relation to software artifact models require a complete engineering framework, including composition, decomposition, analysis, configuration and artifact derivation, refactoring, re-engineering, extraction, and testing. This framework can be used both to tame imposed diversity and to manage chosen diversity.

*3.2.2.1. Challenges*

A fundamental problem is that the **number of variants** can be exponential in the number of options (features). Already with 300 boolean configuration options, approximately $10^{90}$ configurations exist – more than estimated count of atoms in the universe. Domains like automotive or operating systems have to manage more than 10000 options (e.g., Linux). Practitioners face the challenge of developing billions of variants. It is easy to forget a necessary constraint, leading to the synthesis of unsafe variants, or to under-approximate the capabilities of the software platform. Scalable modelling techniques are therefore crucial to specify and reason about a very large set of variants.

Model-driven development supports two ways to deal with the increasing number of concerns in complex systems: (1) multi-view modeling, *i.e.* when modeling each concern separately, and variability modeling. However, there is little support to combine both approaches consistently. Techniques to integrate both approaches will enable the construction of a consistent set of views and variation points in each view.

The design, construction and maintenance of software families have a major impact on **software testing**. Among the existing challenges, we can cite: the selection of test cases for a specific variant; the evolution of test suites with integration of new variants; the combinatorial explosion of the number of software configurations to be tested. Novel model-based techniques for test generation and test management in a software product line context are needed to overcome state-of-the-art limits we already observed in some projects.

*3.2.2.2. Scientific objectives*

We aim at developing scalable techniques to automatically analyze variability models and their interactions with other views on the software intensive system (requirements, architecture, design). These techniques provide two major advancements in the state of the art: (1) an extension of the semantics of variability models in order to enable the definition of attributes (*e.g.*, cost, quality of service, effort) on features and to include these attributes in the reasoning; (2) an assessment of the consistent specification of variability models with respect to system views (since variability is orthogonal to system modeling, it is currently possible to specify the different models in ways that are semantically meaningless). The former aspect of analysis is tackled through constraint solving and finite-domain constraint programming, while the latter aspect is investigated through automatic search-based techniques (similar to genetic algorithms) for the exploration of the space of interaction between variability and view models.

We aim to develop procedures to reverse engineer dependencies and features' sets from existing software artefacts – be it source code, configuration files, spreadsheets (e.g., product comparison matrices) or requirements. We expect to scale up (e.g., for extracting a very large number of variation points) and guarantee some properties (e.g., soundness of configuration semantics, understandability of ontological semantics). For instance, when building complex software-intensive systems, textual requirements are captured in very large quantities of documents. In this context, adequate models to formalize the organization of requirements documents and automated techniques to support impact analysis (in case of changes in the requirements) have to be developed.

We aim at developing sound methods and tools to integrate variability management in model-based testing activities. In particular, we will leverage requirement models as an essential asset to establish formal relations between variation points and test models. These relations will form the basis for novel algorithms that drive the systematic selection of test configurations that satisfy well-defined test adequacy criteria as well as the generation of test cases for a specific product in the product line.

### *3.2.3. Heterogeneous and dynamic software architectures*

Flexible yet dependable systems have to cope with heterogeneous hardware execution platforms ranging from smart sensors to huge computation infrastructures and data centers. Evolutions range from a mere change in the system configuration to a major architectural redesign, for instance to support addition of new features

or a change in the platform architecture (new hardware is made available, a running system switches to low bandwidth wireless communication, a computation node battery is running low, etc). In this context, we need to devise formalisms to reason about the impact of an evolution and about the transition from one configuration to another. It must be noted that this axis focuses on the use of models to drive the evolution from design time to run-time. Models will be used to (i) systematically define predictable configurations and variation points through which the system will evolve; (ii) develop behaviors necessary to handle unpredicted evolutions.

*3.2.3.1. Challenges*

The main challenge is to provide new homogeneous architectural modelling languages and efficient techniques that enable continuous software reconfiguration to react to changes. This work handles the challenges of handling the diversity of runtime infrastructures and managing the cooperation between different stakeholders. More specifically, the research developed in this axis targets the following dimensions of software diversity.

Platform architectural heterogeneity induces a first dimension of imposed diversity (type diversity). Platform reconfigurations driven by changing resources define another dimension of diversity (deployment diversity). To deal with these imposed diversity problems, we will rely on model based runtime support for adaptation, in the spirit of the dynamic distributed component framework developed by the Triskell team. Since the runtime environment composed of distributed, resource constrained hardware nodes cannot afford the overhead of traditional runtime adaptation techniques, we investigate the design of novel solutions relying on models@runtime and on specialized tiny virtual machines to offer resource provisioning and dynamic reconfigurations. In the next two years this research will be supported by the InfraJVM project.

Diversity can also be an asset to optimize software architecture. Architecture models must integrate multiple concerns in order to properly manage the deployment of software components over a physical platform. However, these concerns can contradict each other (*e.g.*, accuracy and energy). In this context, we investigate automatic solutions to explore the set of possible architecture models and to establish valid trade-offs between all concerns in case of changes.

*3.2.3.2. Scientific objectives*

**Automatic synthesis of optimal software architectures.** Implementing a service over a distributed platform (*e.g.*, a pervasive system or a cloud platform) consists in deploying multiple software components over distributed computation nodes. We aim at designing search-based solutions to (i) assist the software architect in establishing a good initial architecture (that balances between different factors such as cost of the nodes, latency, fault tolerance) and to automatically update the architecture when the environment or the system itself change. The choice of search-based techniques is motivated by the very large number of possible software deployment architectures that can be investigated and that all provide different trade-offs between qualitative factors. Another essential aspect that is supported by multi-objective search is to explore different architectural solutions that are not necessarily comparable. This is important when the qualitative factors are orthogonal to each other, such as security and usability for example.

**Flexible software architecture for testing and data management.** As the number of platforms on which software runs increases and different software versions coexist, the demand for testing environments also increases. For example, to test a software patch or upgrade, the number of testing environments is the product of the number of running environments the software supports and the number of coexisting versions of the software. Based on our first experiment on the synthesis of cloud environment using architectural models, our objective is to define a set of domain specific languages to catch the requirement and to design cloud environments for testing and data management of future internet systems from data centers to things. These languages will be interpreted to support dynamic synthesis and reconfiguration of a testing environment.

**Runtime support for heterogeneous environments.** Execution environments must provide a way to account or reserve resources for applications. However, current execution environments such as the Java Virtual Machine do not clearly define a notion of application: each framework has its own definition. For example, in OSGi, an application is a component, in JEE, an application is most of the time associated to a class loader, in the Multi-Tasking Virtual machine, an application is a process. The challenge consists in defining an execution environment that provides direct control over resources (CPU, Memory, Network I/O) independently from the

definition of an application. We propose to define abstract resource containers to account and reserve resources on a distributed network of heterogeneous devices.

### 3.2.4. Diverse implementations for resilience

Open software-intensive systems have to evolve over their lifetime in response to changes in their environment. Yet, most verification techniques assume a closed environment or the ability to predict all changes. Dynamic changes and evolutions thus represent a major challenge for these techniques that aim at assessing the correctness and robustness of the system. On the one hand, DIVERSE will adapt V&V techniques to handle diversity imposed by the requirements and the execution environment, on the other hand we leverage diversity to increase the robustness of software in face of unpredicted situations. More specifically, we address the following V&V challenges.

#### *3.2.4.1. Challenges*

One major challenge to build flexible and open yet dependable systems is that current software engineering techniques require architects to foresee all possible situations the system will have to face. However, openness and flexibility also mean unpredictability: unpredictable bugs, attacks, environmental evolutions, etc. Current fault-tolerance [113] and security [87] techniques provide software systems with the capacity of detecting accidental and deliberate faults. However, existing solutions assume that the set of bugs or vulnerabilities in a system does not evolve. This assumption does not hold for open systems, thus it is essential to revisit fault-tolerance and security solutions to account for diverse and unpredictable faults.

Diversity is known to be a major asset for the robustness of large, open, and complex systems (*e.g.*, economical or ecological systems). Following this observation, the software engineering literature provides a rich set of work that choose to implement diversity in software systems in order to improve robustness to attacks or to changes in quality of service. These works range from N-version programming to obfuscation of data structures or control flow, to randomization of instruction sets. An essential remaining challenge is to support the automatic synthesis and evolution of software diversity in open software-intensive systems. There is an opportunity to further enhance these techniques in order to cope with a wider diversity of faults, by multiplying the levels of diversity in the different software layers that are found in software-intensive systems (system, libraries, frameworks, application). This increased diversity must be based on artificial program transformations and code synthesis, which increase the chances of exploring novel solutions, better fitted at one point in time. The biological analogy also indicates that diversity should emerge as a side-effect of evolution, to prevent over-specialization towards one kind of diversity.

#### *3.2.4.2. Scientific objectives*

The main objective is to address one of the main limitations of N-version programming for fault-tolerant systems: the manual production and management of software diversity. Through automated injection of artificial diversity we aim at systematically increasing failure diversity and thus increasing the chances of early error detection at run-time. A fundamental assumption for this work is that software-intensive systems can be “good enough” [114], [126].

**Proactive program diversification.** We aim at establishing novel principles and techniques that favor the emergence of multiple forms of software diversity in software-intensive systems, in conjunction with the software adaptation mechanisms that leverage this diversity. The main expected outcome is a set of meta-design principles that maintain diversity in systems and the experimental demonstration of the effects of software diversity on the adaptive capacities of CASs. Higher levels of diversity in the system provide a pool of software solutions that can eventually be used to adapt to situations unforeseen at design time (bugs, crash, attacks, etc.). Principles of automated software diversification rely on the automated synthesis of variants in a software product line, as well as finer-grained program synthesis combining unsound transformations and genetic programming to explore the space of mutational robustness.

**Multi-tier software diversification.** We call multi-tier diversification the fact of diversifying several application software components simultaneously. The novelty of our proposal, with respect to the software diversity state of the art, is to diversify the application-level code (for example, diversify the business logics of the application), focusing on the technical layers found in web applications. The diversification of application software

code is expected to provide a diversity of failures and vulnerabilities in web server deployment. Web server deployment usually adopts a form of the Reactor architecture pattern, for scalability purposes: multiple copies of the server software stack, called request handlers, are deployed behind a load balancer. This architecture is very favorable for diversification, since by using the multiplicity of request handlers running in a web server we can simultaneously deploy multiple combinations of diverse software components. Then, if one handler is hacked or crashes the others should still be able to process client requests.

**FOCUS Project-Team**

# 3. Research Program

## 3.1. Models

The objective of Focus is to develop concepts, techniques, and possibly also tools, that may contribute to the analysis and synthesis of CBUS. Fundamental to these activities is *modeling*. Therefore designing, developing and studying computational models appropriate for CBUS is a central activity of the project. The models are used to formalise and verify important computational properties of the systems, as well as to propose new linguistic constructs.

The models we study are in the process calculi (e.g., the $\pi$-calculus) and $\lambda$-calculus tradition. Such models, with their emphasis on algebra, well address compositionality—a central property in our approach to problems. Accordingly, the techniques we employ are mainly operational techniques based on notions of behavioural equivalence, and techniques based on algebra, mathematical logics, and type theory.

**INDES Project-Team**

# 3. Research Program

## 3.1. Parallelism, concurrency, and distribution

Concurrency management is at the heart of diffuse programming. Since the execution platforms are highly heterogeneous, many different concurrency principles and models may be involved. Asynchronous concurrency is the basis of shared-memory process handling within multiprocessor or multicore computers, of direct or fifo-based message passing in distributed networks, and of fifo- or interrupt-based event handling in web-based human-machine interaction or sensor handling. Synchronous or quasi-synchronous concurrency is the basis of signal processing, of real-time control, and of safety-critical information acquisition and display. Interfacing existing devices based on these different concurrency principles within HOP or other diffuse programming languages will require better understanding of the underlying concurrency models and of the way they can nicely cooperate, a currently ill-resolved problem.

## 3.2. Web and functional programming

We are studying new paradigms for programming Web applications that rely on multi-tier functional programming [8]. We have created a Web programming environment named HOP. It relies on a single formalism for programming the server-side and the client-side of the applications as well as for configuring the execution engine.

HOP is a functional language based on the SCHEME programming language. That is, it is a strict functional language, fully polymorphic, supporting side effects, and dynamically type-checked. HOP is implemented as an extension of the BIGLOO compiler that we develop [9]. In the past, we have extensively studied static analyses (type systems and inference, abstract interpretations, as well as classical compiler optimizations) to improve the efficiency of compilation in both space and time.

## 3.3. Security of diffuse programs

The main goal of our security research is to provide scalable and rigorous language-based techniques that can be integrated into multi-tier compilers to enforce the security of diffuse programs. Research on language-based security has been carried on before in former Inria teams [2], [1]. In particular previous research has focused on controlling information flow to ensure confidentiality.

Typical language-based solutions to these problems are founded on static analysis, logics, provable cryptography, and compilers that generate correct code by construction [6]. Relying on the multi-tier programming language HOP that tames the complexity of writing and analysing secure diffuse applications, we are studying language-based solutions to prominent web security problems such as code injection and cross-site scripting, to name a few.

**PHOENIX Project-Team**

# 3. Research Program

## 3.1. Design-Driven Software Development

Raising the level of abstraction beyond programming is a very active research topic involving a range of areas, including software engineering, programming languages and formal verification. The challenge is to allow design dimensions of a software system, both functional and non-functional, to be expressed in a high-level way, instead of being encoded with a programming language. Such design dimensions can then be leveraged to verify conformance properties and to generate programming support.

Our research on this topic is to take up this challenge with an approach inspired by programming languages, introducing a full-fledged language for designing software systems and processing design descriptions both for verification and code generation purposes. Our approach is also DSL-inspired in that it defines a conceptual framework to guide software development. Lastly, to make our approach practical to software developers, we introduce a methodology and a suite of tools covering the development life-cycle.

To raise the level of abstraction beyond programming, the key approaches are model-driven engineering and architecture description languages. A number of *architecture description languages* have been proposed; they are either (1) coupled with a programming language (*e.g.,* [37]), providing some level of abstraction above programming, or (2) integrated into a programming language (*e.g.,* [33], [38]), mixing levels of abstraction. Furthermore, these approaches poorly leverage architecture descriptions to support programming, they are crudely integrated into existing development environments, or they are solely used for verification purposes. *Model-driven software development* is another actively researched area. This approach often lacks code generation and verification support. Finally, most (if not all) approaches related to our research goal are *general purpose*; their universal nature provides little, if any, guidance to design a software system. This situation is a major impediment to both reasoning about a design artifact and generating programming support.

## 3.2. Integrating Non-Functional Concerns into Software Design

Most existing design approaches do not address non-functional concerns. When they do, they do not provide an approach to non-functional concerns that covers the entire development life-cycle. Furthermore, they usually are general purpose, impeding the use of non-functional declarations for verification and code generation. For example, the Architecture Analysis & Design Language (AADL) is a standard dedicated to real-time embedded systems [34]. AADL provides language constructs for the specification of software systems (*e.g.,* component, port) and their deployment on execution platforms (*e.g.,* thread, process, memory). Using AADL, designers specify non-functional aspects by adding properties on language constructs (*e.g.,* the period of a thread) or using language extensions such as the Error Model Annex. [0] The software design concepts of AADL are still rather general purpose and give little guidance to the designer.

Beyond offering a conceptual framework, our language-based approach provides an ideal setting to address non-functional properties (*e.g.,* performance, reliability, security, ...). Specifically, a design language can be enriched with non-functional declarations to pursue three goals: (1) expanding further the type of conformance that can be checked between the design of a software system and its implementation or execution infrastructure, (2) enabling additional programming support and guidance, and (3) leveraging the design declarations to optimize the generated implementation.

We are investigating this idea by extending our design language with non-functional declarations. For example, we have addressed error handling [9], access conflicts to resources [36], quality of service constraints [35], and more recently, data delivery models and parallel computation models for masses of sensors citekaba:hal-01319730.

[0] The Error Model Annex is a standardized AADL extension for the description of errors [39].

Following our approach to paradigm-oriented software development, non-functional declarations are verified at design time, they generate support that guides and constrains programming, they produce a runtime system that preserves invariants and performs efficiently.

## 3.3. Human-Driven Software Design

Knowledge of the human characteristics (individual, social and organizational) allow the design of complex system and artifacts for increasing their efficacy. In our approach of assistive computing, a main challenge is the integration of facets of Human Factors in order to design technology support adapted to user needs in term of ergonomic properties (acceptability, usability, utility etc) and delivered functionalities (oriented task under user abilities contraints).

We adapt this approach to improve the independent living and self-determination of users with cognitive impairments by developing a variety of orchestration scenarios of networked objects (hardware/software) to provide a pervasive support to their activities. Human factors methodologies are adopted in our approach with the direct purpose the reliability and efficiency of the performance of digital support systems in respect of objectives of health and well-being of the person (monitoring, evaluation, and rehabilitation).

Precisely, our methodologies are based on a closed iterative loop, as described in the figure below :

- Identifying the person needs in a natural situation (*i.e.,* desired but problematic activities) according to Human Factors Models of activity (*i.e.,* environmental constraints; social support networks - caregivers and family; person's abilities)
- Designing environmental support that will assist the users to bypass their cognitive impairment (according to environmental models of cognitive compensatory mechanisms); and then implement this support in terms of technological solutions (scenarios of networked objects, hardware interface, software interface, interaction style, *etc*)
- Empirically evaluating the assistive solution based on human experimentations that includes ergonomic assessments (acceptability, usability, usefulness, *etc*) as well as longitudinal evaluations of use's efficacy in terms of activities performed by the individual, of satisfaction and well-being provided to the individual but also to his/her entourage (family and caregivers).

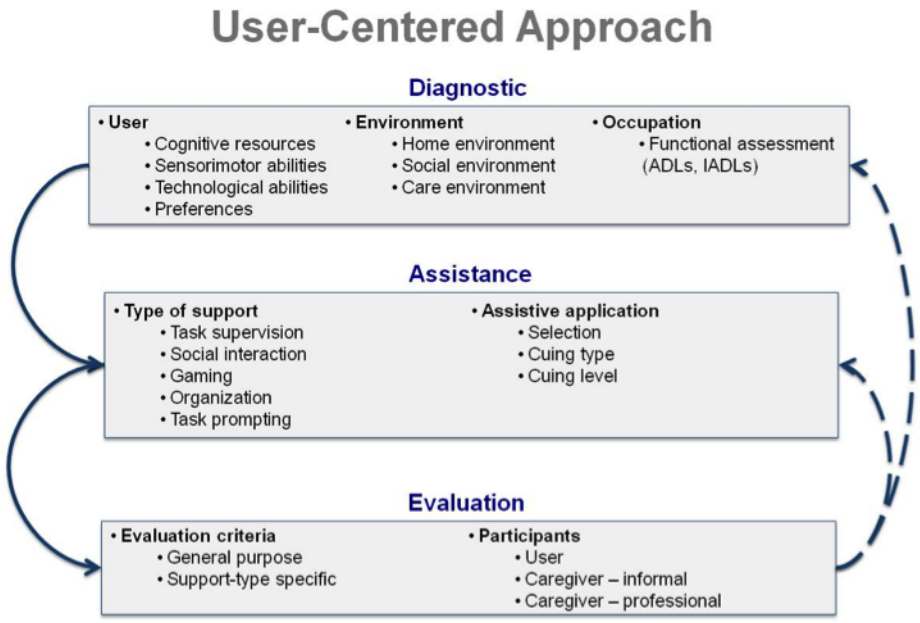


*Figure 1. User-Centered Approach*

**RMOD Project-Team**

# 3. Research Program

## 3.1. Software Reengineering

Strong coupling among the parts of an application severely hampers its evolution. Therefore, it is crucial to answer the following questions: How to support the substitution of certain parts while limiting the impact on others? How to identify reusable parts? How to modularize an object-oriented application?

Having good classes does not imply a good application layering, absence of cycles between packages and reuse of well-identified parts. Which notion of cohesion makes sense in presence of late-binding and programming frameworks? Indeed, frameworks define a context that can be extended by subclassing or composition: in this case, packages can have a low cohesion without being a problem for evolution. How to obtain algorithms that can be used on real cases? Which criteria should be selected for a given remodularization?

To help us answer these questions, we work on enriching Moose, our reengineering environment, with a new set of analyses [45], [44]. We decompose our approach in three main and potentially overlapping steps:

1. Tools for understanding applications,
2. Remodularization analyses,
3. Software Quality.

### *3.1.1. Tools for understanding applications*

**Context and Problems.** We are studying the problems raised by the understanding of applications at a larger level of granularity such as packages or modules. We want to develop a set of conceptual tools to support this understanding.

Some approaches based on Formal Concept Analysis (FCA) [75] show that such an analysis can be used to identify modules. However the presented examples are too small and not representative of real code.

**Research Agenda.**

FCA provides an important approach in software reengineering for software understanding, design anomalies detection and correction, but it suffers from two problems: (i) it produces lattices that must be interpreted by the user according to his/her understanding of the technique and different elements of the graph; and, (ii) the lattice can rapidly become so big that one is overwhelmed by the mass of information and possibilities [34]. We look for solutions to help people putting FCA to real use.

### *3.1.2. Remodularization analyses*

**Context and Problems.** It is a well-known practice to layer applications with bottom layers being more stable than top layers [61]. Until now, few works have attempted to identify layers in practice: Mudpie [77] is a first cut at identifying cycles between packages as well as package groups potentially representing layers. DSM (dependency structure matrix) [76], [69] seems to be adapted for such a task but there is no serious empirical experience that validates this claim. From the side of remodularization algorithms, many were defined for procedural languages [57]. However, object-oriented programming languages bring some specific problems linked with late-binding and the fact that a package does not have to be systematically cohesive since it can be an extension of another one [78], [48].

As we are designing and evaluating algorithms and analyses to remodularize applications, we also need a way to understand and assess the results we are obtaining.

**Research Agenda.** We work on the following items:

Layer identification. We propose an approach to identify layers based on a semi-automatic classification of package and class interrelationships that they contain. However, taking into account the wish or knowledge of the designer or maintainer should be supported.

Cohesion Metric Assessment. We are building a validation framework for cohesion/coupling metrics to determine whether they actually measure what they promise to. We are also compiling a number of traditional metrics for cohesion and coupling quality metrics to evaluate their relevance in a software quality setting.

### *3.1.3. Software Quality*

**Research Agenda.** Since software quality is fuzzy by definition and a lot of parameters should be taken into account we consider that defining precisely a unique notion of software quality is definitively a Grail in the realm of software engineering. The question is still relevant and important. We work on the two following items:

Quality models. We studied existing quality models and the different options to combine indicators — often, software quality models happily combine metrics, but at the price of losing the explicit relationships between the indicator contributions. There is a need to combine the results of one metric over all the software components of a system, and there is also the need to combine different metric results for any software component. Different combination methods are possible that can give very different results. It is therefore important to understand the characteristics of each method.

Bug prevention. Another aspect of software quality is validating or monitoring the source code to avoid the emergence of well known sources of errors and bugs. We work on how to best identify such common errors, by trying to identify earlier markers of possible errors, or by helping identifying common errors that programmers did in the past.

## 3.2. Language Constructs for Modular Design

While the previous axis focuses on how to help remodularizing existing software, this second research axis aims at providing new language constructs to build more flexible and recomposable software. We will build on our work on traits [73], [46] and classboxes [35] but also start to work on new areas such as isolation in dynamic languages. We will work on the following points: (1) Traits and (2) Modularization as a support for isolation.

### *3.2.1. Traits-based program reuse*

**Context and Problems.** Inheritance is well-known and accepted as a mechanism for reuse in object-oriented languages. Unfortunately, due to the coarse granularity of inheritance, it may be difficult to decompose an application into an optimal class hierarchy that maximizes software reuse. Existing schemes based on single inheritance, multiple inheritance, or mixins, all pose numerous problems for reuse.

To overcome these problems, we designed a new composition mechanism called Traits [73], [46]. Traits are pure units of behavior that can be composed to form classes or other traits. The trait composition mechanism is an alternative to multiple or mixin inheritance in which the composer has full control over the trait composition. The result enables more reuse than single inheritance without introducing the drawbacks of multiple or mixin inheritance. Several extensions of the model have been proposed [43], [65], [36], [47] and several type systems were defined [49], [74], [66], [59].

Traits are reusable building blocks that can be explicitly composed to share methods across unrelated class hierarchies. In their original form, traits do not contain state and cannot express visibility control for methods. Two extensions, stateful traits and freezable traits, have been proposed to overcome these limitations. However, these extensions are complex both to use for software developers and to implement for language designers.

**Research Agenda: Towards a pure trait language.** We plan distinct actions: (1) a large application of traits, (2) assessment of the existing trait models and (3) bootstrapping a pure trait language.

- To evaluate the expressiveness of traits, some hierarchies were refactored, showing code reuse [38]. However, such large refactorings, while valuable, may not exhibit all possible composition problems, since the hierarchies were previously expressed using single inheritance and following certain patterns. We want to redesign from scratch the collection library of Smalltalk (or part of it). Such a redesign should on the one hand demonstrate the added value of traits on a real large and redesigned library and on the other hand foster new ideas for the bootstrapping of a pure trait-based language.

  In particular we want to reconsider the different models proposed (stateless [46], stateful [37], and freezable [47]) and their operators. We will compare these models by (1) implementing a trait-based collection hierarchy, (2) analyzing several existing applications that exhibit the need for traits. Traits may be flattened [64]. This is a fundamental property that confers to traits their simplicity and expressiveness over Eiffel's multiple inheritance. Keeping these aspects is one of our priority in forthcoming enhancements of traits.
- Alternative trait models. This work revisits the problem of adding state and visibility control to traits. Rather than extending the original trait model with additional operations, we use a fundamentally different approach by allowing traits to be lexically nested within other modules. This enables traits to express (shared) state and visibility control by hiding variables or methods in their lexical scope. Although the traits' "flattening property" no longer holds when they can be lexically nested, the combination of traits with lexical nesting results in a simple and more expressive trait model. We formally specify the operational semantics of this combination. Lexically nested traits are fully implemented in AmbientTalk, where they are used among others in the development of a Morphic-like UI framework.
- We want to evaluate how inheritance can be replaced by traits to form a new object model. For this purpose we will design a minimal reflective kernel, inspired first from ObjVlisp [42] then from Smalltalk [52].

### *3.2.2. Reconciling Dynamic Languages and Isolation*

**Context and Problems.** More and more applications require dynamic behavior such as modification of their own execution (often implemented using reflective features [56]). For example, F-script allows one to script Cocoa Mac-OS X applications and Lua is used in Adobe Photoshop. Now in addition more and more applications are updated on the fly, potentially loading untrusted or broken code, which may be problematic for the system if the application is not properly isolated. Bytecode checking and static code analysis are used to enable isolation, but such approaches do not really work in presence of dynamic languages and reflective features. Therefore there is a tension between the need for flexibility and isolation.

**Research Agenda: Isolation in dynamic and reflective languages.** To solve this tension, we will work on *Sure*, a language where isolation is provided by construction: as an example, if the language does not offer field access and its reflective facilities are controlled, then the possibility to access and modify private data is controlled. In this context, layering and modularizing the meta-level [39], as well as controlling the access to reflective features [40], [41] are important challenges. We plan to:

- Study the isolation abstractions available in erights (http://www.erights.org) [63], [62], and Java's class loader strategies [58], [53].
- Categorize the different reflective features of languages such as CLOS [55], Python and Smalltalk [67] and identify suitable isolation mechanisms and infrastructure [50].
- Assess different isolation models (access rights, capabilities [68]...) and identify the ones adapted to our context as well as different access and right propagation.
- Define a language based on
  - the decomposition and restructuring of the reflective features [39],

- the use of encapsulation policies as a basis to restrict the interfaces of the controlled objects [72],
- the definition of method modifiers to support controlling encapsulation in the context of dynamic languages.

An open question is whether, instead of providing restricted interfaces, we could use traits to grant additional behavior to specific instances: without trait application, the instances would only exhibit default public behavior, but with additional traits applied, the instances would get extra behavior. We will develop *Sure*, a modular extension of the reflective kernel of Smalltalk (since it is one of the languages offering the largest set of reflective features such as pointer swapping, class changing, class definition...) [67].

**TACOMA Team**

# 3. Research Program

## 3.1. Collecting pertinent information

In our model, applications adapt their behavior (for instance, the level of automation) to the quality of their perception of the environment. This is important to alleviate the development constraint we usually have on automated system. We "just" have to be sure a given process will always operate at the right automation level given the precision, the completeness or the confidence it has on its own perception. For instance, a car passing through a cross would choose its speed depending on the confidence it has gained during perception data gathering. When it has not enough information or when it could not trust it, it should reduce the automation level, therefore the speed, to only rely on its own sensors. Such adaptation capability shift requirements from the design and deployment (availability, robustness, accuracy, etc.) to the **assessment of the environment perception** we aim to facilitate in this first research axis.

*Data characterization*. The quality (freshness, accuracy, confidence, reliability, confidentiality, etc.) of the data are of crucial importance to assess the quality of the perception and therefore to ensure proper behavior. The way data is produced, consolidated, and aggregated while flowing to the consumer has an impact on its quality. Moreover part of these quality attributes requires to gather information at several communication layers from various entities. For this purpose, we want to design **lightweight cross-layer interactions** to collect relevant data. As a "frugality" principle should guide our approach, it is not appropriate to build all attributes we can imagine. It is therefore necessary to identify attributes relevant to the application and to have mechanisms to activate/deactivate at run-time the process to collect them.

*Data fusion*. Raw data should be directly used only to determine low-level abstraction. Further help in abstracting from low-level details can be provided by **data fusion** mechanisms. A good (re)construction of a meaningful information for the application reduces the complexity of the pervasive applications and helps the developers to concentrate on the application logic rather on the management of raw data. Moreover, the reactivity required in pervasive systems and the aggregation of large amounts of data (and its processing) are antagonists. We study **software services that can be deployed closer to the edge of the network**. The exploration of data fusion technics will be guided by different criteria: relevance of abstractions produced for pervasive applications, anonymization of exploited raw data, processing time, etc.

*Assessing the correctness of the behavior*. To ease the design of new applications and to align the development of new products with the ever faster standard developments, continuous integration could be used in parallel with continuous conformance and interoperability testing. We already participate in the design of new shared platforms that aims at facilitating this providing remote testing tools. Unfortunately, it is not possible to be sure that all potential peers in the surrounding have a conform behavior. Moreover, upon failure or security breach, a piece of equipment could stop to operate properly and lead to global mis-behavior. We want to propose conceptual tools for **testing at runtime devices in the environment**. The result of such conformance or interoperability tests could be stored safely in the environment by authoritative testing entity. Then application could interact with the device with a higher confidence. The confidence level of a device could be part of the quality attribute of the information it contributed to generate. The same set of tools could be used to identify misbehaving device for maintenance purpose or to trigger further testing.

## 3.2. Building relevant abstraction for new interactions

The pervasive applications are often designed in an ad hoc manner depending on the targeted application area. Ressources (sensors / actuators, connected objets etc.) are often used in silos which complexify the implementation of rich pervasive computing scenarios. In the second research axis, we want to get away from technical aspects identifying **common and reusable system mechanisms** that could be used in various applications.

*Tagging the environment*. Information relative to environment could be stored by the application itself, but it could be complex to manage for mobile application since it could cross a large number of places with various features. Moreover the developer has to build its own representation of information especially when he wants to share information with other instances of the same application or with other applications. A promising approach is to store and to maintain this information associated to an object or to a place, in the environment itself. The infrastructure should provide services to application developers: add/retrieve information in the environment, share information and control who can access it, add computed properties to object for further usage. We want to study an **extensible model to describe and augment the environment**. Beyond a simple distributed storage, we have in mind a new kind of interaction between pervasive applications and changing environment and between applications themselves.

*Taking advantages of the spatial and temporal relationships*. To understand the world they have to interact with, pervasive applications often have to (re)built a model of it from the exchange they have with others or from their own observations. A part of the programmer's task consists in building a model of the spatial layout of the objects in the surrounding. The term *layout* can be understood in several ways: the co-location of multiple objects in the same vicinity, the physical arrangement of two objects relative to each other, or even the crossing of an object of a physical area to another, etc. Determining remotely these spatial properties (see figure 1 -a) is difficult without exchanging a lot of information. Properties related to the spatial layout are far easier to characterize locally. They could be abstracted from interaction pattern without any complex virtual representation of the environment (see figure 1 -b). We want to be able to rely on this type of spatial layout in a pervasive environment. In the prior years, the members of TACOMA already worked on **models for processing object interactions** in the physical world to automatically trigger processing. This was the case in particular of the spatial programming principle: physical space is treated as a tuple-space in which objects are automatically synchronized according to their spatial arrangement. We want to follow this approach by considering **richer and more expressive programming models.**

## 3.3. Acting on the environment

The conceptual tools we aim to study must be *frugal*: they use as less as possible resources, while having the possibility to use much more when it is required. Data needed by an application are not made available for "free"; for example, it costs energy to measure a characteristic of the environment, or to transmit it. So this "design frugality" requires **a fine-grained control** on how data is actually collected from the environment. The third research axis aims at designing solutions that give this control to application developers by **acting on the environment**.

*Acting on the data collection*. We want to be able to identify which information are reality needed during the perception elaboration process. If a piece of data is missing to build a given information with the appropriate quality level, the data collection mechanism should find relevant information in the environment or modify the way it aggregate it. These could lead to a modification of the behavior of the network layer and the path the piece of data use in the aggregation process.

*Acting on object interactions*. Object in the environment could adapt their behavior in a way that strongly depend on the object itself and that is difficult to generalize. Beyond the specific behaviors of actuators triggered through specialized or standard interfaces, the production of information required by an application could necessitate an adaptation at the object level (eg. calibration, sampling). The environment should then be able to initiate such adaption transparently to the application, which may not know all objects it passes by.

*Adapting object behaviors*. The radio communication layers become more flexible and able to adapt the way they use energy to what is really required for a given transmission. We already study how beamforming technics could be used to adapt multicast strategy for video services. We want to show how playing with these new parameters of transmissions (eg. beamforming, power, ...) allows to control spatial relationships objects could have. There is a tradeoff to find between the capacity of the medium, the electromagnetic pollution and the reactivity of the environment. We plan to expend our previous on interface selection and more generally on what we call **opportunistic networking**.

**COATI Project-Team**

# 3. Research Program

## 3.1. Research Program

Members of COATI have a strong expertise in the design and management of wired and wireless backbone, backhaul, broadband, and complex networks. On the one hand, we cope with specific problems such as energy efficiency in backhaul and backbone networks, routing reconfiguration in connection oriented networks (MPLS, WDM), traffic aggregation in SONET networks, compact routing in large-scale networks, survivability to single and multiple failures, etc. These specific problems often come from questions of our industrial partners. On the other hand, we study fundamental problems mainly related to routing and reliability that appear in many networks (not restricted to our main fields of applications) and that have been widely studied in the past. However, previous solutions do not take into account the constraints of current networks/traffic such as their huge size and their dynamics. COATI thus puts a significant research effort in the following directions:

- **Energy efficiency and Software-Defined Networks (SDN)** at both the design and management levels. More precisely, we plan to study the deployment of energy-efficient routing algorithm within SDN. We developed new algorithms in order to take into account the new constraints of SDN equipments and we evaluate their performance by simulation and by experimentation on a fat-tree architecture.
- **Larger networks:** Another challenge one has to face is the increase in size of practical instances. It is already difficult, if not impossible, to solve practical instances optimally using existing tools. Therefore, we have to find new ways to solve problems using reduction and decomposition methods, characterization of polynomial instances (which are surprisingly often the practical ones), or algorithms with acceptable practical performances.
- **Stochastic behaviors:** Larger topologies mean frequent changes due to traffic and radio fluctuations, failures, maintenance operations, growth, routing policy changes, etc. We aim at including these stochastic behaviors in our combinatorial optimization process to handle the dynamics of the system and to obtain robust designs of networks.

**DANTE Project-Team**

# 3. Research Program

## 3.1. Graph-based signal processing

**Participants:** Christophe Crespelle, Éric Fleury, Paulo Gonçalves Andrade, Márton Karsai, Sarah de Nigris, Sarra Ben Alaya, Hadrien Hours.

**Evolving networks can be regarded as "*out of equilibrium*" systems.** Indeed, their dynamics is typically characterized by non standard and intricate statistical properties, such as non-stationarity, long range memory effects, intricate space and time correlations.

Analyzing, modeling, and even defining adapted concepts for dynamic graphs is at the heart of DANTE. This is a largely open question that has to be answered by keeping a balance between specificity (solutions triggered by specific data sets) and generality (universal approaches disconnected from social realities). We will tackle this challenge from a graph-based signal processing perspective involving signal analysts and computer scientists, together with experts of the data domain application. One can distinguish two different issues in this challenge, one related to the graph-based organisation of the data and the other to the time dependency that naturally exits in the dynamic graph object. In both cases, a number of contributions can be found in the literature, albeit in different contexts. In our application domain, high-dimensional data "naturally reside" on the vertices of weighted graphs. The emerging field of signal processing on graphs merges algebraic and spectral graph theoretic concepts with computational harmonic analysis to process such signals on graphs [70].

As for the first point, adapting well-founded signal processing techniques to data represented as graphs is an emerging, yet quickly developing field which has already received key contributions. Some of them are very general and delineate ambitious programs aimed at defining universal, generally unsupervised methods for exploring high-dimensional data sets and processing them. This is the case for instance of the Â« diffusion wavelets Â» and Â« diffusion maps Â» pushed forward at Yale and Duke [54]. Others are more traditionally connected with standard signal processing concepts, in the spirit of elaborating new methodologies via some bridging between networks and time series, see, *e.g.*, ( [65] and references therein). Other viewpoints can be found as well, including multi-resolution Markov models [73], Bayesian networks or distributed processing over sensor networks [64]. Such approaches can be particularly successful for handling static graphs and unveiling aspects of their organisation in terms of dependencies between nodes, grouping, etc. Incorporating possible time dependencies within the whole picture calls however for the addition of an extra dimension to the problem "as it would be the case when switching from one image to a video sequence", a situation for which one can imagine to take advantage of the whole body of knowledge attached to non-stationary signal processing [55].

## 3.2. Theory and Structure of dynamic Networks

**Participants:** Christophe Crespelle, Éric Fleury, Anthony Busson, Márton Karsai.

**Characterization of the dynamics of complex networks.** We need to focus on intrinsic properties of evolving/dynamic complex networks. New notions (as opposed to classical static graph properties) have to be introduced: rate of vertices or links appearances or disappearances, the duration of link presences or absences. Moreover, more specific properties related to the dynamics have to be defined and are somehow related to the way to model a dynamic graph.

Through the systematic analysis and characterization of static network representations of many different systems, researchers of several disciplines have unveiled complex topologies and heterogeneous structures, with connectivity patterns statistically characterized by heavy-tails and large fluctuations, scale-free properties and non trivial correlations such as high clustering and hierarchical ordering [67]. A large amount of work has been devoted to the development of new tools for statistical characterisation and modelling of networks, in order to identify their most relevant properties, and to understand which growth mechanisms could lead to these properties. Most of those contributions have focused on static graphs or on dynamic process (*e.g.* diffusion) occurring on static graphs. This has called forth a major effort in developing the methodology to characterize the topology and temporal behavior of complex networks [67], [58], [74], [63], to describe the observed structural and temporal heterogeneities [52], [58], [53], to detect and measure emerging community structures [56], [71], [72], to see how the functionality of networks determines their evolving structure [62], and to determine what kinds of correlations play a role in their dynamics [59], [61], [66].

The challenge is now to extend this kind of statistical characterization to dynamical graphs. In other words, links in dynamic networks are temporal events, called contacts, which can be either punctual or last for some period of time. Because of the complexity of this analysis, the temporal dimension of the network is often ignored or only roughly considered. Therefore, fully taking into account the dynamics of the links into a network is a crucial and highly challenging issue.

Another powerful approach to model time-varying graphs is via activity driven network models. In this case, the only assumption relates to the distribution of activity rates of interacting entities. The activity rate is realistically broadly distributed and refers to the probability that an entity becomes active and creates a connection with another entity within a unit time step [69]. Even the generic model is already capable to recover some realistic features of the emerging graph, its main advantage is to provide a general framework to study various types of correlations present in real temporal networks. By synthesizing such correlations (*e.g.* memory effects, preferential attachment, triangular closing mechanisms, ...) from the real data, we are able to extend the general mechanism and build a temporal network model, which shows certain realistic feature in a controlled way. This can be used to study the effect of selected correlations on the evolution of the emerging structure [60] and its co-evolution with ongoing processes like spreading phenomena, synchronisation, evolution of consensus, random walk etc. [60], [68]. This approach allows also to develop control and immunisation strategies by fully considering the temporal nature of the backgrounding network.

## 3.3. Distributed Algorithms for dynamic networks: regulation, adaptation and interaction

**Participants:** Thomas Begin, Anthony Busson, Paulo Gonçalves Andrade, Isabelle Guérin Lassous.

**Dedicated algorithms for dynamic networks.** First, the dynamic network object itself trigger original algorithmic questions. It mainly concerns distributed algorithms that should be designed and deployed to efficiently measure the object itself and get an accurate view of its dynamic behavior. Such distributed measure should be "transparent", that is, it should introduce no bias or at least a bias that is controllable and corrigible. Such problem is encountered in all distributed metrology measures / distributed probes: P2P, sensor network, wireless network, QoS routing... This question raises naturally the intrinsic notion of adaptation and control of the dynamic network itself since it appears that autonomous networks and traffic aware routing are becoming crucial.

Communication networks are dynamic networks that potentially undergo high dynamicity. The dynamicity exhibited by these networks results from several factors including, for instance, changes in the topology and varying workload conditions. Although most implemented protocols and existing solutions in the literature can cope with a dynamic behavior, the evolution of their behavior operates identically whatever the actual properties of the dynamicity. For instance, parameters of the routing protocols (*e.g.* hello packets transmission frequency) or routing methods (*e.g.* reactive / proactive) are commonly hold constant regardless of the nodes mobility. Similarly, the algorithms ruling CSMA/CA (*e.g.* size of the contention window) are tuned identically and they do not change according to the actual workload and observed topology.

Dynamicity in computer networks tends to affect a large number of performance parameters (if not all) coming from various layers (viz. physical, link, routing and transport). To find out which ones matter the most for our intended purpose, we expect to rely on the tools developed by the two former axes. These quantities should capture and characterize the actual network dynamicity. Our goal is to take advantage of this latter information in order to refine existing protocols, or even to propose new solutions. More precisely, we will attempt to associate "fundamental" changes occurring in the underlying graph of a network (reported through graph-based signal tools) to quantitative performance that are matter of interests for networking applications and the end-users. We expect to rely on available testbeds such as Senslab and FIT to experiment our solutions and ultimately validate our approach.

**DIANA Project-Team**

# 3. Research Program

## 3.1. Service Transparency

Transparency is to provide network users and application developers with reliable information about the current or predicted quality of their communication services, and about potential leakages of personal information, or of other information related to societal interests of the user as a "connected citizen" (e.g. possible violation of network neutrality, opinion manipulation). Service transparency therefore means to provide information meaningful to users and application developers, such as quality of experience, privacy leakages, or opinion manipulation, etc. rather than network-level metrics such as available bandwidth, loss rate, delay or jitter.

The Internet is built around a best effort routing service that does not provide any guarantee to end users in terms of quality of service (QoS). The simplicity of the Internet routing service is at the root of its huge success. Unfortunately, a simple service means unpredicted quality at the access. Even though a considerable effort is done by operators and content providers to optimise the Internet content delivery chain, mainly by over-provisioning and sophisticated engineering techniques, service degradation is still part of the Internet. The proliferation of wireless and mobile access technologies, and the versatile nature of Internet traffic, make end users quality of experience (QoE) forecast even harder. As a matter of fact, the Internet is missing a dedicated measurement plane that informs the end users on the quality they obtain and in case of substantial service degradation, on the origin of this degradation. Current state of the art activities are devoted to building a distributed measurement infrastructure to perform active, passive and hybrid measurements in the wired Internet. However, the problem is exacerbated with modern terminals such as smartphones or tablets that do not facilitate the task for end users (they even make it harder) as they focus on simplifying the interface and limiting the control on the network, whereas the Internet behind is still the same in terms of the quality it provides. Interestingly, this same observation explains the existing difficulty to detect and prevent privacy leaks. We argue that the lack of transparency for diagnosing QoE and for detecting privacy leaks have the same root causes and can be solved using common primitives. For instance, in both cases, it is important to be able to link data packets to an application. Indeed, as the network can only access data packets, there must be a way to bind these packets to an application (to understand users QoE for this application or to associate a privacy leak to an application). This is however a complex task as the traffic might be obfuscated or encrypted. Our objectives in the research direction are the following:

- Design and develop measurement tools providing transparency, in spite of current complexity
- Deploy those measurement tools at the Internet's edge and make them useful for end users
- Propose measurements plane as an overlay or by exploiting in-network functionalities
- Adapt measurements techniques to network architectural change
- Provide measurements as native functionality in future network architecture

## 3.2. Open network architecture

We are surrounded by personal content of all types: photos, videos, documents, etc. The volume of such content is increasing at a fast rate, and at the same time, the spread of such content among all our connected devices (mobiles, storage devices, set-top boxes, etc) is also increasing. All this complicates the control of personal content by the user both in terms of access and sharing with other users. The access of the personal content in a seamless way independently of its location is a key challenge for the future of networks. Proprietary solutions exist, but apart from fully depending on one of them, there is no standard plane in the Internet for a seamless access to personal content. Therefore, providing network architectural support to design and develop content access and sharing mechanisms is crucial to allow users control their own data over heterogeneous underlying network or cloud services.

On the other hand, privacy is a growing concern for states, administrations, and companies. Indeed, for instance the French CNIL (entity in charge of citizens privacy in computer systems) puts privacy at the core of its activities by defining rules on any stored and collected private data. Also, companies start to use privacy preserving solutions as a competitive advantage. Therefore, understanding privacy leaks and preventing them is a problem that can already find support. However, all end-users do not *currently* put privacy as their first concern. Indeed, in face of two services with one of higher quality, they usually prefer the highest quality one whatever the privacy implication. This was, for instance, the case concerning the Web search service of Google that is more accurate but less privacy preserving than Bing. This is also the case for cloud services such as iCloud or Dropbox that are much more convenient than open source solutions, but very bad in terms of privacy. Therefore, to reach end-users, any privacy preserving solutions must offer a service equivalent to the best existing services.

We consider that it will be highly desirable for Internet users to be able to *easily* move their content from a provider to another and therefore not to depend on a content provider or a social network monopoly. This requires that the network provides built-in architectural support for content networking.

In this research direction, we will define a new *service abstraction layer* (SAL) that could become the new waist of the network architecture with network functionalities below (IP, SDN, cloud) and applications on top. SAL will define different services that are of use to all Internet users for accessing and sharing data (seamless content localisation and retrieval, privacy leakage protection, transparent vertical and horizontal handover, etc.). The biggest challenge here is to cope in the same time with large number of content applications requirements and high underlying networks heterogeneity while still providing efficient applications performance. This requires careful definition of the services primitives and the parameters to be exchanged through the service abstraction layer.

Two concurring factors make the concept behind SAL feasible and relevant today. First, the notion of scalable network virtualization that is a required feature to deploy SAL in real networks today has been discussed recently only. Second, the need for new services abstraction is recent. Indeed, fifteen years ago the Internet for the end-users was mostly the Web. Only eight years ago smartphones came into the picture of the Internet boosting the number of applications with new functionalities and risks. Since a few years, many discussions in the network communities took place around the actual complexity of the Internet and the difficulty to develop applications. Many different approaches have been discussed (such as CCN, SDN) that intend to solve only part of the complexity. SAL takes a broader architectural look at the problem and considers solutions such as CCN as mere use cases. Our objectives in this research direction include the following:

- Identify common key networking services required for content access and sharing
- Detect and prevent privacy leaks for content communication
- Enhance software defined networks for large scale heterogeneous environments
- Design and develop open Content Networking architecture
- Define a service abstraction layer as the thin waist for the future content network architecture
- Test and deploy different applications using SAL primitives on heterogeneous network technologies

## 3.3. Methodology

We follow an experimental approach that can be described in the following techniques:

- Measurements: the aim is to get a better view of a problem in quantifiable terms. Depending on the field of interest, this may involve large scale distributed systems crawling tools; active probing techniques to infer the status and properties of a complex and non controllable system as the Internet; or even crowdsourcing-based deployments for gathering data on real-users environments or behaviours.
- Experimental evaluation: once a new idea has been designed and implemented, it is of course very desirable to assess and quantify how effective it can be, before being able to deploy it on any realistic scale. This is why a wide range of techniques can be considered for getting early, yet as significant as possible, feedback on a given paradigm or implementation. The spectrum for such techniques span from simulations to real deployments in protected and/or controlled environments.

**DIONYSOS Project-Team**

# 3. Research Program

## 3.1. Introduction

The scientific foundations of our work are those of network design and network analysis. Specifically, this concerns the principles of packet switching and in particular of IP networks (protocol design, protocol testing, routing, scheduling techniques), and the mathematical and algorithmic aspects of the associated problems, on which our methods and tools are based.

These foundations are described in the following paragraphs. We begin by a subsection dedicated to Quality of Service (QoS) and Quality of Experience (QoE), since they can be seen as unifying concepts in our activities. Then we briefly describe the specific sub-area of model evaluation and about the particular multidisciplinary domain of network economics.

## 3.2. Quality of Service and Quality of Experience

Since it is difficult to develop as many communication solutions as possible applications, the scientific and technological communities aim towards providing general *services* allowing to give to each application or user a set of properties nowadays called "Quality of Service" (QoS), a terminology lacking a precise definition. This QoS concept takes different forms according to the type of communication service and the aspects which matter for a given application: for performance it comes through specific metrics (delays, jitter, throughput, etc.), for dependability it also comes through appropriate metrics: reliability, availability, or vulnerability, in the case for instance of WAN (Wide Area Network) topologies, etc.

QoS is at the heart of our research activities: We look for methods to obtain specific "levels" of QoS and for techniques to evaluate the associated metrics. Our ultimate goal is to provide tools (mathematical tools and/or algorithms, under appropriate software "containers" or not) allowing users and/or applications to attain specific levels of QoS, or to improve the provided QoS, if we think of a particular system, with an optimal use of the resources available. Obtaining a good QoS level is a very general objective. It leads to many different areas, depending on the systems, applications and specific goals being considered. Our team works on several of these areas. We also investigate the impact of network QoS on multimedia payloads to reduce the impact of congestion.

Some important aspects of the behavior of modern communication systems have subjective components: the quality of a video stream or an audio signal, *as perceived by the user*, is related to some of the previous mentioned parameters (packet loss, delays, ...) but in an extremely complex way. We are interested in analyzing these types of flows from this user-oriented point of view. We focus on the *user perceived quality*, in short, PQ, the main component of what is nowadays called Quality of Experience (in short, QoE), to underline the fact that, in this case, we want to center the analysis on the user. In this context, we have a global project called PSQA, which stands for Pseudo-Subjective Quality Assessment, and which refers to a technology we have developed allowing to automatically measure this PQ.

Another special case to which we devote research efforts in the team is the analysis of qualitative properties related to interoperability assessment. This refers to the act of determining if end-to-end functionality between at least two communicating systems is as required by the base standards for those systems. Conformance is the act of determining to what extent a single component conforms to the individual requirements of the standard it is based on. Our purpose is to provide such a formal framework (methods, algorithms and tools) for interoperability assessment, in order to help in obtaining efficient interoperability test suites for new generation networks, mainly around IPv6-related protocols. The interoperability test suites generation is based on specifications (standards and/or RFCs) of network components and protocols to be tested.

## 3.3. Stochastic modeling

The scientific foundations of our modeling activities are composed of stochastic processes theory and, in particular, Markov processes, queuing theory, stochastic graphs theory, etc. The objectives are either to develop numerical solutions, or analytical ones, or possibly discrete event simulation or Monte Carlo (and Quasi-Monte Carlo) techniques. We are always interested in model evaluation techniques for dependability and performability analysis, both in static (network reliability) and dynamic contexts (depending on the fact that time plays an explicit role in the analysis or not). We look at systems from the classical so-called *call level*, leading to standard models (for instance, queues or networks of queues) and also at the *burst level*, leading to *fluid models*.

In recent years, our work on the design of the topologies of WANs led us to explore optimization techniques, in particular in the case of very large optimization problems, usually formulated in terms of graphs. The associated methods we are interested in are composed of simulated annealing, genetic algorithms, TABU search, etc. For the time being, we have obtained our best results with GRASP techniques.

Network pricing is a good example of a multi-disciplinary research activity half-way between applied mathematics, economy and networking, centered on stochastic modeling issues. Indeed, the Internet is facing a tremendous increase of its traffic volume. As a consequence, real users complain that large data transfers take too long, without any possibility to improve this by themselves (by paying more, for instance). A possible solution to cope with congestion is to increase the link capacities; however, many authors consider that this is not a viable solution as the network must respond to an increasing demand (and experience has shown that demand of bandwidth has always been ahead of supply), especially now that the Internet is becoming a commercial network. Furthermore, incentives for a fair utilization between customers are not included in the current Internet. For these reasons, it has been suggested that the current flat-rate fees, where customers pay a subscription and obtain an unlimited usage, should be replaced by usage-based fees. Besides, the future Internet will carry heterogeneous flows such as video, voice, email, web, file transfers and remote login among others. Each of these applications requires a different level of QoS: for example, video needs very small delays and packet losses, voice requires small delays but can afford some packet losses, email can afford delay (within a given bound) while file transfer needs a good average throughput and remote login requires small round-trip times. Some pricing incentives should exist so that each user does not always choose the best QoS for her application and so that the final result is a fair utilization of the bandwidth. On the other hand, we need to be aware of the trade-off between engineering efficiency and economic efficiency; for example, traffic measurements can help in improving the management of the network but is a costly option. These are some of the various aspects often present in the pricing problems we address in our work. More recently, we have switched to the more general field of network economics, dealing with the economic behavior of users, service providers and content providers, as well as their relations.

**DYOGENE Project-Team**

# 3. Research Program

## 3.1. Network Calculus

Network calculus [53] is a theory for obtaining deterministic upper bounds in networks that has been developed by R. Cruz [41], [42]. From the modelling point of view, it is an algebra for computing and propagating constraints given in terms of envelopes. A flow is represented by its cumulative function $R(t)$ (that is, the amount of data sent by the flow up to time $t$). A constraint on a flow is expressed by an arrival curve $\alpha(t)$ that gives an upper bound for the amount of data that can be sent during any interval of length $t$. Flows cross service elements that offer guarantees on the service. A constraint on a service is a service curve $\beta(t)$ that is used to compute the amount of data that can be served during an interval of length t. It is also possible to define in the same way minimal arrival curves and maximum service curves. Then such constraints envelop the processes and the services. Network calculus enables the following operations:
• computing the exact output cumulative function or at least bounding functions;
• computing output constraints for a flow (like an output arrival curve);
• computing the remaining service curve (that is, the service that of not used by the flows crossing a server);
• composing several servers in tandem;
• giving upper bounds on the worst-case delay and backlog (bounds are tight for a single server or a single flow).
The operations used for this are an adaptation of filtering theory to $(\min, +)$: $(\min, +)$ convolution and deconvolution, sub-additive closure.

We investigate the complexity of computing exact worst-case performance bounds in network calculus and to develop algorithms that present a good trade off between algorithmic efficiency and accuracy of the bounds.

## 3.2. Perfect Simulation

Simulation approaches can be used to efficiently estimate the stationary behavior of Markov chains by providing independent samples distributed according to their stationary distribution, even when it is impossible to compute this distribution numerically.

The classical Markov Chain Monte Carlo simulation techniques suffer from two main problems:
• The convergence to the stationary distribution can be very slow, and it is in general difficult to estimate;
• Even if one has an effective convergence criterion, the sample obtained after any finite number of iterations is biased.

To overcome these issues, Propp and Wilson [56] have introduced a perfect sampling algorithm (PSA) that has later been extended and applied in various contexts, including statistical physics [47], stochastic geometry [52], theoretical computer science [33], and communications networks [30], [46] (see also the bibliography at http://dimacs.rutgers.edu/~dbwilson/exact.html/ annotated by David B. Wilson.

Perfect sampling uses coupling arguments to give an unbiased sample from the stationary distribution of an ergodic Markov chain on a finite state space $\mathcal{X}$. Assume the chain is given by an update function $\Phi$ and an i.i.d. sequence of innovations $(U_n)_{n\in\mathbb{Z}}$, so that

$$X_{n+1} = \Phi(X_n, U_{n+1}). \tag{1}$$

The algorithm is based on a backward coupling scheme: it computes the trajectories from all $x \in \mathcal{X}$ at some time in the past $t = -T$ until time $t = 0$, using the same innovations. If the final state is the same for all trajectories (i.e. $|\{\Phi(x, U_{-T+1}, ..., U_0) : x \in \mathcal{X}\}| = 1$, where $\Phi(x, U_{-T+1}, ..., U_0) := \Phi(\Phi(x, U_{-T+1}), U_{-T+2}, ..., U_0)$ is defined by induction on $T$), then we say that the chain has globally coupled and the final state has the stationary distribution of the Markov chain. Otherwise, the simulations are started further in the past.

Any ergodic Markov chain on a finite state space has a representation of type (1 ) that couples in finite time with probability 1, so Propp and Wilson's PSA gives a "perfect" algorithm in the sense that it provides an *unbiased* sample in *finite time*. Furthermore, the stopping criterion is given by the coupling from the past scheme, and knowing the explicit bounds on the coupling time is not needed for the validity of the algorithm.

However, from the computational side, PSA is efficient only under some monotonicity assumptions that allow reducing the number of trajectories considered in the coupling from the past procedure only to extremal initial conditions. Our goal is to propose new algorithms solving this issue by exploiting semantic and geometric properties of the event space and the state space.

## 3.3. Stochastic Geometry

Stochastic geometry [40] is a rich branch of applied probability which allows one to quantify random phenomena on the plane or in higher dimension. It is intrinsically related to the theory of point processes. Initially its development was stimulated by applications to biology, astronomy and material sciences. Nowadays it is also widely used in image analysis. It provides a way of estimating and computing "spatial averages". A typical example, with obvious communication implications, is the so called Boolean model, which is defined as the union of discs with random radii (communication ranges) centered at the points of a Poisson point process (user locations) of the Euclidean plane (e.g., a city). A first typical question is that of the prediction of the fraction of the plane which is covered by this union (statistics of coverage). A second one is whether this union has an infinite component or not (connectivity). Further classical models include shot noise processes and random tessellations. Our research consists of analyzing these models with the aim of better understanding wireless communication networks in order to predict and control various network performance metrics. The models require using techniques from stochastic geometry and related fields including point processes, spatial statistics, geometric probability, percolation theory.

F. Baccelli, B. Blaszczyszyn in collaboration with M. Karray (Orange Labs) are preparing a new book focusing on the mathematical tools at the basis of stochastic geometry. The book will cover the main mathematical foundations of the field, namely the theory of point processes and random measures as well as the theory of random closed sets. The basis will be the graduate classes and the research courses taught by the authors at a variety of places worldwide.

The collaboration of F. Baccelli with V. Anantharam (UC Berkeley) continues in new directions on high dimensional stochastic geometry, primarily in relation with Information Theory, cf. Section 7.23 .

The collaboration of B. Blaszczyszyn with D. Yogeshwaran (Indian Statistical Institute) and Y. Yukich (Lehigh University) led to the development of the limit theory for geometric statistics on general input processes, cf. Section 7.22 .

## 3.4. Information Theory and Wireless Networks

Classical models of stochastic geometry (SG) are not sufficient for analyzing wireless networks as they ignore the specific nature of radio channels.

Consider a wireless communication network made of a collection of nodes which in turn can be transmitters or receivers. At a given time, some subset of this collection of nodes simultaneously transmit, each toward its own receiver. Each transmitter–receiver pair in this snapshot requires its own wireless link. For each such wireless link, the power of the signal received from the link transmitter is jammed by the powers of the signals received from the other transmitters. Even in the simplest model where the power radiated from a

point decays in some isotropic way with Euclidean distance, the geometry of the location of nodes plays a key role within this setting since it determines the signal to interference and noise ratio (SINR) at the receiver of each such link and hence the possibility of establishing simultaneously this collection of links at a given bit rate, as shown by information theory (IT). In this definition, the interference seen by some receiver is the sum of the powers of the signals received from all transmitters excepting its own. The SINR field, which is of an essentially geometric nature, hence determines the connectivity and the capacity of the network in a broad sense. The essential point here is that the characteristics and even the feasibilities of the radio links that are simultaneously active are strongly interdependent and determined by the geometry. Our work is centered on the development of an IT-aware stochastic geometry addressing this interdependence. Dyogene members published in 2009 a two-volume book [1], [2] on Stochastic Geometry and Wireless Networks that became a reference publication in this domain.

In collaboration with Martin Haenggi (University of Notre Dame Notre Dame, IN, USA), Paul Keeler (Weierstrass Institute for Applied Analysis and Stochastics Berlin, Germany) and Sayandev Mukherjee (DOCOMO Innovations, Inc. Palo Alto, CA, USA), B. Blaszczyszyn is currently working on a book project that is intended to bridge a gap between academic and industrial approach to the design of next-generation cellular networks. In fact, simulation-only approach adopted by a majority of industry practitioners does not scale up with the increasing network complexity and analytical treatment is still yet not widely accepted in various bodies working out future standards specifications. The monograph is intended to bridge that gap, and make the methods, tools, approaches, and results of stochastic geometry available to a wide group of researchers (both in academia and in industry), systems engineers, and network designers. We expect that academic researchers and graduate students will appreciate that the book collects and organizes the most recent research results in a convenient way.

## 3.5. The Cavity Method for Network Algorithms

The cavity method combined with geometric networks concepts has recently led to spectacular progresses in digital communications through error-correcting codes. More than fifty years after Shannon's theorems, some coding schemes like turbo codes and low-density parity-check codes (LDPC) now approach the limits predicted by information theory. One of the main ingredients of these schemes is message-passing decoding strategies originally conceived by Gallager, which can be seen as direct applications of the cavity method on a random bipartite graph (with two types of nodes representing information symbols and parity check symbols, see [57]).

Modern coding theory is only one example of application of the cavity method. The concepts and techniques developed for its understanding have applications in theoretical computer science and a rich class of *complex systems*, in the field of networking, economics and social sciences. The cavity method can be used both for the analysis of randomized algorithms and for the study of random ensembles of computational problems representative real-world situations. In order to analyze the performance of algorithms, one generally defines a family of instances and endows it with a probability measure, in the same way as one defines a family of samples in the case of spin glasses or LDPC codes. The discovery that the hardest-to-solve instances, with all existing algorithms, lie close to a *phase transition* boundary has spurred a lot of interest. Theoretical physicists suggest that the reason is a structural one, namely a change in the geometry of the set of solutions related to the *replica symmetry breaking* in the cavity method. Phase transitions, which lie at the core of statistical physics, also play a key role in computer science [60], signal processing [44] and social sciences [49]. Their analysis is a major challenge, that may have a strong impact on the design of related algorithms.

We develop mathematical tools in the theory of discrete probabilities and theoretical computer science in order to contribute to a rigorous formalization of the cavity method, with applications to network algorithms, statistical inference, and at the interface between computer science and economics (EconCS).

## 3.6. Statistical Learning

Sparse graph structures are useful in a number of information processing tasks where the computational problem can be described as follows: infer the values of a large collection of random variables, given a set

of constraints or observations, that induce relations among them. Similar design ideas have been proposed in sensing and signal processing and have applications in coding [38], network measurements, group testing or multi-user detection. While the computational problem is generally hard, sparse graphical structures lead to low-complexity algorithms that are very effective in practice. We develop tools in order to contribute to a precise analysis of these algorithms and of their gap to optimal inference which remains a largely open problem.

A second line of activities concerns the design of protocols and algorithms enabling a transmitter to learn its environment (the statistical properties of the channel quality to the corresponding receiver, as well as their interfering neighbouring transmitters) so as to optimise their transmission strategies and to fairly and efficiently share radio resources. This second objective calls for the development and use of machine learning techniques (e.g. bandit optimisation).

EVA Project-Team

# 3. Research Program

## 3.1. Generalities

EVA inherits its expertise in designing algorithms and protocols from HiPERCOM2 (e.g. OLSR). EVA also inherit know-how in modeling, simulation, experimentation and standardization. Through this know-how and experience, the results obtained are both far-reaching and useful.

## 3.2. Physical Layer

We plan to study how advanced physical layers can be used in low-power wireless networks. For instance, collaborative techniques such as multiple antennas (e.g. the Massive MIMO technology) can improve communication efficiency. The idea is to use a massive network densification by drastically increasing the number of sensors in a given area in a Time Division Duplex (TDD) mode with time reversal. The first period allows the sensors to estimate the channel state and, after time reversal, the second period is to transmit the data sensed. Other techniques, such as interference cancellation, are also possible.

## 3.3. Wireless Access

Medium sharing in wireless systems has received substantial attention throughout the last decade. HiPERCOM2 has provided models to compare TDMA and CSMA. HiPERCOM2 has also studied how network nodes must be positioned to optimize the global throughput.

EVA will pursue modeling tasks to compare access protocols, including multi-carrier access, adaptive CSMA (particularly in VANETs), as well as directional and multiple antennas. There is a strong need for determinism in industrial networks. The EVA team will focus particularly on scheduled medium access in the context of deterministic industrial networks; this will involve optimizing the joint time slot and channel assignment. Distributed approaches will be considered, and the EVA team will determine their limits in terms of reliability, latency and throughput. Furthermore, adaptivity to application or environment changes will be taken into account.

## 3.4. Coexistence of Wireless Technologies

Wireless technologies such as cellular, low-power mesh networks, (Low-Power) WiFi, and Bluetooth (low-energy) can reasonably claim to fit the requirements of the IoT. Each, however, uses different trade-offs between reliability, energy consumption and throughput. The EVA team will study the limits of each technology, and will develop clear criteria to evaluate which technology is best suited to a particular set of constraints.

Coexistence between these different technologies (or different deployments of the same technology in a common radio space) is a valid point of concern.

The EVA team aims at studying such coexistence, and, where necessary, propose techniques to improve it. Where applicable, the techniques will be put forward for standardization. Multiple technologies can also function in a symbiotic way.

For example, to improve the quality of experience provided to end users, a wireless mesh network can transport sensor and actuator data in place of a cellular network, when and where cellular connectivity is poor.

The EVA team will study how and when different technologies can complement one another. A specific example of a collaborative approach is Cognitive Radio Sensor Networks (CRSN).

## 3.5. Energy-Efficiency and Determinism

Reducing the energy consumption of low-power wireless devices remains a challenging task. The overall energy budget of a system can be reduced by using less power-hungry chips, and significant research is being done in that direction. Nevertheless, power consumption is mostly influenced by the algorithms and protocols used in low-power wireless devices, since they influence the duty-cycle of the radio.

EVA will search for energy-efficient mechanisms in low-power wireless networks. One new requirement concerns the ability to predict energy consumption with a high degree of accuracy. Scheduled communication, such as the one used in the IEEE 802.15.4e TSCH (Time Slotted CHannel Hopping) standard, and by IETF 6TiSCH, allows for a very accurate prediction of the energy consumption of a chip. Power conservation will be a key issue in EVA.

To tackle this issue and match link-layer resources to application needs, EVA's 5-year research program around Energy-Efficiency and Determinism centers around 3 studies:

- Performance Bounds of a TSCH network. We propose to study a low-power wireless TSCH network as a Networked Control System (NCS), and use results from the NCS literature. A large number of publications on NCS, although dealing with wireless systems, consider wireless links to have perfect reliability, and do not consider packet loss. Results from these papers can not therefore be applied directly to TSCH networks. Instead of following a purely mathematical approach to model the network, we propose to use a non-conventional approach and build an empirical model of a TSCH network.
- Distributed Scheduling in TSCH networks. Distributed scheduling is attractive due to its scalability and reactivity, but might result in a sub-optimal schedule. We continue this research by designing a distributed solution based on control theory, and verify how this solution can satisfy service level agreements in a dynamic environment.

## 3.6. Network Deployment

Since sensor networks are very often built to monitor geographical areas, sensor deployment is a key issue. The deployment of the network must ensure full/partial, permanent/intermittent coverage and connectivity. This technical issue leads to geometrical problems which are unusual in the networking domain.

We can identify two scenarios. In the first one, sensors are deployed over a given area to guarantee full coverage and connectivity, while minimizing the number of sensor nodes. In the second one, a network is re-deployed to improve its performance, possibly by increasing the number of points of interest covered, and by ensuring connectivity. EVA will investigate these two scenarios, as well as centralized and distributed approaches. The work starts with simple 2D models and will be enriched to take into account more realistic environment: obstacles, walls, 3D, fading.

## 3.7. Data Gathering and Dissemination

A large number of WSN applications mostly do data gathering (a.k.a "convergecast"). These applications usually require small delays for the data to reach the gateway node, requiring time consistency across gathered data. This time consistency is usually achieved by a short gathering period.

In many real WSN deployments, the channel used by the WSN usually encounters perturbations such as jamming, external interferences or noise caused by external sources (e.g. a polluting source such as a radar) or other coexisting wireless networks (e.g. WiFi, Bluetooth). Commercial sensor nodes can communicate on multiple frequencies as specified in the IEEE 802.15.4 standard. This reality has given birth to the multichannel communication paradigm in WSNs.

Multichannel WSNs significantly expand the capability of single-channel WSNs by allowing parallel transmissions, and avoiding congestion on channels or performance degradation caused by interfering devices.

In EVA, we will focus on raw data convergecast in multichannel low-power wireless networks. In this context, we are interested in centralized/distributed algorithms that jointly optimize the channel and time slot assignment used in a data gathering frame. The limits in terms of reliability, latency and bandwidth will be evaluated. Adaptivity to additional traffic demands will be improved.

## 3.8. Self-Learning Networks

To adapt to varying conditions in the environment and application requirements, the EVA team will investigate self-learning networks. Machine learning approaches, based on experts and forecasters, will be investigated to predict the quality of the wireless links in a WSN. This allows the routing protocol to avoid using links exhibiting poor quality and to change the route before a link failure. Additional applications include where to place the aggregation function in data gathering. In a content delivery network (CDN), it is very useful to predict the popularity, expressed by the number of solicitations per day, of a multimedia content. The most popular contents are cached near the end-users to maximize the hit ratio of end-users' requests. Thus the satisfaction degree of end-users is maximized and the network overhead is minimized.

## 3.9. Security Trade-off in Constrained Wireless Networks

Ensuring security is a sine qua non condition for the widespread acceptance and adoption of the IoT, in particular in industrial and military applications. While the Public-Key Infrastructure (PKI) approach is ubiquitous on the traditional Internet, constraints in terms of embedded memory, communication bandwidth and computational power make translating PKI to constrained networks non-trivial.

Two related standardization working groups were created in 2013 to address this issue. DICE (DTLS In Constrained Environments) is defining a DTLS (Datagram Transport Layer Security) profile that is suitable for IoT applications, using the (Constrained Application Protocol) CoAP protocol. ACE is standardizing authentication and authorization mechanisms for constrained environments.

The issue is to find the best trade-off between a communication and computation overhead compatible with the limited capacity of sensor nodes and the level of protection required by the application.

**FUN Project-Team**

# 3. Research Program

## 3.1. Introduction

We will focus on wireless ubiquitous networks that rely on constrained devices, i.e. with limited resources in terms of storage and computing capacities. They can be sensors, small robots, RFID readers or tags. A wireless sensor retrieves a physical measure such as light. A wireless robot is a wireless sensor that in addition has the ability to move by itself in a controlled way. A drone is a robot with the ability to manoeuvre in 3D (in the air or in the water). RFID tags are passive items that embed a unique identifier for a place or an object allowing accurate traceability. They can communicate only in the vicinity of an RFID reader. An RFID reader can be seen as a special kind of sensor in the network which data is the one read on tags. These devices may run on batteries that are not envisaged to be changed or recharged. These networks may be composed of ten to thousands of such heterogeneous devices for which energy is a key issue.

Today, most of these networks are homogeneous, i.e. composed of only one kind of devices. They have mainly been studied in application and technology silos. Because of this, they are approaching fundamental limitations especially in terms of topology deployment, management and communications, while exploiting the complementarity of heterogeneous devices and communication technologies would enlarge their capacities and the set of applications. Finally, these networks must work efficiently even in dynamic and realistic situations, i.e. they must consider by design the different dynamic parameters and automatically self-adapt to their variations.

Our overall goal is represented by Figure 1 . We will investigate wireless ubiquitous IoT services for constrained devices by smartly combining **different frequency bands** and **different medium access and routing techniques**over **heterogeneous devices** in a **distributed** and **opportunistic** fashion. Our approach will always deal with **hardware constraints** and take care of **security** and **energy** issues to provide protocols that ride on **synergy** and **self-organization** between devices.

***The goal of the FUN project team is to provide these next generation networks with a set of innovative and distributed self-organizing cooperative protocols to raise them to a new level of scalability, autonomy, adaptability, manageability and performance. We aim to break these silos to exploit the full synergy between devices, making them cooperate in a single holistic network. We will consider them as networks of heterogeneous devices rather than a collection of heterogeneous networks.***

To realize the full potential of these ubiquitous networks, there is a need to provide them with a set of tools that allow them to *(i)* (self-)deploy, *(ii)* self-organize, *(iii)* discover and locate each other, resources and services and *(iv)* communicate. These tools will be the basics for enabling cooperation, co-existence and witnessing a global efficient behavior. The deployment of these mechanisms is challenging since it should be achieved in spite of several limitations. The main difficulties are to provide such protocols in a **secured** and **energy-efficient** fashion in spite of :

- dynamic topology changes due to various factors such as the unreliability of the wireless medium, the wireless interferences between devices, node mobility and energy saving mechanisms;
- hardware constraints in terms of CPU and memory capacities that limit the operations and data each node can perform/collect;
- lacks of interoperability between applicative, hardware and technological silos that may prevent from data exchange between different devices.

### *3.1.1. Objectives and methodology*

To reach our overall goal, we will pursue the two following objectives, similar to the ones we set for the previous evaluation period. These two objectives are othogonal and can be carried on jointly :

1. Providing realistic complete self-organizing tools *e.g. vertical perspective*.
2. Going to heterogeneous energy-efficient performing wireless networks *e.g. horizontal perspective*,

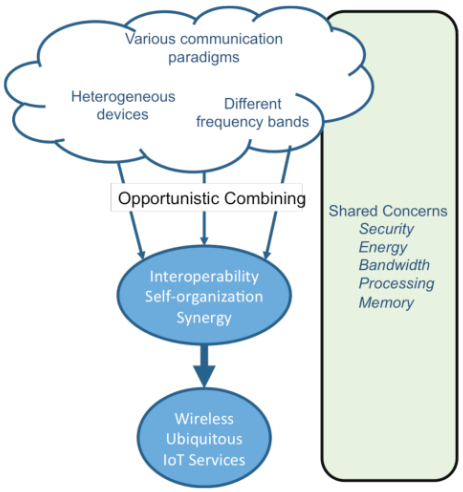


*Figure 1. FUN's overal goal.*

We give more details on these two objectives below. To achieve our main objectives, we will mainly apply the methodology depicted in Figure 2 combining both theoretical analysis and experimental validation. Mathematical tools will allow us to properly dimension a problem, formally define its limitations and needs to provide suitable protocols in response. Then, they will allow us to qualify the outcome solutions before we validate and stress them in real scenarios with regards to applications requirements. For this, we will realize proofs-of-concept with real scenarios and real devices. Differences between results and expectations will be analyzed in return in order to well understand them and integrate them by design for a better protocol self-adaptation capability.

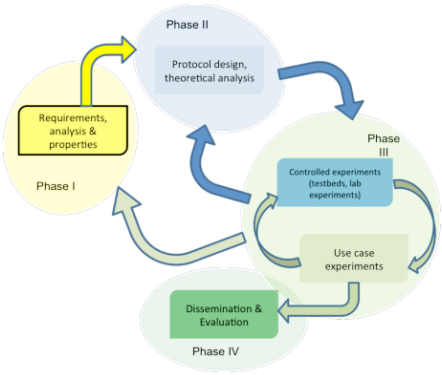


*Figure 2. Methodology to be applied in FUN.*

## 3.2. Vertical Perspective

As mentioned, future ubiquitous networks evolve in dynamic and unpredictable environments. Also, they can be used in a large scope of applications that have several expectations in terms of performance and different contextual limitations. In this heterogeneous context, IoT devices must support multiple applications and relay traffic with non-deterministic pattern.

To make our solutions practical and efficient in real conditions, we will adopt the dual approach both *top-down* and *bottom-up*. The *top-down* approach will ensure that we consider the application (such as throughput, delay, energy consumption, etc) and environmental limitations (such as deployment constraints, etc). The *bottom-up* approach will ensure that we take account of the physical and hardware characteristics such as memory, CPU, energy capacities but also physical interferences and obstacles. With this integrated perpective, we will be

in capacity to design **cross-layer** integrated protocols well adapted [39]. We will design jointly routing and MAC layers by taking dynamics occurring at the physical layer into account with a constant concern for energy and security. We will investigate new adaptive frequency hopping techniques combined with routing protocols [41], [50], [24]. Also, we will work on new scheduling techniques for TSCH (a MAC layer of IEEE 802.15.4e) that are able work under the above-mentionned assumptions and bring the robustness of TSCH to IoT scenarios. We will investigate the performance boundaries of TSCH in particular in terms of energy-efficiency of time synchronization [63], and will propose alternative approaches such as capture effect-based time synchronization in TSCH or opportunistic routing. Another technology we will consider is IEEE 802.15.4g, which provides communication ranges in the order of tens of kilometers. We will propose mechanisms to support scaling up to networks with a density of hundreds of nodes, at the MAC layer and above. We will also consider dual-technology networks where both long and short-range communication cooperate for increased robustness.

This vision will also allow us to integrate external factors by design in our protocols, in an opportunistic way. Yet, we will leverage on the occurrence of any of these phenomena rather than perceiving them as obstacles or limitations. As an example, we will rely on node undergone mobility to enhance routing performance as we have started to investigate in [74], [59]. On the same idea, when specific features are available like controlled mobility, we will exploit it to improve connectivity or coverage quality like in [46] [67].

## 3.3. Horizontal perspective

We aim at designing efficient tools for a plethora of wireless devices supporting highly heterogeneous technologies. We will thus investigate these networks from a horizontal perspective, e.g. by considering heterogeneity in low level communications layers.

Given the spectrum scarcity, they will probably need to coexist in the same frequency bands and sometimes for different purposes (RFID tag reading may use the same frequency bands as the wireless sensors). One important aspect to consider in this setting is how these different access technologies will interact with each other and what are the mechanisms needed to be put in place to guarantee that all services obtain the required share of resources when needed. This problem appears in different application domains, ranging from traffic offloading to unlicensed bands by cellular networks and the need to coexist with WiFi and radars, from a scenario in which multiple-purpose IoT clouds coexist in a city [75]. We will thus explore the dynamics of these interactions and devise ways to ensure smooth coexistence while considering the heterogeneity of the devices involved, the access mechanisms used as well as the requirements of the services provided.

To face the spectrum scarcity, we will also investigate new alternative communication paradigms such as phonon-based or light-based communications as we have initiated in [70], [71][16] and we will work on the coexistence of these technologies with traditional communication techniques, specifically by investigating efficient switching techniques from one communication technology to the other (they were most focused on the security aspects, to prevent jamming attacks). Resilience and reliability of the whole system will be the key factors to be taken into account [50], [24].

As a more prospective activity, we consider exploring software and communication security for IoT. This is challenging given that existing solutions do not address systems that are both constrained and networked [63]. Finally, in order to contribute to a better interoperability between all these technologies, we will continue to contribute to standardization bodies such as IETF and EPC Global.

**GANG Project-Team**

# 3. Research Program

## 3.1. Graph and Combinatorial Algorithms

We focus on two approaches for designing algorithms for large graphs: decomposing the graph and relying on simple graph traversals.

### *3.1.1. Graph Decompositions*

We study new decompositions schemes such as 2-join, skew partitions and others partition problems. These graph decompositions appeared in the structural graph theory and are the basis of some well-known theorems such as the Perfect Graph Theorem. For these decompositions there is a lack of efficient algorithms. We aim at designing algorithms working in $O(nm)$ since we think that this could be a lower bound for these decompositions.

### *3.1.2. Graph Search*

We more deeply study multi-sweep graph searches. In this domain a graph search only yields a total ordering of the vertices which can be used by the subsequent graph searches. This technique can be used on huge graphs and do not need extra memory. We already have obtained preliminary results in this direction and many well-known graph algorithms can be put in this framework. The idea behind this approach is that each sweep discovers some structure of the graph. At the end of the process either we have found the underlying structure (for example an interval representation for an interval graph) or an approximation of it (for example in hard discrete optimization problems). We envision applications to exact computations of centers in huge graphs, to underlying combinatorial optimization problems, but also to networks arising in biology.

### *3.1.3. Graph Exploration*

In the course of graph exploration, a mobile agent is expected to regularly visit all the nodes of an unknown network, trying to discover all its nodes as quickly as possible. Our research focuses on the design and analysis of agent-based algorithms for exploration-type problems, which operate efficiently in a dynamic network environment, and satisfy imposed constraints on local computational resources, performance, and resilience. Our recent contributions in this area concern the design of fast deterministic algorithms for teams of agents operating in parallel in a graph, with limited or no persistent state information available at nodes. We plan further studies to better understand the impact of memory constraints and of the availability of true randomness on efficiency of the graph exploration process.

## 3.2. Distributed Computing

The distributed community can be viewed as the union of two sub-communities. This is true even in our team. Even though they are not completely disjoint, they are disjoint enough not to leverage each others' results. At a high level, one is mostly interested in timing issues (clock drifts, link delays, crashes, etc.) while the other one is mostly interested in spatial issues (network structure, memory requirements, etc.). Indeed, one sub-community is mostly focusing on the combined impact of asynchronism and faults on distributed computation, while the other addresses the impact of network structural properties on distributed computation. Both communities address various forms of computational complexities, through the analysis of different concepts. This includes, e.g., failure detectors and wait-free hierarchy for the former community, and compact labeling schemes and computing with advice for the latter community. We have the ambitious project to achieve the reconciliation between the two communities by focusing on the same class of problems, the yes/no-problems, and establishing the scientific foundations for building up a consistent theory of computability and complexity for distributed computing. The main question addressed is therefore: is the absence of globally coherent computational complexity theories covering more than fragments of distributed computing, inherent

to the field? One issue is obviously the types of problems located at the core of distributed computing. Tasks like consensus, leader election, and broadcasting are of very different nature. They are not *yes-no* problems, neither are they minimization problems. Coloring and Minimal Spanning Tree are optimization problems but we are often more interested in constructing an optimal solution than in verifying the correctness of a given solution. Still, it makes full sense to analyze the *yes-no* problems corresponding to checking the validity of the output of tasks. Another issue is the power of individual computation. The FLP impossibility result as well as Linial's lower bound hold independently from the individual computational power of the involved computing entities. For instance, the individual power of solving NP-hard problems in constant time would not help overcoming these limits which are inherent to the fact that computation is distributed. A third issue is the abundance of models for distributed computing frameworks, from shared memory to message passing, spanning all kinds of specific network structures (complete graphs, unit-disk graphs, etc.) and or timing constraints (from complete synchronism to full asynchronism). There are however models, typically the wait-free model and the LOCAL model, which, though they do not claim to reflect accurately real distributed computing systems, enable focusing on some core issues. Our research program is ongoing to carry many important notions of Distributed Computing into a *standard* computational complexity.

## 3.3. Network Algorithms and Analysis

Based on our scientific foundation on both graph algorithms and distributed algorithms, we plan to analyze the behavior of various networks such as future Internet, social networks, overlay networks resulting from distributed applications or online social networks.

### *3.3.1. Information Dissemination*

One of the key aspects of networks resides in the dissemination of information among the nodes. We aim at analyzing various procedures of information propagation from dedicated algorithms to simple distributed schemes such as flooding. We also consider various models, where noise can alter information as it propagates or where memory of nodes is limited for example.

### *3.3.2. Routing Paradigms*

We try to explore new routing paradigms such as greedy routing in social networks for example. We are also interested in content centric networking where routing is based on content name rather than content address. One of our target is multiple path routing: how to design forwarding tables providing multiple disjoint paths to a destination?

### *3.3.3. Beyond Peer-to-Peer*

Based on our past experience of peer-to-peer application design, we would like to broaden the spectrum of distributed applications where new efficient algorithms and analysis can be performed. We especially target online social networks if we see them as collaborative tools for exchanging information. A basic question resides in making the right connections for gathering filtered and accurate information with sufficient coverage.

### *3.3.4. SAT and Forwarding Information Verification*

As forwarding tables of networks grow and are sometimes manually modified, the problem of verifying forwarding information becomes critical and has recently gained in interest. Some problems that arise in network verification such as loop detection for example, may be naturally encoded as Boolean Satisfiability problems. Beside the theoretical interest of this encoding in complexity proofs, it has also a practical value for solving these problems by taking advantage of the many efficient Satisfiability testing solvers. Indeed, SAT solvers have proved to be very efficient in solving problems coming from various areas (Circuit Verification, Dependency and Conflicts in Software distributions...) and encoded in Conjunctive Normal Form. To test an approach using SAT solvers in network verification, one need to collect data sets from real network and to develop good models for generating realistic networks. The technique of encoding and the solvers themselves need to be adapted to this kind of problems. All this represent a rich experimental field of future research.

### *3.3.5. Network Analysis*

Finally, we are interested in analyzing the structural properties of practical networks. This can include diameter computation or ranking of nodes. As we mostly consider large networks, we are often interested in efficient heuristics. Ideally, we target heuristics that give exact answer although fast computation time is not guaranteed for all networks. We already have designed such heuristics for diameter computation; understanding the structural properties that enable fast computation time in practice is still an open question.

**INFINE Project-Team**

# 3. Research Program

## 3.1. Online Social Networks (OSN)

Large-scale online social networks such as Twitter or FaceBook provide a powerful means of selecting information. They rely on "social filtering", whereby pieces of information are collectively evaluated and sorted by users. This gives rise to information cascades when one item reaches a large population after spreading much like an epidemics from user to user in a viral manner. Nevertheless, such OSNs expose their users to a large amount of content of no interest to them, a sign of poor "precision" according to the terminology of information retrieval. At the same time, many more relevant content items never reach those users most interested in them. In other words, OSNs also suffer from poor "recall" performance.

This leads to a first challenge: *what determines the optimal trade-off between precision and recall in OSNs? And what mechanisms should be deployed in order to approach such an optimal trade-off?* We intend to study this question at a theoretical level, by elaborating models and analyses of social filtering, and to validate the resulting hypotheses and designs through experimentation and processing of data traces. More specifically, we envision to reach this general objective by solving the following problems.

### *3.1.1. Community Detection*

Identification of implicit communities of like-minded users and contact recommendation for helping users "rewire" the information network for better performance. Potential schemes may include variants of spectral clustering and belief propagation-style message passing. Limitations / relative merits of candidate schemes, their robustness to noise in the input data, will be investigated.

### *3.1.2. Incentivization*

Design of incentive mechanisms to limit the impact of users' selfishness on system behavior: efficiency should be maintained even when users are gaming the system to try and increase their estimated expertise. By offering rewards to users on the basis of their involvement in filtering and propagation of content, one might encourage them to adjust their action and contribute to increase the overall efficiency of the OSN as a content access platform.

One promising direction will be to leverage the general class of Vickrey-Clarke-Groves incentive-compatible mechanisms of economic theory to design so-called marginal utility reward mechanisms for OSN users.

### *3.1.3. Social Recommendation and Privacy*

So far we have only alluded to the potential benefits of OSNs in terms of better information access. We now turn to the risks they create. Privacy breaches constitute the greatest of these risks: OSN users disclose a wealth of personal information and thereby expose themselves to discrimination by potential employers, insurers, lenders, government agencies...Such privacy concerns are not specific to OSNs: internauts' online activity is discretely tracked by companies such as Bluekai, and subsequently monetized to advertisers seeking better ad targeting. While disclosure of personal data creates a privacy risk, on the other hand it fuels personalized services and thereby potentially benefits everyone.

One line of research will be to focus on the specific application scenario of content categorization, and to characterize analytically the trade-off between user privacy protection (captured by differential privacy), accuracy of content categorization, and sample complexity (measured in number of probed users).

## 3.2. Traffic and Resource Management

Despite the massive increases in transmission capacity of the last few years, one has every reason to believe that networks will remain durably congested, driven among other factors by the steadily increasing demand for video content, the proliferation of smart devices (i.e., smartphones or laptops with mobile data cards), and the forecasted additional traffic due to machine-to-machine (M2M) communications. Despite this rapid traffic growth, there is still a rather limited understanding of the features protocols have to support, the characteristics of the traffic being carried and the context where it is generated. There is thus a strong need for smart protocols that transport requested information at the cheapest possible cost on the network as well as provide good quality of service to network subscribers. One particularly new aspect of up-and-coming networks is that networks are now used to not only (i) access information, but also (ii) distributively process information, en-route.

We intend to study these issues at the theoretical and protocol design levels, by elaborating models and analysis of content demands and/or mobility of network subscribers. The resulting hypothesis and designs will be validated through experimentation, simulation, or data trace processing. It is also worth mentioning the provided solutions may bring benefits to different entities in the network: to content owners (if applied at the core of Internet) or to subscribers or network operators (if applied at the edge of the Internet).

### *3.2.1. At the Internet Core*

One important optimization variable consists in content replication: users can access the closest replica of the content they are interested in. Thus the memory resource can be used to create more replicas and reduce the usage of the bandwidth resource. Another interesting arbitrage between resources arises because content is no longer static but rather dynamic. Here are two simple examples: i) a video could be encoded at several resolutions. There is then a choice between pre-recording all possible resolutions, or alternatively synthesizing a lower-resolution version on the fly from a higher resolution version when a request arises. ii) A user requests the result of a calculation, say the average temperature in a building; this can either be kept in memory, or recomputed each time such a query arises. Optimizing the joint use of all three resources, namely bandwidth, memory, computation, is a complex task. Content Delivery Network companies such as Akamai or Limelight have worked on the memory/bandwidth trade-off for some years, but as we will explain more can be done on this. On the other hand optimizing the memory/computation trade-off has received far less attention. We aim to characterize the best possible content replication strategies by leveraging fine-grained prediction of i) users' future requests, and ii) wireless channels' future bandwidth fluctuations. In the past these two determining inputs have only been considered at a coarse-grained, aggregate level. It is important to assess how much bandwidth saving can be had by conducting finer-grained prediction. We are developing light-weight protocols for conducting these predictions and automatically instantiating the corresponding optimal replication policies. We are also investigating generic protocols for automatically trading replication for computation, focusing initially on the above video transcoding scenario.

### *3.2.2. At the Internet Edge*

Cellular and wireless data networks are increasingly relied upon to provide users with Internet access on devices such as smartphones, laptops or tablets. In particular, the proliferation of handheld devices equipped with multiple advanced capabilities (e.g., significant CPU and memory capacities, cameras, voice to text, text to voice, GPS, sensors, wireless communication) has catalyzed a fundamental change in the way people are connected, communicate, generate and exchange data. In this evolving network environment, users' social relations, opportunistic resource availability, and proximity between users' devices are significantly shaping the use and design of future networking protocols.

One consequence of these changes is that mobile data traffic has recently experienced a staggering growth in volume: Cisco has recently foreseen that the mobile data traffic will increase 18-fold within 2016, in front of a mere 9-fold increase in connection speeds. Hence, one can observe today that the inherently centralized and terminal-centric communication paradigm of currently deployed cellular networks cannot cope with the increased traffic demand generated by smartphone users. This mismatch is likely to last because (1) forecasted

mobile data traffic demand outgrows the capabilities of planned cellular technological advances such as 4G or LTE, and (2) there is strong skepticism about possible further improvements brought by 5G technology.

Congestion at the Internet's edge is thus here to stay. Solutions to this problem relates to: densify the infrastructure, opportunistically forward data among neighbors wireless devices, to offload data to alternate networks, or to bring content from the Internet closer to the subscribers. Our recent work on leveraging user mobility patterns, contact and inter-contact patterns, or content demand patterns constitute a starting point to these challenges. The projected increase of mobile data traffic demand pushes towards additional complementary offloading methods. Novel mechanisms are thus needed, which must fit both the new context that Internet users experience now, and their forecasted demands. In this realm, we will focus on new approaches leveraging ultra-distributed, user-centric approaches over IP.

## 3.3. Spontaneous Wireless Networks (SWN) and Internet of Things (IoT)

The unavailability of end-to-end connectivity in emergent wireless mobile networks is extremely disruptive for IP protocols. In fact, even in simpler cases of spontaneous wireless networks where end-to-end connectivity exists, such networks are still disruptive for the standard IP protocol stack, as many protocols rely on atomic link-local services (such as link-local multicast/broadcast), while these services are inherently unavailable in such networks due to their opportunistic, wireless multi hop nature. In this domain, we will aim to characterize the achievable performance in such IP-disruptive networks and to actively contribute to the design of new, deployable IP protocols that can tolerate these disruptions, while performing well enough compared to what is achievable and remaining interoperable with the rest of the Internet.

Spontaneous wireless networking is also a key aspect of the Internet of Things (IoT). The IoT is indeed expected to massively use this networking paradigm to gradually connect billions of new devices to the Internet, and drastically increase communication without human source or destination – to the point where the amount of such communications will dwarf communications involving humans. Large scale user environment automation require communication protocols optimized to efficiently leverage the heterogeneous and unreliable wireless vicinity (the scope of which may vary according to the application). In fact, extreme constraints in terms of cost, CPU, battery and memory capacities are typically experienced on a substantial fraction of IoT devices. We expect that such constraints will not vanish any time soon for two reasons. On one hand the progress made over the last decade concerning the cost/performance ratio for such small devices is quite disappointing. On the other hand, the ultimate goal of the IoT is ubiquitous Internet connectivity between devices as tiny as dust particles. These constraints actually require to redesign not only the network protocol stack running on these devices, but also the software platform powering these machines. In this context, we will aim at contributing to the design of novel network protocols and software platforms optimized to fit these constraints while remaining compatible with legacy Internet.

### *3.3.1. Design & Development of Open Experimental IoT Platforms*

Manufacturers announce on a regular basis the availability of novel tiny devices, most of them featuring network interfaces: the Internet of Things (IoT) is already here, from the hardware perspective, and it is expected in the near future that we will see a massive increase of the number of muti-purpose smart objects (from tiny sensors in industrial automation to devices like smart watches and tablets). Thus, one of the challenges is to be able to test architectures, protocols and applications, in realistic conditions and at large scale.

One necessity for research in this domain is to establish and improve IoT hardware platforms and testbeds, that integrate representative scenarios (such as Smart Energy, Home Automation etc.) and follow the evolution of technology, including radio technologies, and associated experimentation tools. For that, we plan to build upon the IoT-LAB federated testbeds, that we have participated in designing and deploying recently. We plan to further develop IoT-LAB with more heterogeneous, up-to-date IoT hardware and radios that will provide a usable and realistic experimentation environment. The goal is to provide a tool that enables testing a validation of upcoming software platforms and network stacks targeting concrete IoT deployments.

In parallel, on the software side, IoT hardware available so far made it uneasy for developers to build apps that run across heterogeneous hardware platforms. For instance Linux does not scale down to small, energy-constrained devices, while microcontroller-based OS alternatives were so far rudimentary and yield a steep learning curve and lengthy development life-cycles because they do not support standard programming and debugging tools. As a result, another necessity for research in this domain is to allow the emergence of it more powerful, unifying IOT software platforms, to bridge this gap. For that, we plan to build upon RIOT, a new open source software platform which provides a portable, Linux-like API for heterogeneous IoT hardware. We plan to continue to develop the systems and network stacks aspects of RIOT, within the open source developer community currently emerging around RIOT, which we co-founded together with Freie Universitaet Berlin. The key challenge is to improve usability and add functionalities, while maintaining architectural consistency and a small enough memory footprint. The goal is to provide an IoT software platform that can be used like Linux is used for less constrained machines, both (i) in the context of research and/or teaching, as well as (ii) in industrial contexts. Of course, we plan to use it ourselves for our own experimental research activities in the domain of IoT e.g., as an API to implement novel network protocols running on IoT hardware, to be tested and validated on IoT-LAB testbeds.

### *3.3.2. Design & Standardization of Architectures and Efficient Protocols for Internet of Things*

As described before, and by definition, the Internet of Things will integrate not only a massive number of homogeneous devices (e.g., networks of wireless sensors), but also heterogeneous devices using various communication technologies. Most devices will be very constrained resources (memory resources, computational resources, energy). Communicating with (and amongst) such devices is a key challenge that we will focus on. The ability to communicate efficiently, to communicate reliably, or even just to be able to communicate at all, is non-trivial in many IoT scenarios: in this respect, we intend to develop innovative protocols, while following and contributing to standardization in this area. We will focus and base most of our work on standards developed in the context of the IETF, in working groups such as 6lo, CORE, LWIG etc., as well as IRTF research groups such as NWCRG on network coding and ICNRG on Information Centric Networking. We note however that this task goes far beyond protocol design: recently, radical rearchitecturing of the networks with new paradigms such as Information Centric Networking, ICN, (or even in wired networks, software-defined networks), have opened exciting new avenues. One of our direction of research will be to explore these content-centric approaches, and other novel architectures, in the context of IoT.

**MADYNES Project-Team**

# 3. Research Program

## 3.1. Evolutionary needs in network and service management

The foundation of the MADYNES research activity is the ever increasing need for automated monitoring and control within networked environments. This need is mainly due to the increasing dependency of both people and goods towards communication infrastructures as well as the growing demand towards services of higher quality. Because of its strategic importance and crucial requirements for interoperability, the management models were constructed in the context of strong standardization activities by many different organizations over the last 15 years. This has led to the design of most of the paradigms used in today's deployed approaches. These paradigms are the Manager/Agent interaction model, the Information Model paradigm and its container, together with a naming infrastructure called the Management Information Base. In addition to this structure, five functional areas known under Fault, Configuration, Accounting, Performance and Security are associated to these standards.

While these models were well suited for the specific application domains for which they were designed (telecommunication networks or dedicated protocol stacks), they all show the same limits. Especially they are unable:

1. to deal with any form of dynamicity in the managed environment,
2. to master the complexity, the operating mode and the heterogeneity of the emerging services,
3. to scale to new networks and service environments.

These three limits are observed in all five functional areas of the management domain (fault, configuration, accounting, performance and security) and represent the major challenges when it comes to enable effective automated management and control of devices, networks and services in the next decade.

MADYNES addresses these challenges by focusing on the design of management models that rely on inherently dynamic and evolving environments. The project is centered around two core activities. These activities are, as mentioned in the previous section, the design of an autonomous management framework and its application to three of the standard functional areas namely security, configuration and performance.

**MAESTRO Project-Team**

# 3. Research Program

## 3.1. Research Directions

MAESTRO's research directions belong to five main themes motivated by direct applications: network science, wireless networks, network engineering games, green networking and smart grids, content-oriented systems. These directions are very connected: network engineering games find applications in many networking fields, from wireless protocols to applications such as social networks. Green IT studies are often concerned with wireless networks, etc. The study of these applications often raises questions of methodological nature, less close to direct applications; these advances are reported in a separate section.

### *3.1.1. Network Science*

MAESTRO contributes to this new fast growing research subject. "Network Science" or "Complex Network Analysis" aims at understanding the structural properties and the dynamics of a variety of large-scale networks in telecommunications (e.g. the graph of autonomous systems, the Web graph), social science (e.g. community of interest, advertisement, reputation, recommendation systems), bibliometrics (e.g. citations, co-authors), biology (e.g. spread of an epidemic, protein-protein interactions), and physics. It has been observed that the complex networks encountered in these areas share common properties such as power law degree distribution, small average distances, community structure, etc. It also appears that many general questions/applications (e.g. community detection, epidemic spreading, search, anomaly detection) are common in various disciplines which study networks. In particular, we aim at understanding the evolution of complex networks with the help of game theoretical tools in connection with Network Engineering Games, as described below. We design efficient tools for measuring specific properties of large scale complex networks and their dynamics. More specifically, we work on the problem of distributed optimization in large networks where nodes cooperatively solve an optimization problem relying only on local information exchange.

### *3.1.2. Wireless Networks*

The amazing technological advances in wireless devices has led networks to become heterogeneous and very complex. Many research groups worldwide investigate performance evaluation of wireless technologies. MAESTRO's specificity relies on the use of a large variety of analytic tools from applied probability, control theory and distributed optimization to study and improve wireless networks functionalities. We investigate in particular problems of self-organization, channel selection and power control, the association problem and others.

### *3.1.3. Network Engineering Games*

The foundations of *Network Engineering Games* are currently being laid. These are games arising in telecommunications engineering at all the networking layers. This includes considerations from information and communications theory for dealing with the physical and link layers, along with cross layer approaches. MAESTRO's focus is on three areas: *routing games*, *evolutionary games* and *epidemic games*. In routing games we progress on the theory for costs that are not additive over links (such as packet losses or call blocking probabilities). We pursue their research in the stochastic extension of evolutionary game theory, namely the "anonymous sequential games" in which we study the total expected costs and the average cost. Within epidemic games they study epidemics that compete against each other. We apply this to social networks, considering in particular the coupling between various social networks (e.g. propagation strategies that combine Twitter, FaceBook and other social networks).

### 3.1.4. Green Networking and Smart Grids

The ICT (Information and Communications Technology) sector is becoming one of the main energy consumers worldwide. There is awareness that networks should have a reduced environmental footprint. Our objective is to have a systematically "green" approach when solving optimization problems. The energy cost and the environmental impact should be considered in optimization functions along with traditional performance metrics such as throughput, fairness or delay. We aim at contributing to the design and the analysis of future green networks, in particular those using renewable energy.

Researchers envision that future electricity distribution network will be "smart", with a large number of small generators (due to an extensive use of renewable energies) and of consumer devices able to adapt their energy needs to a time-varying offer. Generators and devices will be able to locally communicate through the electrical grid itself (or more traditional communication networks), in order to optimize production, transport and use of the energy. This is definitely a new application scenario for MAESTRO, to which we hope to be able to contribute with our expertise on analytic models and performance evaluation.

### 3.1.5. Content-Oriented Systems

We generally study problems related with the placement and the retrieval of data in communication networks.

We are particularly interested in In-network caching, a widely adopted technique to provide an efficient access to data or resources on a world-wide deployed system while ensuring scalability and availability. For instance, caches are integral components of the Domain Name System, the World Wide Web, Content Distribution Networks, or the recently proposed Information-Centric Network (ICN) architectures. We analyze network of caches, study their optimal placement in the network and optimize data placement in caches/servers.

We also study other aspects related to replication and placement of data: how much to replicate it and on which servers to place it? Finally, we study optimal ways of retrieving the data through prefetching.

### 3.1.6. Advances in Methodological Tools

MAESTRO has a methodological activity that aims at advancing the state of the art in the tools used for the general performance evaluation and control of systems. We contribute to such fields as perturbation analysis, Markov processes, queueing theory, control theory and game theory. Another objective is to enhance our activity on general-purpose modeling algorithms and software for controlled and uncontrolled stochastic systems.

## 3.2. Scientific Foundations

The main mathematical tools and formalisms used in MAESTRO include:

- theory of stochastic processes: Markov process, renewal process, branching process, point process, Palm measure, large deviations, mean-field approximation, fluid approximation;
- theory of dynamical discrete-event systems: queues, pathwise and stochastic comparisons, random matrix theory;
- theory of control and scheduling: dynamic programming, Markov decision process, game theory, deterministic and stochastic scheduling; stochastic approximation algorithms;
- theory of singular perturbations.

**MUSE Team**

# 3. Research Program

## 3.1. Active probing methods

We are developing methods that actively introduce probes in the network to discover properties of the connected devices and network segments. We are focusing in particular on methods to discover properties of home networks (connected devices and their types) and to distinguish if performance bottlenecks lie within the home network versus outside. Our goal is to develop adaptative methods that can leverage the collaboration of the set of available devices (including end-user devices and the home router, depending on which devices are running the measurement software).

## 3.2. Passive monitoring methods

This part our research develops methods that simply observe network traffic to infer the performance of networked applications and the location of performance bottlenecks, as well as to extract patterns of web content consumption. We are working on techniques to collect network traffic both at user's end-devices and at home routers. We also have access to network traffic traces collected on a campus network and on a large European broadband access provider.

## 3.3. Inferring user online experience

We are developing hybrid measurement methods that combine passive network measurement techniques to infer application performance with techniques from HCI to measure user perception. We will later use the resulting datasets to build models of user perception of network performance based only on data that we can obtain automatically from the user device or from user's traffic observed in the network.

## 3.4. Filtering real-time Web streams

The Web has become a large-scale real-time information system forcing us to revise both how to effectively assess relevance of information for a user and how to efficiently implement information retrieval and dissemination functionality. To increase information relevance, Real-time Web applications such as Twitter and Facebook, extend content and social-graph relevance scores with "real-time" user generated events (e.g. re-tweets, replies, likes). To accommodate high arrival rates of information items and user events we explore a publish/subscribe paradigm in which we index queries and update on the fly their results each time a new item and relevant events arrive. In this setting, we need to process continuous top-k text queries combining both static and dynamic scores. To the best of our knowledge, this is the first work addressing how non-predictable, dynamic scores can be handled in a continuous top-k query setting.

## 3.5. Flexible online drift detection

Monitoring streaming content is a challenging big data analytics problem, given that very large datasets are rarely (if ever) stationary. In several real world monitoring applications (e.g., newsgroup discussions, network connections, etc.) we need to detect significant change points in the underlying data distribution (e.g., frequency of words, sessions, etc.) and track the evolution of those changes over time. These change points, depending on the research community, are referred to as temporal evolution, non-stationarity, or concept drift and provide valuable insights on real world events (e.g. a discussion topic, an intrusion) to take a timely action. In our work, we adopt a query-based approach to drift detection and address the question of processing drift queries over very large datasets. To the best of our knowledge, our work is the first to formalize flexible drift queries on streaming datasets with varying change rates.

**RAP Project-Team**

# 3. Research Program

## 3.1. Scaling of Markov Processes

The growing complexity of communication networks makes it more difficult to apply classical mathematical methods. For a one/two-dimensional Markov process describing the evolution of some network, it is sometimes possible to write down the equilibrium equations and to solve them. The key idea to overcome these difficulties is to consider the system in limit regimes. This list of possible renormalization procedures is, of course, not exhaustive. The advantages of these methods lie in their flexibility to various situations and to the interesting theoretical problems they raised.

A fluid limit scaling is a particularly important means to scale a Markov process. It is related to the first order behavior of the process and, roughly speaking, amounts to a functional law of large numbers for the system considered.

A fluid limit keeps the main characteristics of the initial stochastic process while some second order stochastic fluctuations disappear. In “good” cases, a fluid limit is a deterministic function, obtained as the solution of some ordinary differential equation. As can be expected, the general situation is somewhat more complicated. These ideas of rescaling stochastic processes have emerged recently in the analysis of stochastic networks, to study their ergodicity properties in particular.

## 3.2. Design and Analysis of Algorithms

Data Structures, Stochastic Algorithms

The general goal of the research in this domain is of designing algorithms to analyze and control the traffic of communication networks. The team is currently involved in the design of algorithms to allocate bandwidth in optical networks and also to allocate resources in large distributed networks. See the corresponding sections below.

The team also pursues analysis of algorithms and data structures in the spirit of the former Algorithms team. The team is especially interested in the ubiquitous divide-and-conquer paradigm and its applications to the design of search trees, and stable collision resolution protocols.

## 3.3. Structure of random networks

This line of research aims at understanding the global structure of stochastic networks (connectivity, magnitude of distances, etc) via models of random graphs. It consists of two complementary foundational and applied aspects of connectivity.

RANDOM GRAPHS, STATISTICAL PHYSICS AND COMBINATORIAL OPTIMIZATION. The connectivity of usual models for networks based on random graphs models (Erdős–Rényi and random geometric graphs) may be tuned by adjusting the average degree. There is a *phase transition* as the average degree approaches one, a *giant* connected component containing a positive proportion of the nodes suddenly appears. The phase of practical interest is the *supercritical* one, when there is at least a giant component, while the theoretical interest lies at the *critical phase*, the break-point just before it appears.

At the critical point there is not yet a macroscopic component and the network consists of a large number of connected component at the mesoscopic scale. From a theoretical point of view, this phase is most interesting since the structure of the clusters there is expected (heuristically) to be *universal*. Understanding this phase and its universality is a great challenge that would impact the knowledge of phase transitions in all high-dimensional models of *statistical physics* and *combinatorial optimization*.

RANDOM GEOMETRIC GRAPHS AND WIRELESS NETWORKS. The level of connection of the network is of course crucial, but the *scalability* imposes that the underlying graph also be *sparse*: trade offs must be made, which required a fine evaluation of the costs/benefits. Various direct and indirect measures of connectivity are crucial to these choices: What is the size of the overwhelming connected component? When does complete connectivity occur? What is the order of magnitude of distances? Are paths to a target easy to find using only local information? Are there simple broadcasting algorithms? Can one put an end to viral infections? How much time for a random crawler to see most of the network?

NAVIGATION AND POINT LOCATION IN RANDOM MESHES. Other applications which are less directly related to networks include the design of improved navigation or point location algorithms in geometric meshes such as the Delaunay triangulation build from random point sets. There the graph model is essentially fixed, but the constraints it imposes raise a number of challenging problems. The aim is to prove performance guarantees for these algorithms which are used in most manipulations of the meshes.

**SOCRATE Project-Team**

# 3. Research Program

## 3.1. Research Axes

In order to keep young researchers in an environment close to their background, we have structured the team along the three research axes related to the three main scientific domains spanned by Socrate. However, we insist that a *major objective* of the Socrate team is to *motivate the collaborative research between these axes*, this point is specifically detailed in Section 3.5 . The first one is entitled "Flexible Radio Front-End" and will study new radio front-end research challenges brought up by the arrival of MIMO technologies, and reconfigurable front-ends. The second one, entitled "Multi-user communication", will study how to couple the self-adaptive and distributed signal processing algorithms to cope with the multi-scale dynamics found in cognitive radio systems. The last research axis, entitled "Software Radio Programming Models" is dedicated to embedded software issues related to programming the physical protocols layer on these software radio machines. Figure 3 illustrates the three regions of a transceiver corresponding to the three Socrate axes.

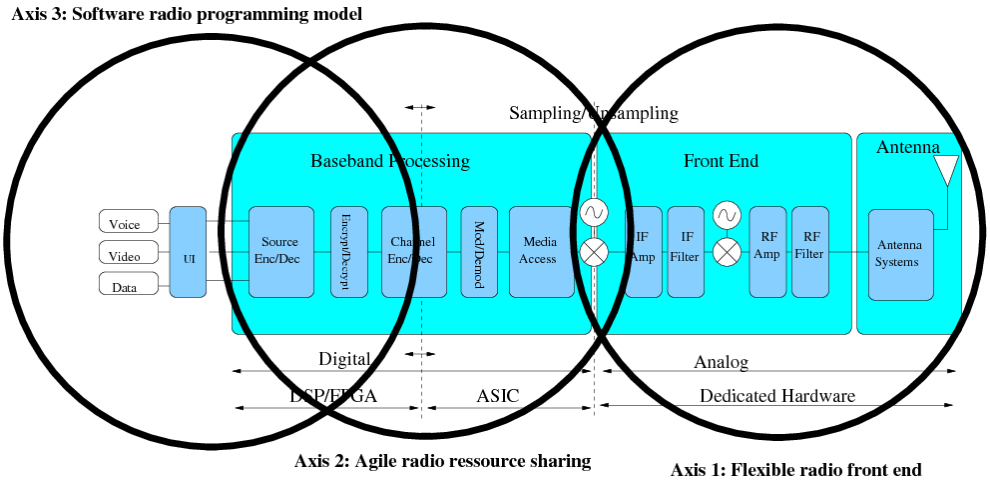


*Figure 3. Center of interest for each of the three Socrate research axes with respect to a generic software radio terminal.*

## 3.2. Flexible Radio Front-End

**Participants:** Guillaume Villemaud, Florin Hutu.

This axis mainly deals with the radio front-end of software radio terminals (right of Fig 3 ). In order to ensure a high flexibility in a global wireless network, each node is expected to offer as many degrees of freedom as possible. For instance, the choice of the most appropriate communication resource (frequency channel, spreading code, time slot,...), the interface standard or the type of antenna are possible degrees of freedom. The *multi-** paradigm denotes a highly flexible terminal composed of several antennas providing MIMO features to enhance the radio link quality, which is able to deal with several radio standards to offer interoperability and efficient relaying, and can provide multi-channel capability to optimize spectral reuse. On the other hand, increasing degrees of freedom can also increase the global energy consumption, therefore for energy-limited terminals a different approach has to be defined.

In this research axis, we expect to demonstrate optimization of flexible radio front-end by fine grain simulations, and also by the design of home made prototypes. Of course, studying all the components deeply would not be possible given the size of the team, we are currently not working in new technologies for DAC/ADC and power amplifiers which are currently studied by hardware oriented teams. The purpose of this axis is to build system level simulation taking into account the state of the art of each key component.

## 3.3. Multi-User Communications

**Participants:** Jean-Marie Gorce, Claire Goursaud, Nikolai Lebedev, Samir Perlaza, Leonardo Sampaio-Cardoso.

While the first and the third research axes deal with the optimization of the cognitive radio nodes themselves from system and programming point of view, an important complementary objective is to consider the radio nodes in their environments. Indeed, cognitive radio does not target the simple optimization of point to point transmissions, but the optimization of simultaneous concurrent transmissions. The tremendous development of new wireless applications and standards currently observed calls for a better management of the radio spectrum with opportunistic radio access, cooperative transmissions and interference management. This challenge has been identified as one of the most important issue for 5G to guarantee a better exploitation of the spectrum. In addition, mobile internet is going to support a new revolution that is the *tactile internet*, with real time interactions between the virtual and the real worlds, requiring new communication objectives to be met such as low latency end to end communications, distributed learning techniques, in-the-network computation, and many more. The future network will be heterogeneous in terms of technologies, type of data flows and QoS requirements. To address this revolution two work directions have naturally formed within the axis. The first direction concerns the theoretical study of fundamental limits in wireless networks. Introduced by Claude Shannon in the 50s and heavily developed up to today, Information Theory has provided a theoretical foundation to study the performance of wireless communications, not from a practical design view point, but using the statistical properties of wireless channels to establish the fundamental trade-offs in wireless communications. Beyond the classical *energy efficiency - spectral efficiency* tradeoff, information theory and its many derivations, i.e., network information theory, may also help to address additional questions such as determining the optimal rates under decentralized policies, asymptotic behavior when the density of nodes increases, latency controled communication with finite block-length theory, etc... In these cases, information theory is often associated to other theoretical tools such as game theory, stochastic geometry, control theory, graph theory and many others.

Our first research direction consists in evaluating specific mulit-user scenarios from a network information theory perspective, inspired by practical scenarios from various applicative frameworks (e.g. 5G, Wifi, sensor networks, IoT, etc...), and to establish fundamental limits for these scenarios. The second research direction is related to algorithmic and protocol design (PHY/MAC), applied to practical scenarios. Exploiting signal processing, linear algebra inspired models and distributed algorithms, we develop and evaluate various distributed algorithms allowing to improve many QoS metrics such as communication rates, reliability, stability, energy efficiency or computational complexity.

It is clear that both research directions are symbiotic with respect to each other, with the former providing theoretical bounds that serves as a reference to the performance of the algorithms created in the later. In the other way around, the later offers target scenarios for the former, through identifying fundamental problems that are interesting to be studied from the fundamental side. Our contributions of the year in these two directions are summarized further in the document.

## 3.4. Software Radio Programming Model

**Participants:** Tanguy Risset, Kevin Marquet, Lionel Morel, Guillaume Salagnac, Florent de Dinechin.

Finally the third research axis is concerned with software aspect of the software radio terminal (left of Fig 3 ). We have currently two actions in this axis, the first one concerns the programming issues in software defined radio devices, the second one focusses on low power devices: how can they be adapted to integrate some reconfigurability.

The expected contributions of Socrate in this research axis are :

- The design and implementation of a "middleware for SDR", probably based on a Virtual Machine.
- Prototype implementations of novel software radio systems, using chips from Leti and/or Lyrtech software radio boards.
- Development of a *smart node*: a low-power Software-Defined Radio node adapted to WSN applications.
- Methodology clues and programming tools to program all these prototypes.

## 3.5. Inter-Axes Collaboration

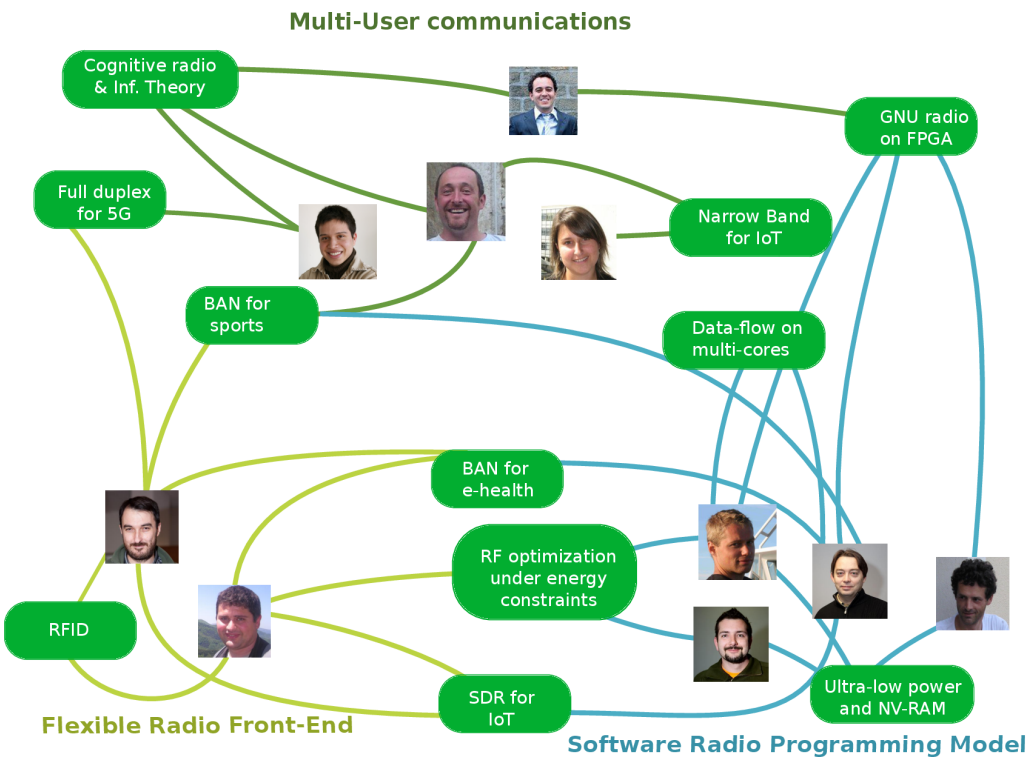


*Figure 4. Inter-Axis Collaboration in Socrate: we expect innovative results to come from this pluri-disciplinary research*

Innovative results come from collaborations between the three axes. To highlight the fact that this team structure does not limit the ability of inter-axes collaborations between Socrate members, we list below the *on-going* research actions that *already* involve actors from two or more axes, this is also represented on Fig 4 .

- *Optimizing network capacity of very large scale networks*. 2 Phds started in October/November 2011 with Guillaume Villemaud (axis 1) and Claire Goursaud (axis 2), respectively.
- *SDR for sensor networks*. A PhD started in 2012 in collaboration with FT R&D, involving people from axis 3 (Guillaume Salagnac, Tanguy Risset) and axis 1 (Guillaume Villemaud).
- *CorteXlab*. The 3 axes also collaborate on the design and the development of CorteXlab.

- *body area networks applications*. Axis 2 and axis 3 collaborate on the development of body area networks applications in the framework of the FUI Smacs project. Jean-Marie Gorce and Tanguy Risset co-advised Matthieu Lauzier.
- *Wiplan and NS3*. The MobiSim ADT involves Guillaume Villemaud (axis 1) and Jean-Marie Gorce (axis 2).
- *Resource allocation and architecture of low power multi-band front-end*. The EconHome project involves people from axis 2 (Jean-Marie Gorce,Nikolai Lebedev) and axis 1 (Florin Hutu). 1 Phd started in 2011.
- *Virtual machine for SDR*. In collaboration with CEA, a PhD started in October 2011, involving people from axis 3 (Tanguy Risset, Kevin Marquet) and Leti's engineers closer to axis 2.
- *Relay strategy for cognitive radio*. Guillaume Villemaud and Tanguy Risset were together advisers of Cedric Levy-Bencheton PhD Thesis (defense last June).

Finally, we insist on the fact that the *FIT project* will involve each member of Socrate and will provide many more opportunities to perform cross layer SDR experimentations. FIT is already federating all members of the Socrate team.

**URBANET Team**

# 3. Research Program

## 3.1. Capillary networks

The definition of Smart Cities is still constantly redefined and expanded so as to comprehensively describe the future of major urban areas. The Smart City concept mainly refers to granting efficiency and sustainability in densely populated metropolitan areas while enhancing citizens' life and protecting the environment. The Smart City vision can be primarily achieved by a clever integration of ICT in the urban tissue. Indeed, ICTs are enabling an evolution from the current duality between the "real world" and its digitalized counterpart to a continuum in which digital contents and applications are seamlessly interacting with classical infrastructures and services. The general philosophy of smart cities can also be seen as a paradigm shift combining the Internet of Things (IoT) and Machine-to-Machine (M2M) communication with a citizen-centric model, all together leveraging massive data collected by pervasive sensors, connected mobile or fixed devices, and social applications.

The fast expansion of urban digitalization yields new challenges that span from social issues to technical problems. Therefore, there is a significant joint effort by public authorities, academic research communities and industrial companies to understand and address these challenges. Within that context, the application layer, i.e., the novel services that ICT can bring to digital urban environments, have monopolized the attention. Lower-layer network architectures have gone instead quite overlooked. We believe that this might be a fatal error, since the communication network plays a critical role in supporting advanced services and ultimately in making the Smart City vision a reality. The UrbaNet project deals precisely with that aspect, and the study of network solutions for upcoming Smart Cities represents the core of our work.

Most network-related challenges along the road to real-world Smart Cities deal with efficient mobile data communication, both at the backbone and at the radio access levels. It is on the latter that the UrbaNet project is focused. More precisely, the scope of the project maps to that of capillary networks, an original concept we define next.

The capillary networking concept represents a unifying paradigm for wireless last-mile communication in smart cities. The term we use is reminiscent of the pervasive penetration of different technologies for wireless communication in future digital cities. Indeed, capillary networks represent the very last portion of the data distribution and collection network, bringing Internet connectivity to every endpoint of the urban tissue in the same exact way capillary blood vessels bring oxygen and collect carbon dioxide at tissues in the human body. Capillary networks inherit concepts from the self-configuring, autonomous, ad hoc networks so extensively studied in the past decade, but they do so in a holistic way. Specifically, this implies considering multiple technologies and applications at a time, and doing so by accounting for all the specificities of the urban environment.

## 3.2. Specific issues and new challenges of capillary networks

Capillary networks are not just a collection of independent wireless technologies that can be abstracted from the urban environment and/or studied separately. That approach has been in fact continued over the last decade, as technologies such as sensor, mesh, vehicular, opportunistic, and – generally speaking – M2M networks have been designed and evaluated in isolation and in presence of unrealistic mobility and physical layer, simplistic deployments, random traffic demands, impractical application use cases and non-existent business models. In addition, the physical context of the network has a significant impact on its performances and cannot be reduced to a simple random variable. Moreover, one of the main element of a network never appears in many studies: the user. To summarize, networks issues should be addressed from a user- and context-centric perspective.

Such abstractions and approximations were necessary for understanding the fundamentals of wireless network protocols. However, real world deployments have shown their limits. The finest protocols are often unreliable and hardly applicable to real contexts. That also partially explains the marginal impact of multi-hop wireless technologies on today's production market. Industrial solutions are mostly single-hop, complex to operate, and expensive to maintain.

In the UrbaNet project we consider the capillary network as an ensemble of strongly intertwined wireless networks that are expected to coexist and possibly co-operate in the context of arising digital cities. This has three major implications:

- Each technology contributing to the overall capillary network should not be studied apart. As a matter of fact, mobile devices integrate today a growing number of sensors (e.g., environment sensing, resource consumption metering, movement, health or pollution monitoring) and multiple radio interfaces (e.g., LTE, WiFi, ZigBee,. . . ), and this is becoming a trend also in the case of privately owned cars, public transport vehicles, commercial fleets, and even city bikes. Similarly, access network sites tend to implement heterogeneous communication technologies so as to limit capital expenses. Enabling smart-cities needs a dense sensing of its activities, which cannot be achieved without multi-service sensor networks. Moreover, all these devices are expected to inter-operate so as to make the communication more sustainable and reliable. Thus, the technologies that build up the capillary network shall be studied as a whole in the future.
- The capillary network paradigm necessarily accounts for actual urban mobility flows, city land-use layouts, metropolitan deployment constraints, and expected activity of the citizens. Often, these specificities do not arise from purely networking features, but relate to the study of city topologies and road layouts, social acceptability, transportation systems, energy management, or urban economics. Therefore, addressing capillary network scenarios cannot but rely on strong multidisciplinary interactions.
- Digital and smart cities are often characterized by arising M2M applications. However, a city is, before all, the gathering of citizens, who use digital services and mobile Internet for increasing their quality of life, empowerment, and entertainment opportunities. Some data flows should be gathered to, or distributed from, an information system. Some other should be disseminated to a geographically or time constrained perimeter. Future usage may induce peer-to-peer like traffics. Moreover these services are also an enabler of new usages of the urban environment. Solutions built within the capillary network paradigm have to manage this heterogeneity of traffic requirements and user behaviors.

By following these guidelines, the UrbaNet ambition is to go one step beyond traditional approaches discussed above. The capillary network paradigm for Smart Cities is tightly linked to the specificities of the metropolitan context and the citizens' activity. Our proposal is thus to re-think the way capillary network technologies are developed, considering a broader and more practical perspective.

## 3.3. Characterizing urban networks

Our first objective is to understand and model those properties of real-world urban environments that have an impact on the design, deployment and operation of capillary networks. It means to collect and analyze data from actual deployments and services, as well as testbeds experiments. These data have then to be correlated with urban characteristics, e.g. topography, density of population and activities. The objective is to deduce analytical models, simulations and traces of realistic scenarios that can be leveraged afterward. We structure the axis into three tasks that correspond to the three broad categories of networking aspects affected by the urban context.

- **Topological characteristics**. Nowadays, the way urban wireless network infrastructures are typically represented in the literature is dissatisfying. As an example, wireless links are mostly represented as symmetric, lossless channels whose signal quality depends continuously on the distance between the transmitter and the receiver. No need to say, real-world behaviors are very far from

these simplified representations. Another example, topologies are generally modeled according to deterministic (e.g., regular grids and lattices, or perfect hexagonal cell coverages) or stochastic (e.g., random uniform distributions over unbound surfaces) approaches. These make network problems mathematically tractable and simulations easier to set up, but are hardly representative of the layouts encountered in the real world. Employing simplistic models helps understanding some fundamental principles but risks to lead to unreliable results, both from the viewpoint of the network architecture design and from that of its performance evaluation. It is thus our speculation that the actual operations and the real-world topologies of infrastructured capillary networks are key to the successful deployment of these technologies, and, in this task, we aim at characterizing them. To that end, we leverage existing collaborations with device manufacturers (Alcatel-Lucent, HiKob) and operators (Orange), as well as collaboration such as the Sense City project and testbed experiments, in order to provide models that faithfully mimic the behavior of real world network devices. The goal is to understand the important features of the topologies, including, e.g., their overall connectivity level, spatial density, degree distribution, regularity, etc. Building on these results, we try to define network graph models that reproduce such major features and can be employed for the development and evaluation of capillary network solutions.

- **Mobilities**. We aim at understanding and modeling the mobile portion of capillary networks as well as the impact of the human mobility on the network usage. Our definition of "mobile portion" includes traditional mobile users as well as all communication-enabled devices that autonomously interact with Internet-based servers and among themselves. There have been efforts to collect real-world movement traces, to generate synthetic mobility dataset and to derive mobility models. However, real-world traces remain limited to small scenarios or circumstantial subsets of the users (e.g., cabs instead of the whole road traffic). Synthetic traces are instead limited by their scale and by their level of realism, still insufficient. Finally, even the most advanced models cannot but provide a rough representation of user mobility in urban areas, as they do not consider the street layout or the human activity patterns. In the end, although often deprecated, random or stochastic mobility models (e.g., random walks, exponential inter-arrivals and cell residence times) are still the common practice. We are well aware of the paramount importance of a faithful representation of device and user mobility within capillary networks and, in order to achieve it, we leverage a number of realistic sources, including Call Detail Records (CDR) collected by mobile operators, Open Data initiatives, real-world social network data, and experiments. We collect data and analyze it, so as to infer the critical properties of the underlying mobility patterns.
- **Data traffic patterns**. The characterization of capillary network usages means understanding and modeling when, where and how the wireless access provided by the diverse capillary network technologies is exploited by users and devices. In other words, we are interested in learning which applications are used at different geographical locations and day times, which urban phenomena generate network usage, and which kind of data traffic load they induce on the capillary network. Properly characterizing network usages is as critical as correctly modeling network topology and mobility. Indeed, the capillary networks being the link directly collecting the data from end devices, we cannot count on statistical smoothing which yields regular distributions. Unfortunately, the common practice is to consider, e.g., that each user or device generates a constant data traffic or follows on/off models, that the offered load is uniform over space and does not vary over time, that there is small difference between uplink and downlink behaviors, or that source/destination node pairs are randomly distributed in the network. We plan to go further on the specific scenarios we address, such as smart-parking, floating car data, tele-metering, road traffic management of pollution detection. To that end, we collect real-world data, explore it and derive properties useful to the accurate modeling of content consumption.

## 3.4. Autonomic networking protocols

While the capillary networks concept covers a large panel of technologies, network architectures, applications and services, common challenges remain, regardless the particular choice of a technology or architecture.

Our record of research on spontaneous and multi-hop networks let us think that autonomic networking appears as the main issue: the connectivity to Internet, to cyber-physical systems, to Information Systems should be transparent for the user, context-aware and location-aware. To address these challenges, a capillary network model is required. Unfortunately, very few specific models fit this task today. However, a number of important, specific capillary networks properties can already be inferred from recent experiments: distributed and localized topologies, very high node degree, dynamic network diameter, unstable / asymmetric / non-transitive radio links, concurrent topologies, heterogeneous capabilities, etc. These properties can already be acknowledged in the design of networking solutions, and they are particularly challenging for the functioning of the MAC layer and QoS support. Clearly, capillary networks provide new research opportunities with regard to networking protocols design.

- **Self-* protocols**. In this regard, self-configuration, self-organization and self-healing are some of the major concerns within the context of capillary networks. Solving such issues would allow spontaneous topologies to appear dynamically in order to provide a service depending of the location and the context, while also adapting to the interactions imposed by the urban environment. Moreover, these mechanisms have the capacity to alleviate the management of the network and the deployment engineering rules, and can provide efficient support to the network dynamics due to user mobility, environment modifications, etc. The designed protocols have to be able to react to traffic requests and local node densities. We address such self-adaptive protocols as a transversal solution to several scenarios, e.g. pollution monitoring, smart-services depending on human activities, vehicle to infrastructure communications, etc. In architectures where self-* mechanisms govern the protocol design, both robustness and energy are more than ever essential challenges at the network layer. Solutions such as energy-harvesting can significantly increase the network lifetime in this case, therefore we investigate their impact on the mechanisms at both MAC and network layers.
- **Quality of service issues**. The capillary networks paradigm implies a simultaneous deployment of multiple wireless technologies, and by different entities (industry, local community, citizens). This means that some applications and services can be provided concurrently by different parts of the capillary network, while others might require the cooperation of multiple parties. The notion of Service Level Agreement (SLA) for traffic differentiation, quality of service support (delay, reliability, etc.) is a requirement in these cases for scalability purposes and resource sharing. We contribute to a proper definition of this notion and the related network mechanisms in the settings of low power wireless devices. Because of the urban context, but also because of the wireless media itself, network connectivity is always temporary, while applications require a delivery ratio close to 100%. We investigate different techniques that can achieve this objective in an urban environment.
- **Data impact**. Capillary networks suffer from low capacity facing the increasing user request. In order to cope with network saturation, a promising strategy is to consider the nature of the transmitted data in the development of the protocols. Data aggregation and data gathering are two concepts with a major role to play in this context of limited capacity. In particular, combining local aggregation and measurement redundancy for improving on data reliability is a promising idea, which can also be important for energy saving purposes. Even if the data flow is well known and regular, e.g. temperature or humidity metering, developing aggregation schemes tailored to the constraints of the urban environment is a challenge we address within the UrbaNet team. Many urban applications generate data which has limited spatial and temporal perimeters of relevance, e.g. smart-parking applications, community information broadcasting, etc. When solely a spatial range of relevance is considered, the underlying mechanisms are denoted “geocasting”. We also address these spatio-temporal constraints, which combine geocasting approaches with real-time techniques.

## 3.5. Optimizing cellular network usage

The capacity of cellular networks, even those that are now being planned, does not seem able to cope with the increasing demands of data users. Moreover, new applications with high bandwidth requirements are also foreseen, for example in the intelligent transportation area, and an exponential growth in signaling traffic is

expected in order to enable this data growth. Cumulated with the lack of available new spectrum, this leads to an important challenge for mobile operators, who are looking at both licensed and unlicensed technologies for solutions. The usual strategy consists in a dramatic densification of micro-cells coverage, allowing both to minimize the transmission power of cellular networks as well as to increase the network capacity. However, this solution has obvious physical limits, which we work on determining, and we propose exploiting the capillarity of network interfaces as a complementary solution.

- **Green cellular network**. Increasing the density of micro-cells means multiplying the energy consumption issues. Indeed, the energy consumption of actual LTE eNodeBs and relays, whatever their state, idle, transmitting or receiving, is a major and growing part of the access network energy consumption. For a sustainable deployment of such micro-cell infrastructures and for a significative decrease of the overall energy consumption, an operator needs to be able to switch off cells when they are not absolutely needed. The densification of the cells induces the need for an autonomic control of the on/off state of cells. One solution in this sense can be to adapt the WSN mechanisms to the energy models of micro-cells and to the requirements of a cellular network. The main difficulty here is to be able to adapt and assess the proposed solutions in a realistic environment (in terms of radio propagation, deployment of the cells, user mobility and traffic dynamics).
- **Offloading**. Offloading the cellular infrastructure implies taking advantage of the wealth of connectivity provided by capillary networks instead of relying solely on 4G connectivity. Cellular operators usually possess an important ADSL or cable infrastructure for wired services, the development of femtocell solutions thus becomes very popular. However, while femtocells can be an excellent solution in zones with poor coverage, their extensive use in areas with a high density of mobile users leads to serious interference problems that are yet to be solved. Taking advantage of capillarity for offloading cellular data relies on using IEEE 802.11 Wi-Fi (or other similar technologies) access points or direct device-to-device communications. The ubiquity of Wi-Fi access in urban areas makes this solution particularly interesting, and many studies have focused on its potential. However, these studies fail to take into account the usually low quality of Wi-Fi connections in public areas, and they consider that a certain data rate can be sustained by the Wi-Fi network regardless of the number of contending nodes. In reality, most public Wi-Fi networks are optimized for connectivity, but not for capacity, and more research in this area is needed to correctly assess the potential of this technology. Direct opportunistic communication between mobile users can also be used to offload an important amount of data. This solution raises a number of major problems related to the role of social information and multi-hop communication in the achievable offload capacity. Moreover, in this case the business model is not yet clear, as operators would indeed offload traffic, but also lose revenue as direct ad-hoc communication would be difficult to charge and privacy issues may arise. However, combining hotspot connectivity and multi-hop communications is an appealing answer to broadcasting geo-localized informations efficiently.