INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team regal

Large-Scale Distributed Systems and Applications

Paris - Rocquencourt

Theme : Distributed Systems and Services

Activity Report

2010

Table of contents

Regal is a joint research group with CNRS and Université Paris 6 through the "Laboratoire d'Informatique de Paris 6", LIP6 *(UMR 7606).*

1. Team

Research Scientists

Gilles Muller [DR, HdR]
Marc Shapiro [DR, HdR]
Mesaac Makpangou [CR, HdR]

Faculty Members

Pierre Sens [Team Leader, Professor Université Paris 6, HdR]
Luciana Arantes [Associate Professor Université Paris 6]
Bertil Folliot [Professor Université Paris 6, HdR]
Maria Gradinariu Potop-Butucaru [Associate Professor Université Paris 6, HdR]
Olivier Marin [Associate Professor Université Paris 6]
Sébastien Monnet [Associate Professor Université Paris 6]
Franck Petit [Professor Université Paris 6, HdR]
Julien Sopena [Associate Professor Université Paris 6]
Gaël Thomas [Associate Professor Université Paris 6]

Technical Staff

Pierpaolo Cincilla [Junior Engineer]

PhD Students

Lamia Benmouffok [Microsoft research grant - Université Paris 6]
Mathieu Bouillaguet [Université Paris 6]
Raja Boujbel [Université Paris 6 since September]
Swan Dubois [Université Paris 6]
Ruijing Hu [Université Paris 6 since September]
Nicolas Hidalgo [Université Paris 6]
Anissa Lamani [Université Amiens]
Sergey Legtchenko [Université Paris 6]
Jean-Pierre Lozi [Université Paris 6 since September]
Corentin Méhat [Université Paris 6]
Nicolas Pierron [Université Paris 6 since September]
Thomas Preud'homme [Université Paris 6]
Erika Rosas [Université Paris 6]
Masoud Saeida Ardekani [Université Paris 6 since September]
Guthemberg Silvestre [Université Paris 6 since September]
Suman Saha [Université Paris 6 since September]
Pierre Sutra [Université Paris 6]
Mathieu Valéro [Université Paris 6]
Marek Zawirski [Université Paris 6 since September]

Post-Doctoral Fellows

Hyun-Gul Roh
Corentin Travers

Visiting Scientists

Julia Lawall
Fabiola Greve

Administrative Assistant

Hélène Milome [Secretary]

2. Overall Objectives

2.1. Overall Objectives

The main focus of the Regal team is research on large-scale distributed computing systems, and addresses the challenges of automated adminstration of highly dynamic networks, of fault tolerance, of consistency in large-scale distributed systems, of information sharing in collaborative groups, of dynamic content distribution, and of operating system adaptation. Regal is a joint research team between LIP6 and INRIA-Paris-Rocquencourt.

3. Scientific Foundations

3.1. Fondation 1

Scaling to large configurations is one of the major challenges addressed by the distributed system community lately. The basic idea is how to efficiently and transparently use and manage resources of millions of hosts spread over a large network. The problem is complex compared to classical distributed systems where the number of hosts is low (less than a thousand) and the inter-host links are fast and relatively reliable. In such "classical" distributed architectures, it is possible and reasonable to build a single image of the system so as to "easily" control resource allocation.

In large configurations, there is no possibility to establish a global view of the system. The underlying operating system has to make decisions (on resource allocation, scheduling ...) based only on partial and possibly wrongs view of the resources usage.

Scaling introduces the following problems:

- Failure: as the number of hosts increases, the probability of a host failure converges to one. [1] Compared to classical distributed systems, failures are more common and have to be efficiently processed.
- Asynchronous networks: on the Internet, message delays vary considerably and are unbounded.
- Impossibility of consensus:In such an asynchronous network with failures, consensus cannot be solved deterministically (the famous Fischer-Lynch-Patterson impossibility result of 1985).The system can only approximate, suspecting hosts that are not failed, or failing to suspect hosts that have failed. As a result, no host can form a consistent view of system state.
- Failure models: the classical view of distributed systems considers only crash and omission failures. In the context of large-scale, open networks, the failure model must be generalised to include stronger attacks. For instance, a host can be taken over ("zombie") and become malicious. Arbitrary faults, so-called Byzantine behaviours, are to be expected and must be tolerated.
- Managing distributed state: In contrast to a local-area network, establishing a global view of a large distributed system system is unfeasible. The operating system must make its decisions, regarding resource allocation or scheduling, based on partial and incomplete views of system state.

Three architectures in relation with the scaling problem have emerged during the last years:

Grid computing: Grid computing offers a model for solving massive computational problems using large numbers of computers arranged as clusters interconnected by a telecommunications infrastructure as internet, renater or VTHD.

[1] For instance if we consider a classical host MTBF (Mean Time Between Failure) equals to 13 days, in a middle scale system composed of only 10000 hosts, a failure will occur every 4 minutes.

If the number of involved hosts can be high (several thousands), the global environment is relatively controlled and users of such systems are usually considered safe and only submitted to host crash failures (typically, Byzantine failures are not considered).

Peer-to-peer overlay network: Generally, a peer-to-peer (or P2P) computer network is any network that does not rely on dedicated servers for communication but, instead, mostly uses direct connections between clients (peers). A pure peer-to-peer network does not have the notion of clients or servers, but only equal peer nodes that simultaneously function as both "clients" and "servers" with respect to the other nodes on the network.

This model of network arrangement differs from the client-server model where communication is usually relayed by the server. In a peer-to-peer network, any node is able to initiate or complete any supported transaction with any other node. Peer nodes may differ in local configuration, processing speed, network bandwidth, and storage capacity.

Different peer-to-peer networks have varying P2P overlays. In such systems, no assumption can be made on the behavior of the host and Byzantine behavior has to be considered.

Cloud computing Cloud computing offers a conceptually infinite amount of computing, storage and network resources for rent. From a user's perspective, cloud computing has many advantages, including low upfront investment and outsourcing of system administration. From the provider's perspective, cloud computing shares many characteristics with both grid computing (e.g., very large geographical and numeric scale, or service-oriented interfaces) and with P2P computing (e.g., self-administration). It also has some unique characteristics, such as systematic virtualisation of all resources, highly variable load, fast elastic adaptation, and quality-of-service objectives that are negociated with clients via SLAs.

Regal is interested in how to adapt distributed middleware to these large scale configurations. We target Grid and Peer-to-peer configurations. This objective is ambitious and covers a large spectrum. To reduce its spectrum, Regal focuses on fault tolerance, replication management, and dynamic adaptation.

We concentrate on the following research themes:

Data management: the goal is to be able to deploy and locate effectively data while maintening the required level of consistency between data replicas.

System monitoring and failure detection: we envisage a service providing the follow-up of distributed information. Here, the first difficulty is the management of a potentially enormous flow of information which leads to the design of dynamic filtering techniques. The second difficulty is the asynchronous aspect of the underlying network which introduces a strong uncertainty on the collected information.

Adaptive replication: we design parameterizable techniques of replication aiming to tolerate the faults and to reduce information access times. We focus on the runtime adaptation of the replication scheme by (1) automatically adjusting the internal parameters of the strategies and (2) by choosing the replication protocol more adapted to the current context.

The dynamic adaptation of application execution support: the adaptation is declined here to the level of the execution support (in either of the high level strategies). We thus study the problem of dynamic configuration at runtime of the low support layers.

4. Application Domains

4.1. Application Domains

As we already mentioned, we focus on two kinds of large scale environments: computational grids and peer-to-peer (P2P) systems. Although both environments have the same final objective of sharing large sets of resources, they initially emerged from different communities with different context assumptions and hence they have been designed differently. Grids provide support for a large number of services needed by scientific communities. They usually target thousands of hosts and hundreds of users. Peer-to-peer environments address millions of hosts with hundreds of thousands of simultaneous users but they offer limited and specialized functionalities (file sharing, parallel computation).

In peer-to-peer configurations we focus on the following applications:

- Internet services such as web caches or content distribution network (CDN) which aim at reducing the access time to data shared by many users,
- Data storage of mutable data. Data storage is a classical peer-to-peer application where users can share documents (audio and video) across the Internet. A challenge for the next generation of data sharing systems is to provide update management in order to develop large cooperative applications.
- multi-player games. The recent involvement of REGAL in the PLAY ALL project gives us the opportunity to consider distributed interactive video games. Theses applications are very interesting for us since they bring new constraints, most specifically on latency.

In Grid configurations we address resource management for two kinds of applications:

- Multi-agent applications which model complex cooperative behaviors.
- Application Service Provider (ASP) environments in cooperation with the DIET project of the GRAAL team.

Our third application domain is based on data sharing. Whereas most work on P2P applications focuses on write-once single-writer multiple-reader applications, we consider the (more demanding) applications that share mutable data in large-scale distributed settings. Some examples are co-operative engineering, collaborative authoring, or entreprise information libraries: for instance co-operative code development tools or decentralized wikis. Such applications involve users working from different locations and at different times, and for long durations. In such settings, each user *optimistically* modifies his private copy, called a replica, of a shared datum. As replicas may diverge, this poses the problem of reconciliation. Our research takes into account a number of issues not addressed by previous work, for instance respecting application semantics, high-level operations, dependence, atomicity and conflict, long session times, etc.

5. Software

5.1. Telex

Participants: Marc Shapiro [correspondent], Lamia Benmouffok, Pierre Sutra, Pierpaolo Cincilla.

Developing write-sharing applications is challenging. Developers must deal with difficult problems such as managing distributed state, disconnection, and conflicts. Telex is an application-independent platform to ease development and to provide guarantees. Telex is guided by application-provided parameters: actions (operations) and constraints (concurrency control statements). Telex takes care of replication and persistence, drives application progress, and ensures that replicas eventually agree on a correct, common state. Telex supports partial replication, i.e., sites only receive operations they are interested in. The main data structure of Telex is a large, replicated, highly dynamic graph; we discuss the engineering trade-offs for such a graph and our solutions. Our novel agreement protocol runs Telex ensures, in the background, that replicas converge to a safe state. We conducted an experimental evaluation of the Telex based on a cooperative calendar application and on benchmarks.

We report on application experience, building a collaborative application for model-oriented software engineering above Telex, in SAC 2011 [52]. Future work includes extending Telex to cloud computing, opportunistic mobile networks, and real-time collaboration, within several ANR projects: PROSE (Section 7.1.4), STREAMS (Section 7.1.3) and ConcoRDanT (Section 7.1.2).

The code is freely available on http://gforge.inria.fr/ under a BSD license.

5.2. Treedoc

Participants: Marc Shapiro [correspondent], Marek Zawirski, Hyun-Gul Roh.

A Commutative Replicated Data Type (CRDT) is one where all concurrent operations commute. The replicas of a CRDT converge automatically, without complex concurrency control. We designed and developed a novel CRDT design for cooperative text editing, called Treedoc. It is designed over a dense identifier space based on a binary trees. Treedoc also includes an innovative garbage collection algorithm based on tree rebalancing. In the best case, Treedoc incurs no overhead with respect to a linear text buffer. The implementation has been validated with performance measurements, based on real traces of social text editing in Wikipedia and SVN.

Work in 2010 has focused on studying large-scale garbage collection for Treedoc, and design improvements. Future work includes engineering a large-scale collaborative Wiki, and studying CRDTs more generally. This is the subject the PROSE, STREAMS and ConcoRDanT ANR projects (Sections 7.1.4, 7.1.3 and 7.1.2 respectively).

The code is freely available on http://gforge.inria.fr/ under a BSD license.

5.3. VMKit and .Net runtimes for LLVM

Participants: Bertil Folliot [correspondent], Jean-Pierre Lozi, Gaël Thomas [correspondent], Gilles Muller, Thomas Preud'homme.

Many systems research projects now target managed runtime environments (MRE) because they provide better productivity and safety compared to native environments. Still, developing and optimizing an MRE is a tedious task that requires many years of development. Although MREs share some common functionalities, such as a Just In Time Compiler or a Garbage Collector, this opportunity for sharing hash not been yet exploited in implementing MREs. We are working on VMKit, a first attempt to build a common substrate that eases the development and experimentation of high-level MREs and systems mechanisms. VMKit has been successfully used to build two MREs, a Java Virtual Machine and a Common Language Runtime, as well as a new system mechanism that provides better security in the context of service-oriented architectures.

VMKit project is an implementation of a JVM and a CLI Virtual Machines (Microsoft .NET is an implementation of the CLI) using the LLVM compiler framework and the MMTk garbage collectors. The JVM, called J3, executes real-world applications such as Tomcat, Felix or Eclipse and the DaCapo benchmark. It uses the GNU Classpath project for the base classes. The CLI implementation, called N3, is its in early stages but can execute simple applications and the "pnetmark" benchmark. It uses the pnetlib project or Mono as its core library. The VMKit VMs compare in performance with industrial and top open-source VMs on CPU-intensive applications. VMKit is publicly available under the LLVM license.

http://vmkit.llvm.org/

6. New Results

6.1. Introduction

In 2010, we focused our research on the following areas:

- distributed algorithms for large and dynamic networks,
- Peer-to-peer storage
- dynamic adaptation of virtual machines,
- services management in large scale environments,
- Formal and practical study of optimistic replication, incorporating application semantics.
- Decentralized commitment protocols for semantic optimistic replication.
- dynamic replication in distributed multi-agent systems.

6.2. Distributed algorithms

Participants: Luciana Arantes [correspondent], Maria Gradinariu Potop-Butucaru [correspondent], Mathieu Bouillaguet, Pierre Sens, Julien Sopena.

Our current research in the context of distributed algorithms focuses on two main axes. We are interested in providing fault-tolerant and self*(self-organizing, self-healing and self-stabilizing) solutions for fundamental problems in distributed computing. More precisely, we target the following basic blocks: mutual exclusion, resources allocation, agreement and communication primitives. We propose solutions for both static (eg. grid) and dynamic networks (P2P and mobile networks).

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

- Since 2009, we start exploiting the dynamics of MANETs in order to propose a distributed computing model that characterize as much as possible the dynamic and self-organizing behavior of MANETs'. The temporal variations in the network topology implies that MANET can not be viewed as a static connected graph over which paths between nodes are established beforehand. Path between two nodes is in fact built over the time. Furthermore, lack of connectivity between nodes (temporal or not) makes of MANET a *partitionable system*, i.e., a system in which nodes that do not crash or leave the system might be not capable to communicate between themselves. To this end, a first work has been published in [28].
- The main challenges of our research activity over 2010 year were to develop self* (self-stabilizing, self-organizing and self-healing) local algorithms for dynamic networks (P2P, sensor and robot networks [17], [70]). We addressed fundamental problems such as constructions of fault tolerant and reliable infrastructures for networks hit by topological dynamicity [47]. Furthermore we investigated fault-tolerant agreement in robot networks under various forms of constraints([20]).
- Another ongoing research work focuses on trust assessment in dynamic systems. Even if it is near impossible to fully trust a node in a P2P system, managing a set of the most trusted nodes in the system can help to implement more trusted and reliable services. Using these nodes can reduce the probability of introducing malicious nodes in distributed computations. Our work aims at the following objectives: 1. To design a distributed membership algorithm for structured Peer to Peer networks in order to build a group of trusted nodes. 2. To design a maintenance algorithm to periodically clean the trusted group so as to avoid nodes whose reputation has decreased under the minimum value. 3. To provide a way for a given node X to find at least one trusted node. 4. To design a prototype of an information system, such as a news dissemination system, that relies on the trusted group.

6.3. Peer-to-peer systems

Participants: Pierre Sens [correspondent], Nicolas Hidalgo, Sergey Legtchenko, Sébastien Monnet, Gilles Muller, Corentin Travers, Mathieu Valéro.

6.3.1. Peer-to-peer storage

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

When connection/disconnection frequency is too high in the system, data-blocks may be lost. This is true for most current DHT-based system's implementations. To avoid this problem, it is necessary to build really efficient replication and maintenance mechanisms. Since 2008 we study the effect of churn on an existing DHT-based P2P system namely PAST/Pastry. We have proposed RelaxDHT [24], a churn-resilient peer-to-peer DHT. RelaxDHT proposes an enhanced replication strategy with relaxed placement constraints, avoiding useless data transfers and improving transfer parallelization. This new replication strategy is able to cut down by 2 the number of data-block losses compared to PAST DHT. We are now starting to study the use of erasure coding mechanisms along with replication within DHTs. Our goal is to propose hybrid mechanisms to find a good tradeoff among 1) churn-resilience, 2) maintenance cost, and 3) storage space.

6.3.2. Peer-to-peer overlay

Peer-to-peer overlays allow distributed applications to work in a wide-area, scalable, and fault-tolerant manner. We are considering a new class of target applications: massively multi-player online games (MMOG) such as virtual worlds. Within the context of a project funded by the LIP6, we have modeled a P2P distribution of such applications. Following this model, groups of object replicas are moving among peers while players evolve in the virtual world. For this kind of applications, it is important that the underlying overlay is flexible in order to remain adapted to the changing application structure. We are investigating an even more dynamic kind of overlay: a malleable overlay that anticipate application needs in order to adapt itself in advance [51]. Thanks to this anticipation, the overlay is already operational and efficient when the application uses it.

6.3.3. Peer-to-peer publish-subscribe

Publish/Subscribe implemented on top of distributed R-trees (DR-trees) overlays offer efficient DHT-free communication primitives. We have then extend the distributed R-trees (DR-trees) in order to reduce event delivery latency in order to meet the requirements of massively distributed video games such that pertinent information is quickly distributed to all the interested parties without degrading the load of nodes neither increasing the number of noisy events. The enhanced structure performs better than the traditional distributed R-tree in terms of delivery latency. Additionally, it does not alter the performances related to the scalability, nor the load balancing of the tree, and neither the rate of false positives and negatives filtered by a node. The results of this work can be found in [27]. In [60] we study the mapping between the DR-tree logical nodes and the physical nodes. We propose mechanisms for placement and dynamic migration of virtual nodes that balances the load of the network without modifying the DR-tree virtual structure. Furthermore, since each single crash can potentially break the tree structure connectivity, DR-trees are crash-sensitivite. We this have proposed a fault tolerant approach which exploits replication of non leaf nodes in order to ensure the tree connectivity in presence of crashes. This work will be published in [61].

6.3.4. Peer-to-peer routing protocol

Structured peer-to-peer (p2p) overlay networks provide a scalable object location and routing substrate for large scale distributed applications. However, due to the great number of nodes of such systems, message complexity of their routing protocol may considerably increase network traffic and load. Therefore, we have proposed in [46] a novel routing protocol for structured P2P systems which is specially suitable for applications whose message delivery delay is not critical. It exploits message aggregation and implement a multi-buffer mechanism in order to route aggregate messages in parallel. Experimental results on top of PeerSim show that our protocol can reduce the total transmission delay of messages, the number of hops of a message, and thus, the global traffic and load of the network.

6.4. Virtual machine (VM)

Participants: Bertil Folliot, Gaël Thomas.

Our research interest are in computer systems, particularly operating systems and virtual machines. We focus on resource management, isolation and concurrency management in virtual machines. Since September 2008, we started with Gilles Muller a new complementary research theme on dynamic patches of operating systems.

6.4.1. Virtual machines

Isolation in OSGi: The OSGi framework is a Java-based, centralized, component oriented platform. It is being widely adopted as an execution environment for the development of extensible applications. However, current Java Virtual Machines are unable to isolate components from each other's. By modifying shared variables or allocating too much memory, a malicious component can freeze the complete platform. We work on I-JVM, a Java Virtual Machines that provides a lightweight approach to isolation while preserving the compatibility with legacy OSGi applications. Our evaluation of I-JVM shows that it solves the 15 known OSGi vulnerabilities due to the Java Virtual Machine with an overhead below 20%. I-JVM has been presented in DSN 2009.

VMKit: Managed Runtime Environments (MREs), such as the JVM and the CLI, form an attractive environment for program execution, by providing portability and safety, via the use of a bytecode language and automatic memory management, as well as good performance, via just-in-time (JIT) compilation. Nevertheless, developing such a fully featured MRE, including features such as a garbage collector and JIT compiler, is a herculean task. As a result, new languages cannot easily take advantage of the benefits of MREs, and it is difficult to experiment with extensions of existing MRE based languages. VMKit is a first attempt to build a common substrate that eases the development of high-level MREs. We have successfully used VMKit to build two MREs: a Java Virtual Machine (J3) and a Common Language Runtime (N3). VMKit has performance comparable to the well established open source MREs Cacao, Apache Harmony and Mono. VMKit is freely distributed under the LLVM licence with the LLVM framework developed by the University of Illinois at Urbana-Champaign and now maintained by Apple. In 2010, our results on VMKit were published in VEE conference [45].

6.4.2. Semantic patches

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

We are developing a comprehensive solution to the problem of finding bugs in API usage in open source infrastructure software based on our experience in using the Coccinelle code matching and transformation tool, and our interactions with the Linux community. Coccinelle targets the problem of documenting and automating collateral evolutions in C code, specifically Linux code. A collateral evolution is a change that is needed in the clients of an API when the API changes in some way that affects its interface. Coccinelle provides a language for expressing collateral evolutions by means of Semantic Patches, and a transformation tool for performing them automatically. Recently, we have begun using Coccinelle to generate traditional patches for improving the safety of Linux. Some Linux developers have also begun to use the tool. Over 750 of these patches developed using Coccinelle have been integrated into the mainline Linux kernel, and more have been accepted by Linux maintainers and are pending integration. Our current work is to build on the results of Coccinelle by designing libraries of semantic patches to identify API protocols and detect violations in their usage. One of the novelty of this work is to explore how to develop these semantic patches in collaborative manner with the community of Linux open-source developers as a target.

6.5. Hosted database replication service

Participant: Mesaac Makpangou [correspondent].

Today, the vast majority of content distributed on the web are produced by web 2.0 applications. Examples of such applications include social networks, virtual universities, multi-players games, e-commerce web sites, and search engines. These applications rely on databases to serve end-users' requests. Hence, the success of these applications/services depends mainly on the scalability and the performance of the database backend.

The objective of our research is to provide a hosted database replication service. With respect to end-users applications, this service offers an interface to create, to register, and to access databases. Internally, each hosted database is fragmented and its fragments are replicated towards a peer-to-peer network. We anticipate that such a service may improve the performance and the availability of popular web applications, thanks to partial replications of backend databases. Partial database replication on top of a peer-to-peer network raises a number of difficult issues: (i) enforcing replica consistency in presence of update transactions, without jeopardizing the scalability and the performance of the system? (ii) accommodating the dynamic and the heterogeneity of a peer-to-peer network with the database requirements?

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

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

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

6.6. Commitment protocols for WAN replication

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

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

6.7. Optimistic approaches in collaborative editing

Participants: Marc Shapiro [correspondent], Marek Zawirski, Hyun-Gul Roh.

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

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

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

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

Treedoc will be used in ANR projects PROSE (Section 7.1.4) and STREAMS, and will be further studied and developed in ANR project ConcoRDanT (Section 7.1.2) and under a Google European Doctoral Fellowship.

6.8. CRDTs, a principled approach to eventual consistency

Participants: Marc Shapiro [correspondent], Marek Zawirski, Hyun-Gul Roh.

Most well-studied approaches to replica consistency maintain a global total order of operations. This serialisation constitutes a performance and scalability bottleneck, while the CAP theorem imposes a trade-off between consistency and partition-tolerance. An alternative approach, *eventual consistency* or *optimistic replication*, is attractive. A replica may execute an operation without synchronising *a priori* with other replicas. The operation is sent asynchronously to other replicas; every replica eventually applies all updates, but possibly in different orders. This approach ensures that data remains available despite network partitions, and is perceived to scale well and to provide acceptable quality of service. The consensus bottleneck remains but is off the critical path. However, reconciliation is generally complex. There is little theoretical guidance on how to design a correct optimistic system, and ad-hoc approaches have proven brittle and error-prone. We propose a simple, theoretically sound approach to eventual consistency, the concept of a *convergent* or *commutative replicated data type* (CRDT), for which some simple mathematical properties ensure eventual consistency. Provably, any CRDT converges to a common state that is equivalent to some sequential execution. A CRDT requires no synchronisation, thus every update can execute immediately, unaffected by network latency, faults, or disconnection. It is extremely scalable and is fault-tolerant, and does not require much mechanism. Previously, only a handful of CRDTs were known. Our current research aims to push the CRDT envelope, to study the principles of CRDTs, and to design a library of useful CRDTs. So far we have designed variations on registers, counters, sets, maps (key-value stores), graphs, and sequences. Potential application areas include computation in delay-tolerant networks, latency tolerance in wide-area networks, disconnected operation, churn-tolerant peer-to-peer computing, and partition-tolerant cloud computing. CRDTs are the main topic of ANR project ConcoRDanT (Section 7.1.2). This research is also funded in part by a Google European Doctoral Fellowship.

6.9. Fault tolerance in multi-agent systems

Participants: Olivier Marin [correspondent], Corentin Méhat.

Distributed agent systems stand out as a powerful tool for designing scalable software. The general outline of distributed agent software consists of computational entities which interact with one another towards a common goal that is beyond their individual capabilities

Our research focuses on middleware to deploy agent on large-scale environments and mobile networks. Our main topics of interest comprise: fault tolerance, process replication, and dynamicity with respect to both environments and applications. The ongoing research projects we are working on are all related to these topics.

The FRAME (Failure Resilient Agents in Mobile Environments) project – funded by LIP6 in 2006 and 2007 – aims at designing a middleware for the deployment of distributed algorithms among mobile devices. The originality of our approach is double: (i) we view partial and total disconnection as types of failures and aim to integrate fault tolerance solutions in order to guarantee the continuity of the computation in such a context, and (ii) we provide a modeling language which is close to Pi-calculus and yet focuses on communication channels in order to represent replicated applications and introduce failures. The ongoing PhD effort of Corentin Méhat is at the core of this project.

Our current work addresses the detection of failures among mobile devices. We are presently implementing a solution based on the partition participant detector algorithm [28] designed in our team. This will allow us to determine how well the algorithm fares in real live conditions.

The DARX (Dynamic Agent Replication eXtension) project aims at building an architecture for fault-tolerant agent computing in multi-cluster networks. The originality of our approach lies in two features: (i) an automated replication service which chooses for the application which of its computational components are to be made dependable, to which degree, and at what point of the execution, and (ii) the hierarchic architecture of the middleware which ought to provide suitable support for large-scale applications. DARX is now a component of the FACOMA project, which is supported in the context of the ANR-SETIN frame. FACOMA was originally supposed to end in 2009, but its extension has been approved by the ANR until 2010.

The latest advances include evaluating the performances of our distributed exception-handling system which can be shared by the agent application and the dynamic replication service, and assessing the heuristics developed for the dynamic replication of agents.

The DDEFCON (Dependable DEployment oF Code in Open eNvironments) project addresses the safe and secure deployment of collaborative software components over large-scale networks. We seek to achieve a deployment platform that can be implemented on top of a structured peer to peer overlay. DDEFCON serves as a basis for the PhD thesis of Nicolas Gibelin.

We are currently working on a probabilistic algorithm to determine the criticality of a process with respect to the rest of the computation.

7. Other Grants and Activities

7.1. National initiatives

7.1.1. MyCloud - (2010–2014)

Members: INRIA Rhones-Alpes (SARDES), LIP6 (REGAL), EMN, WeAreCloud, Elastic Cloud

Funding: MyCloud project is funded by ANR Arpège

Objectives: Cloud Computing is a paradigm for enabling remote, on-demand access to a set of configurable computing resources. The objective of the MyCloud project is to define and implement a novel cloud model: SLAaaS (SLA aware Service). Novel models, control laws, distributed algorithms and languages will be proposed for automated provisioning, configuration and deployment of cloud services to meet SLA requirements, while tackling scalability and dynamics issues. The principal investigators for Regal are Luciana Arantes, Pierre Sens, and Julien Sopena. It involves a grant of 155 000 euros from ANR to LIP6 over three years.

7.1.2. ConcoRDanT - (2010–2013)

Members: INRIA Regal, project leader; LORIA, Universdide Nova de Lisboa

Funding: PROSE project is funded by ANR Blanc

Objectives: CRDTs for consistency without concurrency control in Cloud and Peer-To-Peer systems Massive computing systems and their applications suffer from a fundamental tension between scalability and data consistency. Avoiding the synchronisation bottleneck requires highly skilled programmers, makes applications complex and brittle, and is error-prone. The ConcoRDanT project investigates a promising new approach that is simple, scales indefinitely, and provably ensures eventual consistency. A Commutative Replicated Data Type (CRDT) is a data type where all concurrent operations commute. If all replicas execute all operations, they converge; no complex concurrency control is required. We have shown in the past that CRDTs can replace existing techniques in a number of tasks where distributed users can update concurrently, such as co-operative editing, wikis, and version control. However CRDTs are not a universal solution and raise their own issues (e.g., growth of meta-data). The ConcoRDanT project engages in a systematic and principled study of CRDTs, to discover their power and limitations, both theoretical and practical. Its outcome will be a body of knowledge about CRDTs and a library of CRDT designs, and applications using them. We are hopeful that significant distributed applications can be designed using CRDTs, a radical simplification of software, elegantly reconciling scalability and consistency. The project leader and principal investigator for Regal is Marc Shapiro. ConcoRDanT involves a grant of 192 637 euros from ANR to INRIA over three years.

7.1.3. STREAMS - (2010–2013)

Members: LORIA (Score, Cassis), INRIA (Regal, ASAP), Xwiki

Funding: STREAMS is funded by ANR Arpège

Objectives: Solutions for a peer-To-peer REAl-tiMe Social web The STREAMS project proposes to design peer-to-peer solutions that offer underlying services required by real-time social web applications and that eliminate the disadvantages of centralised architectures. These solutions are meant to replace a central authority-based collaboration with a distributed collaboration that offers support for decentralisation of services. The project aims to advance the state of the art on peer-to-peer networks for social and real-time applications. Scalability is generally considered as an inherent characteristic of peer-to-peer systems. It is traditionally achieved using replication techniques. Unfortunately, the current state of the art in peer-to-peer networks does not address replication of continuously updated content due to real-time user changes. Moreover, there exists a tension between sharing data with friends in a social network deployed in an open peer-to-peer network and ensuring privacy. One of the most challenging issues in social applications is how to balance collaboration with access control to shared objects. Interaction is aimed at making shared objects available to all who need them, whereas access control seeks to ensure this availability only to users with proper authorisation. STREAMS project aims at providing theoretical solutions to these challenges as well as practical experimentation. The principal investigators for Regal is Marc Shapiro. It involves a grant of 57 000 euros from ANR to INRIA over three years.

7.1.4. PROSE - (2009–2011)

Members: Technicolor, INRIA (Regal), EURECOM, PLAYADZ, LIAFA

Funding: PROSE project is funded by ANR VERSO

Objectives: Content Shared Through Peer-to-Peer Recommendation & Opportunistic Social Environment

The Prose project is a collective effort to design opportunistic contact sharing schemes, and characterizes the environmental conditions as well as algorithmic and architecture principles that let them operate. The partners of the Prose project will engage in this exploration through various expertise: network measurement, system design, behavioral study, analysis of distributed algorithms, theory of dynamic graph, networking modeling, and performance evaluation.

The principal investigators for Regal are Sébastien Monnet and Marc Shapiro. It involves a grant of 152 000 euros from ANR to INRIA over three years.

7.1.5. ABL - (2009–2011)

Members: Gilles Muller, Gaël Thomas, Julia Lawall, Saha Suman

Funding: ANR Blanc

Objectives: The goal of the "A Bug's Life" (ABL) project is to develop a comprehensive solution to the problem of finding bugs in API usage in open source infrastructure software. The ABL project has grown out of our experience in using the Coccinelle code matching and transformation tool, which we have developed as part of the former ANR project Blanc Coccinelle, and our interactions with the Linux community. Coccinelle targets the problem of documenting and automating collateral evolutions in C code, specifically Linux code. A collateral evolution is a change that is needed in the clients of an API when the API changes in some way that affects its interface. Coccinelle provides a language for expressing collateral evolutions by means of Semantic Patches, and a transformation tool for performing them automatically.

We have used Coccinelle to reproduce over 60 collateral evolutions in recent versions of Linux, affecting almost 6000 files. Recently, we have begun using Coccinelle to generate traditional patches for improving the safety of Linux. Over 750 of these patches developed using Coccinelle have been integrated into the mainline Linux kernel. Julia Lawall was among the top 10 in terms of the number of contributed patches in Linux 2.6.36. She also presented Coccinelle and our recent results on the robustness of Linux code in a lightning at the Kernel Summit in November. Finally, about 20 semantic patches are integrated into the Linux sources so that developpers can improve the quality of their programs by running Coccinelle as part of the development process.

In the ABL project, we build on the results of Coccinelle by 1) designing libraries of semantic patches to identify API protocols and detect violations in their usage, 2) extending Coccinelle to address the needs of bug finding and reporting, and 3) designing complementary tools to help the programmer to track and fix bugs.

7.1.6. SHAMAN - (2009–2011)

Members: LIP6 (NPA), Inria Saclay (Grand-Large), Inria Bretagne (ASAP), LIP6 (Regal)

Funding: SHAMAN project is funded by ANR TELECOM

Objectives: Large-scale networks (e.g. sensor networks, peer-to-peer networks) typically include several thousands (or even hundred thousand) basic elements (computers, processors) endowed with communication capabilities (low power radio, dedicated fast network, Internet). Because of the large number of involved components, these systems are particularly vulnerable to occurrences of failures or attacks (permanent, transient, intermittent). Our focus in this project is to enable the sustainability of autonomous network functionalities in spite of component failures (lack of power, physical damage, software or environmental interference, etc.) or system evolution (changes in topology, alteration of needs or capacities). We emphasize the self-organization, fault-tolerance, and resource saving properties of the potential solutions. In this project, we will consider two different kinds of large-scale systems: on one hand sensor networks, and on the other hand peer to peer networks.

7.1.7. R-DISCOVER - (2009–2011)

Members: MIS, LASMEA, GREYC, LIP6 (Regal), Thales

Funding: R-DISCOVER project is funded by ANR CONTINT

Objectives: This project considers a set of sensors and mobile robots arbitrarily deployed in a geographical area. Sensors are static. The robots can move and observe the positions of other robots and sensors in the plane and based on these observations they perform some local computations. This project addresses the problem of topological and cooperative navigation of robots in such complex systems.

7.1.8. SPREADS - (2008–2010)

Members: UbiStorage, LACL, Inria Sophia, Inria (Regal)

Funding: SPREADS project is funded by ANR TELECOM

Objectives: This project proposes a collaborative research effort to study and design highly dynamic secure P2P storage systems on large scale networks like the Internet. The scientific program covered by this proposal is mainly the design of new mathematical safety, security and performance models, secure patterns, simulation to evaluate the quality of service of a peer-to-peer storage system in the context of a dynamic large scale network. These models and simulations will eventually be corroborated by experimentation on the Grid 5000 and Grid eXplorer Platforms.

7.1.9. Facoma - (2007–2009)

Members: LIP6, LIRMM, Regal

Funding: Facoma project is funded by ANR SETIN

Objectives: The fault tolerance research community has developped solutions (algorithms and architectures), mostly based on the concept of replication, applied for instance to data bases. But, these techniques are almost always applied explicitely and statically. This is the responsability of the designer of the application to identify explicitly which critical servers should be made robust and also to decide which strategies (active or passive replication) and their configurations (how many replicas, their placement). Meanwhile, regarding new cooperative applications, which are very dynamic, for instance: decision support systems, distributed control, electronic commerce, crisis management systems, and intelligent sensors networks, - such applications increasingly modeled as a set of cooperative agents (multi-agent systems) -, it is very difficult, or even impossible, to identify in advance the most critical agents of the application. This is because the roles and relative importances of the agents can greatly vary during the course of computation, interaction and cooperation, the agents being able to change roles, strategies, plans, and new agents may also join or leave the application (open system). Our approach is in consequence to give the capacity to the multi-agent system itself to dynamically identify the most critical agents and to decide which abilisation strategies to apply to them.

7.1.10. PACTOL - (2009–2011)

Members: LIP6 (NPA, Regal), CNAM

Funding: Digiteo

Objectives: The scope of PACTOL is to propose verification tools for self-stabilizing distributed algorithms.

7.1.11. MOTAR2 - 2009

Members: NPA, Regal (LIP6)

Funding: LIP6

Objectives: The study of fault tolerance in robot networks.

7.2. European Initiatives

7.2.1. Google European Doctoral Fellowship “A principled approach to eventual consistency based on CRDTs

Cloud computing systems suffer from a fundamental tension between scalability and data consistency. Avoiding the synchronisation bottleneck requires highly skilled programmers, makes applications complex and brittle, and is error-prone. The Commutative Replicated Data Type (CRDT) approach, based on commutativity, is a simple and principled solution to this conundrum; however, only a handful of CRDTs are known, and CRDTs are not a universal solution. This PhD research aims to expand our knowledge of CRDTs, to design and implement a re-usable library of composable CRDTs, to maintain study techniques for maintaining strong invariants above CRDTs, and to experiment with CRDTs in applications. We are hopeful that significant distributed applications can be designed using our techniques, which would radically simplify the design of cloud software, reconciling scalability and consistency. This Google European Doctoral Fellowship is awarded to Marek Zawirski, advised by Marc Shapiro. This award includes a grant of 41 000 euros yearly over three years starting September 2010.

7.2.2. FTH-GRID - (2009–2011)

Members: Université de Lisbonne (LASIGE), LIP6 (Regal)

Funding: Egide

Objectives: FTH-Grid, Fault-Tolerant Hierarchical Grid Scheduling, is a cooperation project between the Laboratoire d'Informatique de Paris 6 (LIP6/CNRS, France) and the Large-Scale Informatics Systems Laboratory (LASIGE/FCUL, Portugal).

Its goal is to foster scientific research collaboration between the two research teams. The project aims at rendering Map Reduce on top of Grid tolerant to byzantine failure. Map Reduce is a programming model for large-scale data-parallel applications whose implementation is based on master-slave scheduling of bag-of-tasks. MapReduce breaks a computation into small tasks that run in parallel on different machines, scaling easily to several cluster. The core research activities of the project consist mainly in extending the execution and programming model to make Byzantine fault-tolerant MapReduce applications. The project was extended for another year, after a results assessment by Egide.

7.3. International Initiatives

7.3.1. DEMEDYS - INRIA-CNPq - (2010–2011)

Members: INRIA Bretagne (ASAP), INRIA Paris Rocquencourt (REGAL), UFBA (Bahia, Brazil), IME (Sao Paulo, Brazil)

Funding: INRIA / CNPq

Objectives: DEMEDYS Project (Dependable Mechanisms for Dynamic Systems) will study fundamental aspects of dynamic distributed systems.

7.3.2. Bi-lateral collaborations

JAIST (Japon). With the group of Prof. Xavier Defago we investigate various aspects of self-organization and fault tolerance in the context of robots networks.

UNLV (USA) With the group of Prof. Ajoy Datta we collaborate in designing self* solutions for the computations of connected covers of query regions in sensor networks.

Technion (Israel). We collaborate with Prof. Roy Friedman on divers aspects of dynamic systems ranging from the computation of connected covers to the design of agreement problems adequate for P2P networks.

Ben Gurion (Israel). We collaborate recently with prof. Shlomi Dolev on the implementation of self-stabilizing atomic memory.

Kent University (SUA) With prof. Mikhail Nesterenko we started recently a collaboration on FTSS solutions for dynamic tasks.

Nagoya Institute of Technology (Japon) With prof. Taisuke Izumi we started this year a collaboration on the probabilistic aspects of robot networks.

COFECUB (Brazil). With the group of Prof. F. Greve. (Univ. Federal of Bahia), we investigate various aspects of failure detection for dynamic environement such as MANET of P2P systems.

CONYCIT (Chili). Since 2007, we start on new collaboration with the group of X. Bonnaire Fabre (Universidad Técnica Federico Santa María - Valparaiso). The main goal is to implement trusted services in P2P environment. Even if it is near impossible to fully trust a node in a P2P system, managing a set of the most trusted nodes in the system can help to implement more trusted and reliable services. Using these nodes, can reduce the probability to have some malicious nodes that will not correctly provide the given service. The project will have the following objectives: 1. To design a distributed membership algorithm for structured Peer to Peer networks in order to build a group of trusted nodes. 2. To design a maintenance algorithm to periodically clean the trusted group so as to avoid nodes whose reputation has decreased under the minimum value. 3. To provide a way for a given node X to find at least one trusted node. 4. To design a prototype of an information system, such as a news dissemination system, that relies on the trusted group.

Collaboration with CITI-UNL, Portugal Our collaboration with CITI, the Research Center for Informatics and Information Technologies of UNL, the New University of Lisbon (Portugal), is materialised by several joint articles. Furthermore, Marc Shapiro is an advisor to the project "RepComp - Replicated Components for Improved Performance or Reliability in Multicore Systems," funded by Fundação para a Ciência e a Tecnologia (FCT, Portuguese equivalent of ANR). Finally, Marc Shapiro is a Member of the CITI Advisory Board.

8. Dissemination

8.1. Animation of the scientific community

Luciana Arantes is:

- Member of the program committee of the 6ème Conférence française sur les systèmes d'exploitation, CFSE-6, Friburg, Switzerland, february 2008.
- Member of PC of Workshop de Sistemas Operacionais, SBC, Brésil, 2007-2010.
- Member of PC of WTF - Workshop of Fault Tolerance, Brésil, 2009-201
- Member of PC of International Conference on Grid and Pervasive Computing, 2009-2011.
- Member of PC of LADC - Fifth Latin-American Symposium on Dependable Computing 2011.
- Reviewer for JPDC and TPDS journals.

Bertil Folliot is:

- Head of the Network and Distributed Systems department at LIP6.
- Elected member of the "Commission de spécialistes" of the Paris 6 University.
- Member of a selection committee for recruiting a full professor.
- Member of the scientific committee of LIP6.
- Member of the "Executive Comittee" of GdR ASR (Hardware, System and Network), CNRS.
- Elected member of the IFIP WG10.3 working group (International Federation for Information Processing - Concurrent systems).
- Member of the "Advisory Board" of EuroPar (International European Conference on Parallel and Distributed Computing), IFIP/ACM.
- Member of the "Steering Committee" of the International Symposium on Parallel and Distributed Computing".
- Member of the program committee of the 2010 International Conference on High Performance Computing & Simulation (HPCS 2010), Caen, France, july 2010.
- Member of the program committee of the 8th International Conference on the Principles and Practice of Programming in Java, Vienna, Austria, september 2010.

Maria Gradinariu Potop-Butucaru is:
computing)

- Co-chair of Algotel 2010 (12eme rencontres francophones sur les aspects algorithmiques de telecommunications)
- Member of the program commitee of ICDCN 2010 (International Conference on distributed computing and networking)
- Member of the University Pierre et Marie Curie UFR Computer Science Council
- Reviewer for Distributed Computing, SIAM Journal of Computing, IEEE TPDS, IEEE Transactions on Computers, Theoretical Computer Science Journal ...

Olivier Marin is:

- Reviewer for Distributed Computing journal.
- Elected vice-president of the "Commission de spécialistes" of the Paris 6 University, section 27.
- Elected member of the "Commission des primes de recherche" of the Paris 6 University.
- Elected member of the "Commission des avancements de carrière - rang B" of the Paris 6 University.

Sébastien Monnet is:

- Member of the PC of the first 1st International Workshop on Fault-Tolerance for HPC at Extreme Scale (held with DSN 2010, Chicago, Illinois, USA, June 2010).
- Member of the program commitee of the storage track for the 5th IEEE International Conference on Networking, Architecture, and Storage (NAS 2010) July 2010, Macau SAR, China.
- GDR ASR mailing lists moderator

Gilles Muller is:

- Member of PC of the SRDS 2011 conference, October 2011 http://lsd.ls.fi.upm.es:11888/srds/.
- Member of PC of the SYSTOR 2011 conference, May 2011 https://www.research.ibm.com/haifa/conferences/systor2011/.
- Member of the jury of the best European thesis on systems (EuroSys) 2011.
- Member of PC of the 8ème Conférence française sur les systèmes d'exploitation, CFSE-8, May 2011 http://renpar.irisa.fr/CFP-CFSE.html.
- Member of PC of PC of the 5th EuroSys Doctoral workshop to be held in Saltzburg, April 2011.
- Member of PC of the HotDep 2010 workshop, October 2010 http://www.usenix.org/event/hotdep10/ .
- Member of PC of the DSN 2010 conference, June 2010 http://www.dsn.org/.
- Chair of PC of the EuroSys 2010 conference, April 2010 http://eurosys2010.sigops-france.fr/.

Julia Lawall is:

- Vice chair of the IFIP TC-2 working group on Program Generation (WG 2.11) http://resource-aware.org/do/view/WG211
- Vice chair and then chair of the Nordic chapter of Sigma Xi http://sigmaxinordic.wordpress.com/
- Member of PC of the International Workshop on Formalization of Modeling Languages, June 2011
- Member of PC of the 20th European Symposium on Programming (ESOP-2011), March 2011.
- Member of PC of the 8th IEEE International Conference on Embedded and Ubiquitous Computing (EUC-10), December 2010.

- Member of PC of the 9th AOSD Workshop on Aspects, Components, and Patterns for Infrastructure Software (ACP4IS), March 2010.
- Member of PC of Eurosys 2010 conference. April 2010.
- Member of PC of the 5th Domain-Specific Aspect Languages Workshop. March 2010.
- Member of PC of the Second Workshop on Generative Technologies (WGT-2010). March 2010.
- Member of PC of the International Conference on Compiler Construction (CC-2010). March 2010.
- Member of the "Comité de sélection" for "Chaire" CNRS/IRIT (Toulouse).
- Evaluation for "Fundacao para Ciencia e a Tecnologia" (Portugal) in Computer Sciences and Engineering

Franck Petit is:

- Co-program chair of SSS 2011, 13^{th} International Symposium on Stabilization, Safety, and Security of Distributed Systems, ed. LNCS, Japan, 2011.
- PC Member of LAFT 2011, 2^{nd} International Workshop on Logical Aspects of Fault-Tolerance, co-collated with 26^{th} Annual IEEE Symposium on Logic in Computer Science (LICS 2011) Toronto, Canada.
- PC Member of Renpar 2011, 20^{th} Rencontres francophones du Parallélisme, France, 2011.
- Co-Chair of WRAS 2010, 3^{rd} ACM SIGACT-SIGOPS Workshop on Reliability, Availability, and Security, co-collated with ACM SIGACT-SIGOPS PODC 2010, Zurich, Switzerland, 2010.
- PC Member PODC 2010, 29^{th} ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing, Ed. ACM, Zurich, Switzerland, 2010.
- PC Member of OPODIS 2010, 14^{th} International Conference On Principles Of Distributed Systems, ed. LNCS, Tozeur, Tunisia, 2010.
- Co-chair of Stabilization Track of SSS 2010, International Symposium on Stabilization, Safety, and Security of Distributed Systems, Ed. LNCS, Nov. 2010, New-York City, USA.
- Co-program chair, chair of stabilization track, and co-organizing chair of SSS 2009, International Symposium on Stabilization, Safety, and Security of Distributed Systems, Ed. LNCS, Nov. 2009, Lyon, France.
- Member of PC of WRAS 2009, 2^{nd} International Workshop on Reliability, Availability, and Security, Ed. IEEE, Dec. 2009, Hiroshima, Japan.
- Invited Editor for ACM TAAS, Transactions on Autonomous and Adaptive Systems, Special Issue on Stabilization, Safety, and Security of Distributed Systems,
- Invited Editor for TCS, Theoretical Computer Science Special Issue on Stabilization, Safety, and Security of Distributed Systems,
- Reviewer during 2009-2010 for Journal of the ACM (JACM), ACM Transactions on Algorithms (TALG), ACM Transactions on Autonomous and Adaptive Systems (TAAS), IEEE Transactions on Robotics, The Computer Journal, IEEE Transaction on Parallel and Distributed Systems (TPDS), Information Processing Letters (IPL), Journal of Parallel and Distributed Computing (JPDC), Parallel Processing Letters (PPL), Theoretical Computer Science (TCS).
- Elected member of the administrative committee of LIP6.

Pierre Sens is:

- Global chair of Topic 8 “Distributed systems and algorithms” of EuroPar 2010
- Member of PC of SSS 2010 (International Symposium on Stabilization, Safety, and Security of Distributed Systems).
- vice-chair of LIP6 Laboratory.
- Member of the scientific council of AFNIC.
- Member of the scientific committee of LIP6.
- Member of the evaluation committee of the Digiteo DIM LSC program.
- Member of scientific committee of ANR project Blanc International.
- Member of "Comité de selection" of Universities of Amiens, Grenoble, and Paris X.

Marc Shapiro is:

- Member of Advisory Board for CITI, the Research Center for Informatics and Information Technologies of UNL, the New University of Lisbon (Portugal).
- Member of the steering committee for the LADIS workshop (Large-Scale Distributed Systems and Middleware).
- Member of PC of EuroSys 2010.
- PC Chair of CFSE 2011, and member of PC of CFSE 2009 (Conférence française sur les systèmes d'exploitation).
- Promotion reviewer for various European universities (names confidential).
- Reviewer for European Research Council.
- Reviewer for ANR (Agence Nationale de la Recherche), France.
- Reviewer for National Science Foundation, Switzerland.
- Reviewer for USA-Israel Binational Science Foundation.
- Reviewer for Swedish Research Council (Vetenskapsrådet).
- Reviewer for Springer Distributed Computing.
- Reviewer for IEEE Transactions on Parallel and Distributed Systems (TPDS).
- Member, ACM Distinguished Service Award Committee 2010.
- Member, ACM Europe Council. Co-chair, ACM Europe Council subcommittee on Members and Awards.

8.2. PhD reviews

Bertil Folliot was the president of the jury of:

- Maria Potop-Butucaru. Self* networks: from static to dynamic large scale networks. HdR, Univ. Paris 6, December 2010.

Gilles Muller was part of the committee of:

- Fabien Gaud. PhD. Univ. Grenoble, December 2010, Member
- Vivien Quéma. HDR. Univ. Grenoble, November 2010, Member.
- Riadh Ben Abdallah, PhD, Insa Lyon, October 2010, Reviewer.
- Xavier Guérin, PhD, Univ. Grenoble, May 2010, Reviewer.
- Marc Poulhies, PhD, Univ. Grenoble, March 2010, Reviewer.

Franck Petit was member of thesis commitees of:

- Clémence Magnien, HDR. UPMC, LiP6, July 2010, Commitee Chair
- *Stéphane Rovedakis, PhD. Univ. Evry-Val-d'Essonne, IBISC, December 2009, Reviewer*
- *Akka Zemmari*, HDR. Univ. Bordeaux, October 2009, LaBRI, Bordeaux, Commitee Chair
- Mikhail Nesterenko, Full Professor Promotion of Kent State University, USA, August 2009, Reviewer

Pierre Sens was the reviewer of:

- Sara Bouchenak. HDR. Univ. Grenoble, December 2010
- Cyril Labbé. HDR. Univ. Grenoble, December 2010
- Thierry Monteil. HDR. Laas. December 2010
- Eddy Carron. HDR. ENS Lyon. October 2010
- F. Bonnet. PhD. IRISA, (Advisor : M. Raynal)
- S. Duquennoy. PhD. LIFL, (Advisor : G. Grimaud)
- W. Palma. PhD. LIMA, (Advisor: P. Valduriez)
- J. Arnaud. PhD LIG, (Advisors: JB. Stefani, S. Bouchenak)
- C. Canon. PhD LORIA, (Advisor: E. Jeannot)
- F. Bouabache. PhD. LRI, (Advisor: F. Cappello)

8.3. Teaching

- Luciana Arantes
  - Principles of operating systems in Licence d'Informatique, Université Paris 6
  - Operating systems kernel in Master Informatique, Université Paris 6
  - Distributed algorithms in Master Informatique, Université Paris 6
  - Responsible of Advanced distributed algorithms in Master Informatique, Université Paris 6
  - Unix system programming, Licence and Master d'Informatique, Université Paris 6
- Bertil Folliot
  - Principles of operating systems in Licence d'Informatique, Université Paris 6
  - Distributed algorithms and systems in Master Informatique, Université Paris 6
  - Distributed systems and client/serveur in Master Informatique, Université Paris 6
  - Projects in distributed programming in Master Informatique, Université Paris 6
- Maria Gradinariu Potop-Butucaru
  - Principles of Operating Systems, Licence d'Informatique, Université Paris 6
  - Distributed algorithms, Master d'Informatique, Université Paris 6
- Mesaac Makpangou
  - Client/server architecture, Licence professionelle d'Informatique, Université Paris 6: TD, 16 heures
- Oliver Marin
  - Advanced operating systems programming, Master d'Informatique, Université Paris 6
  - Advanced distributed algorithms in Master Informatique, Université Paris 6
  - Responsible of operating systems programming, Licence d'Informatique, Université Paris 6
  - Parallel and distributed systems, Master d'Informatique, Université Paris 6
  - Responsible of the "Parcours professionnalisant L3 - Développeur d'Applications - Nouvelles Technologies (DANT)"

  - Responsible of Deployment of Cooperative Objects, Licence d'Informatique - parcours DANT, Université Paris 6

- Sébastien Monnet
  - Responsible of "Middleware for advanced computing systems in Master d'Informatique (2), Université Paris 6"
  - Operating systems kernel in Master Informatique (1), Université Paris 6
  - Principles of operating systems in Licence d'Informatique (3), Universié Paris 6
  - System and Internet programmation in Licence d'Informatique (2), Université Paris 6
  - Computer science initiation in Licence d'Informatique (1), Université Paris 6
- Pierre Sens
  - Responsible of "Principles of operating systems" in Licence d'Informatique, Université Paris 6
  - Responsible of "Operating systems kernel in Master Informatique", Université Paris 6
  - Distributed systems and algorithms in Master Informatique, Université Paris 6
- Marc Shapiro
  - Teaches NMV (*Noyaux Multi-cœurs et Virtualisation*, i.e., multicore kernels and virtualisation) at Université Paris 6, Master 2.
- Gaël Thomas
  - Responsible for the Master 1 module "Systèmes Répartis Clients/Serveurs" in Master Informatique at the Univeristy Université Paris 6
  - Responsible for the Master 2 module "Middleware Orientés Composants" in Master Informatique at the Univeristy Université Paris 6
  - Responsible for Master 2 module NMV (*Noyaux Multi-cœurs et Virtualisation*, i.e., multicore kernels and virtualisation)
  - Responsible for the Master 2 module "Répartition et Client/Serveur" in Master Informatique at the Univeristy Université Paris 6
  - "Noyau des Systèmes d'exploitation" in Master Informatique at the Université Paris 6
  - "Systèmes" at PolyTech' Paris

9. Bibliography

Major publications by the team in recent years

[1] E. ANCEAUME, R. FRIEDMAN, M. GRADINARIU. *Managed Agreement: Generalizing two fundamental distributed agreement problems*, in "Inf. Process. Lett.", 2007, vol. 101, n° 5, p. 190-198.

[2] L. ARANTES, D. POITRENAUD, P. SENS, B. FOLLIOT. *The Barrier-Lock Clock: A Scalable Synchronization-Oriented Logical Clock*, in "Parallel Processing Letters", 2001, vol. 11, n° 1, p. 65–76.

[3] J. BEAUQUIER, M. GRADINARIU, C. JOHNEN. *Randomized self-stabilizing and space optimal leader election under arbitrary scheduler on rings*, in "Distributed Computing", 2007, vol. 20, n° 1, p. 75-93.

[4] M. BERTIER, L. ARANTES, P. SENS. *Distributed Mutual Exclusion Algorithms for Grid Applications: A Hierarchical Approach*, in "JPDC: Journal of Parallel and Distributed Computing", 2006, vol. 66, p. 128–144.

[5] M. BERTIER, O. MARIN, P. SENS. *Implementation and performance of an adaptable failure detecto r*, in "Proceedings of the International Conference on Dependable Systems and Networks (DSN '02)", June 2002.

[6] M. BERTIER, O. MARIN, P. SENS. *Performance Analysis of Hierarchical Failure Detector*, in "Proceedings of the International Conference on Dependable Systems and Networks (DSN '03)", San-Francisco (USA), IEEE Society Press, June 2003.

[7] J.-M. BUSCA, F. PICCONI, P. SENS. *Pastis: a Highly-Scalabel Multi-User Peer-to-Peer File Systems*, in "Euro-Par'05 - Parallel Processing", Lisboa, Portugal, Lecture Notes in Computer Science, Springer-Verlag, August 2005.

[8] A.-M. KERMARREC, A. ROWSTRON, M. SHAPIRO, P. DRUSCHEL. *The IceCube approach to the reconciliation of divergent replicas*, in "20th Symp. on Principles of Dist. Comp. (PODC)", Newport RI (USA), ACM SIGACT-SIGOPS, August 2001.

[9] N. KRISHNA, M. SHAPIRO, K. BHARGAVAN. *Brief announcement: Exploring the Consistency Problem Space*, in "Symp. on Prin. of Dist. Computing (PODC)", Las Vegas, Nevada, USA, ACM SIGACT-SIGOPS, July 2005.

[10] O. MARIN, M. BERTIER, P. SENS. *DARX - A Framework For The Fault-Tolerant Support Of Agent S oftware*, in "Proceedings of the 14th IEEE International Symposium on Sofwat are Reliability Engineering (ISSRE '03)", Denver (USA), IEEE Society Press, November 2003.

[11] F. OGEL, G. THOMAS, A. GALLAND, B. FOLLIOT. *MVV : une Plate-forme à Composants Dynamiquement Reconfigurables — La Machine Virtuelle Virtuelle*, in "Numéro Spécial Technique et Science Informatiques (TSI)", 2004.

[12] Y. PADIOLEAU, J. L. LAWALL, R. R. HANSEN, G. MULLER. *Documenting and Automating Collateral Evolutions in Linux Device Drivers*, in "EuroSys 2008", Glasgow, Scotland, March 2008, p. 247–260.

[13] Y. PADIOLEAU, J. L. LAWALL, G. MULLER. *Understanding Collateral Evolution in Linux Device Drivers*, in "The first ACM SIGOPS EuroSys conference (EuroSys 2006)", Leuven, Belgium, April 2006, p. 59-71, Also available as INRIA Research Report RR-5769, http://hal.inria.fr/inria-00070251/en/.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[14] M. G. POTOP-BUTUCARU. *Self* networks: from static to dynamic large scale networks*, Université Pierre et Marie Curie (Paris 6), 4, place Jussieu, Paris, december 2010, Habilitation à Diriger des Recherches.

[15] P. SUTRA. *Protocoles efficaces pour le consensus généralisé et la réplication partielle*, Université Pierre et Marie Curie (Paris 6), 4, place Jussieu, Paris, december 2010.

Articles in International Peer-Reviewed Journal

[16] S. BIANCHI, P. FELBER, M. G. POTOP-BUTUCARU. *Stabilizing Distributed R-Trees for Peer-to-Peer Content Routing*, in "IEEE Trans. Parallel Distrib. Syst.", 2010, vol. 21, n^{o} 8, p. 1175-1187.

[17] Z. BOUZID, M. G. POTOP-BUTUCARU, S. TIXEUIL. *Optimal Byzantine-resilient convergence in unidimensional robot networks*, in "Theor. Comput. Sci.", 2010, vol. 411, n^{o} 34-36, p. 3154-3168.

[18] L. BURGY, L. RÉVEILLÈRE, J. L. LAWALL, G. MULLER. *Zebu: A Language-Based Approach for Network Protocol Message Processing*, in "IEEE Transactions on Software Engineering", 2010.

[19] E. CARON, F. DESPREZ, F. PETIT, C. TEDESCHI. *Snap-Stabilizing Prefix Tree for Peer-to-Peer Systems*, in "Parallel Processing Letters", 2010, vol. 20, n^{o} 1, p. 15-30.

[20] J. CLEMENT, X. DÉFAGO, M. GRADINARIU POTOP-BUTUCARU, T. IZUMI, S. MESSIKA. *The cost of probabilistic agreement in oblivious robot networks*, in "Inf. Process. Lett.", 2010, vol. 110, n^{o} 11, p. 431-438.

[21] Y. DIEUDONNÉ, O. LABBANI-IGBIDA, F. PETIT. *Deterministic Robot-Network Localization is Hard*, in "IEEE Transactions on Robotics", 2010, vol. 26, n^{o} 2, p. 331-339.

[22] E. EIDE, G. MULLER, O. SPINCZYK. *Report on 5th workshop on programming languages and operating systems (PLOS)*, in "Operating System Review", 2010, n^{o} 43:4, p. 31-34.

[23] Z. GUESSOUM, J.-P. BRIOT, N. FACI, O. MARIN. *Towards reliable multi-agent systems: An adaptive replication mechanism*, in "Multiagent and Grid Systems", 2010, vol. 6, n^{o} 1, p. 1-24.

[24] S. LEGTCHENKO, S. MONNET, P. SENS, G. MULLER. *RelaxDHT: a churn-resilient replication strategy for peer-to-peer distributed hash-tables*, in "ACM Transactions on Autonomous and Adaptive Systems", 2011, To appear.

[25] H.-G. ROH, M. JEON, J.-S. KIM, J. LEE. *Replicated Abstract Data Types: Building Blocks for Collaborative Applications*, in "Journal of Parallel and Distributed Computing", 2011, (To appear).

International Peer-Reviewed Conference/Proceedings

[26] N. ALON, H. ATTIYA, S. DOLEV, S. DUBOIS, M. G. POTOP-BUTUCARU, S. TIXEUIL. *Brief Announcement: Sharing Memory in a Self-stabilizing Manner*, in "24th International Symposium Distributed Computing (DISC)", 2010, p. 525-527.

[27] L. ARANTES, M. GRADINARIU POTOP-BUTUCARU, P. SENS, M. VALERO. *Enhanced DR-tree for low latency filtering in publish/subscribe systems*, in "International Conference on. Advanced Information Networking and Applications (AINA)", 2010.

[28] L. ARANTES, P. SENS, G. THOMAS, D. CONAN, L. LIM. *Partition Participant Detector with Dynamic Paths in Mobile Networks*, in "IEEE International Symposium on Networking Computing and Applications (NCA)", 2010.

[29] S. BERNARD, S. DEVISMES, K. PARROUX, M. GRADINARIU POTOP-BUTUCARU, S. TIXEUIL. *Probabilistic Self-stabilizing Vertex Coloring in Unidirectional Anonymous Networks*, in "11th International Conference Distributed Computing and Networking (ICDCN)", 2010, p. 167-177.

[30] S. Bernard, M. G. Potop-Butucaru, S. Tixeuil. *A Framework for Secure and Private P2P Publish/Subscribe*, in "12th International Symposium Stabilization, Safety, and Security of Distributed Systems (SSS)", 2010, p. 531-545.

[31] A. N. Bessani, V. V. Cogo, M. Correia, P. Costa, M. Pasin, F. Silva, L. Arantes, O. Marin, P. Sens, J. Sopena. *Making Hadoop MapReduce Byzantine Fault-Tolerant - Fast Abstract*, in "The 40th Annual IEEE/IFIP International Conference on Dependable Systems and Networks", 2010.

[32] T. F. Bissyandé, L. Réveillère, Y.-D. Bromberg, J. L. Lawall, G. Muller. *Bridging the Gap between Legacy Services and Web Services*, in "Middleware", Bangalore, India, December 2010, p. 273–292.

[33] L. Blin, A. Milani, M. G. Potop-Butucaru, S. Tixeuil. *Exclusive Perpetual Ring Exploration without Chirality*, in "24th International Symposium Distributed Computing (DISC)", 2010, p. 312-327.

[34] L. Blin, M. G. Potop-Butucaru, S. Rovedakis, S. Tixeuil. *Loop-Free Super-Stabilizing Spanning Tree Construction*, in "12th International Symposium Stabilization, Safety, and Security of Distributed Systems (SSS)", 2010, p. 50-64.

[35] B. Bonakdarpour, S. Devismes, F. Petit. *Snap-Stabilizing Committee Coordination*, in "25th IEEE International Parallel and Distributed Processing Symposium (IPDPS)", Anchorage, Alaska, 2011, to appear.

[36] D. Cassou, E. Balland, C. Consel, J. L. Lawall. *Leveraging Architectures to Guide and Verify Development of Sense/Compute/Control Applications*, in "33rd International Conference on Software Engineering (ICSE)", Honolulu, HI, USA, May 2011, To appear.

[37] A. Cournier, S. Dubois, A. Lamani, F. Petit, V. Villain. *Snap-Stabilizing Linear Message Forwarding*, in "12th International Symposium Stabilization, Safety, and Security of Distributed Systems (SSS)", 2010, p. 546-559.

[38] Y. Dieudonné, F. Petit, V. Villain. *Brief announcement: leader election vs pattern formation*, in "29th Annual ACM Symposium on Principles of Distributed Computing (PODC)", 2010, p. 404-405.

[39] Y. Dieudonné, F. Petit, V. Villain. *Leader Election Problem versus Pattern Formation Problem*, in "24th International Symposium Distributed Computing (DISC)", 2010, p. 267-281.

[40] S. Dubois, T. Masuzawa, S. Tixeuil. *On Byzantine Containment Properties of the in + 1 Protocol*, in "12th International Symposium Stabilization, Safety, and Security of Distributed Systems (SSS)", 2010, p. 96-110.

[41] S. Dubois, T. Masuzawa, S. Tixeuil. *The Impact of Topology on Byzantine Containment in Stabilization*, in "24th International Symposium Distributed Computing (DISC)", 2010, p. 495-509.

[42] B. Ducourthial, S. Khalfallah, F. Petit. *Best-effort group service in dynamic networks*, in "22nd Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)", 2010, p. 233-242.

[43] M. Filali-Amine, J. L. Lawall. *Development of a Synchronous Subset of AADL*, in "Abstract State Machines, Alloy, B and Z, Second International Conference (ABZ)", Orford, QC, Canada, Lecture Notes in Computer Science, February 2010, vol. 5977, p. 245–258.

[44] F. GAUD, S. GENEVÈS, R. LACHAIZE, B. LEPERS, F. MOTTET, G. MULLER, V. QUÉMA. *Efficient Workstealing for Multicore Event-Driven Systems*, in "International Conference on Distributed Computing Systems (ICDCS)", 2010, p. 516-525.

[45] N. GEOFFRAY, G. THOMAS, J. L. LAWALL, G. MULLER, B. FOLLIOT. *VMKit: a substrate for managed runtime environments*, in "6th International Conference on Virtual Execution Environments (VEE)", 2010, p. 51-62.

[46] N. HIDALGO, L. ARANTES, P. SENS. *An Aggregation-Based Routing Protocol for Structured Peer to Peer Overlay Networks*, in "The Second International Conference on Advances in P2P Systems", 2010.

[47] T. IZUMI, M. G. POTOP-BUTUCARU, S. TIXEUIL. *Connectivity-Preserving Scattering of Mobile Robots with Limited Visibility*, in "12th International Symposium Stabilization, Safety, and Security of Distributed Systems (SSS)", 2010, p. 319-331.

[48] A. LAMANI, M. G. POTOP-BUTUCARU, S. TIXEUIL. *Optimal Deterministic Ring Exploration with Oblivious Asynchronous Robots*, in "17th International Colloquium Structural Information and Communication Complexity (SIROCCO)", 2010, p. 183-196.

[49] J. L. LAWALL, B. LAURIE, R. R. HANSEN, N. PALIX, G. MULLER. *Finding Error Handling Bugs in OpenSSL Using Coccinelle*, in "Eighth European Dependable Computing Conference (EDCC)", 2010, p. 191-196.

[50] J. L. LAWALL, D. LO. *An automated approach for finding variable-constant pairing bugs*, in "25th IEEE/ACM International Conference on Automated Software Engineering (ASE)", Antwerp, Belgium, September 2010, p. 103–112.

[51] S. LEGTCHENKO, S. MONNET, G. THOMAS. *Blue banana: resilience to avatar mobility in distributed MMOGs*, in "The 40th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)", July 2010.

[52] J. MICHAUX, X. BLANC, P. SUTRA, M. SHAPIRO. *A Semantically Rich Approach for Collaborative Model Edition*, in "Symp. on Applied Computing (SAC)", TaiChung, Taiwan, ACM, March 2011, vol. 26.

[53] U. MIR, S. AKNINE, L. ARANTES. *Multi-agent Coordination in Ad-hoc Networks based on Coalition Formation*, in "International Conference ICAART", Espagne, 2010, 33, http://hal.inria.fr/hal-00446101.

[54] M. C. OLESEN, R. R. HANSEN, J. L. LAWALL, N. PALIX. *Clang and Coccinelle: The best of two worlds for CERT C Secure Coding certification*, in "4th International Workshop on Foundations and Techniques for Open Source Software Certification", Pisa, Italy, September 2010.

[55] N. PALIX, J. L. LAWALL, G. MULLER. *Tracking code patterns over multiple software versions with Herodotos*, in "9th International Conference on Aspect-Oriented Software Development (AOSD)", 2010, p. 169-180.

[56] N. PALIX, S. SAHA, G. THOMAS, C. CALVÈS, J. L. LAWALL, G. MULLER. *Faults in Linux: Ten Years Later*, in "Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)", Newport Beach, CA, USA, March 2011, To appear.

[57] T. PREUDHOMME, J. SOPENA, G. THOMAS, B. FOLLIOT. *BatchQueue: Fast and Memory-Thrifty Core to Core Communication*, in "22nd International Symposium on Computer Architecture and High Performance Computing (SBAC 2010), IFIP/IEEE", 2010.

[58] S. SAHA, J. L. LAWALL, G. MULLER. *An Approach to Improving the Structure of Error-Handling Code in the Linux Kernel*, in "ACM SIGPLAN/SIGBED Conference on Languages, Compilers, Tools and Theory for Embedded Systems (LCTES)", Chicago, IL, USA, April 2011, To appear.

[59] N. SCHIPER, P. SUTRA, F. PEDONE. *P-Store: Genuine Partial Replication in Wide Area Networks*, in "Symp. on Reliable Dist. Sys. (SRDS)", New Dehli, India, IEEE Comp. Society, October 2010, p. 214–224.

[60] M. VALERO, L. ARANTES, M. G. POTOP-BUTUCARU, P. SENS. *Dynamically Reconfigurable Filtering Architectures*, in "12th International Symposium Stabilization, Safety, and Security of Distributed Systems (SSS)", 2010, p. 504-518.

[61] M. VALERO, L. ARANTES, M. G. POTOP-BUTUCARU, P. SENS. *Enhancing fault tolerance of distributed Rtree*, in "The Fifth Latin-American Symposium on Dependable Computing (LADC)", 2011.

National Peer-Reviewed Conference/Proceedings

[62] S. DUBOIS, T. MASUZAWA, S. TIXEUIL. *Construction auto-stabilisante d'arbre couvrant en dépit d'actions malicieuses*, in "12èmes Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications (AlgoTel)", France Belle Dune, 2010, http://hal.inria.fr/inria-00477480.

Scientific Books (or Scientific Book chapters)

[63] L. ARANTES, F. GREVE, P. SENS. *Unreliable Failure Detectors for Mobile Ad-hoc Networks*, in "Handbook of Research on Mobility and Computing: Evolving Technologies and Ubiquitous Impacts", Editors: Maria Manuela Cruz-Cunha, IGI Global, 2010, 20.

Books or Proceedings Editing

[64] A. K. DATTA, F. PETIT, R. GUERRAOUI (editors). *Journal of Theoretical Computer Science, Special Issue on Stabilization, Safety, and Security of Distributed Systems*, Elsevier, 2011.

[65] A. K. DATTA, F. PETIT, R. GUERRAOUI (editors). *Transactions on Autonomous and Adaptive Systems, Special Issue on Stabilization, Safety, and Security of Distributed Systems*, ACM, 2011.

[66] C. MORIN, G. MULLER (editors). *European Conference on Computer Systems, Proceedings of the 5th European confer ence on Computer systems, EuroSys 2010, Paris, France, April 13-16, 2010*, ACM, 2010.

Research Reports

[67] L. BLIN, S. DOLEV, M. GRADINARIU POTOP-BUTUCARU, S. ROVEDAKIS. *Fast Self-Stabilizing Minimum Spanning Tree Construction*, INRIA, 2010, http://hal.inria.fr/hal-00492398.

[68] L. BLIN, A. MILANI, M. GRADINARIU POTOP-BUTUCARU, S. TIXEUIL. *Exclusive perpetual ring exploration without chirality*, INRIA, Mar 2010, http://hal.inria.fr/inria-00464206.

[69] L. BLIN, S. ROVEDAKIS, M. GRADINARIU POTOP-BUTUCARU, S. TIXEUIL. *Universal Loop-Free Super-Stabilization*, INRIA, 2010, http://hal.inria.fr/hal-00492320.

[70] Z. BOUZID, S. DOLEV, M. GRADINARIU POTOP-BUTUCARU, S. TIXEUIL. *RoboCast: Asynchronous Communication in Robot Networks*, INRIA, Apr 2010, http://hal.inria.fr/inria-00496379.

[71] S. DUBOIS, T. MASUZAWA, S. TIXEUIL. *On Byzantine Containment Properties of the* $min + 1$ *Protocol*, INRIA, May 2010, http://hal.inria.fr/inria-00487091.

[72] S. DUBOIS, T. MASUZAWA, S. TIXEUIL. *Self-stabilization with Byzantine tolerance for global tasks*, INRIA, May 2010, http://hal.inria.fr/inria-00484645.

[73] S. DUBOIS, T. MASUZAWA, S. TIXEUIL. *The Impact of Topology on Byzantine Containment in Stabilization*, INRIA, May 2010, http://hal.inria.fr/inria-00481836.

[74] F. GAUD, S. GENEVÈS, R. LACHAIZE, B. LEPERS, F. MOTTET, G. MULLER, V. QUÉMA. *Mely: Efficient Workstealing for Multicore Event-Driven Systems*, INRIA, Jan 2010, RR-7169, http://hal.inria.fr/inria-00449530.

[75] A. LAMANI, A. COURNIER, S. DUBOIS, F. PETIT, V. VILLAIN. *Snap-Stabilizing Linear Message Forwarding*, INRIA, Jun 2010, http://hal.inria.fr/inria-00492870.

[76] N. PALIX, S. SAHA, G. THOMAS, C. CALVÈS, J. L. LAWALL, G. MULLER. *Faults in Linux: Ten Years Later*, INRIA, Aug 2010, RR-7357, http://hal.inria.fr/inria-00509256.

[77] M. SANTOS DE LIMA, F. GREVE, L. ARANTES, P. SENS. *Byzantine Failure Detection for Dynamic Distributed Systems*, INRIA, Mar 2010, RR-7222, http://hal.inria.fr/inria-00461518.

[78] P. SUTRA, M. SHAPIRO. *Fast Genuine Generalized Consensus*, INRIA, 2010, http://hal.inria.fr/hal-00476885.

Other Publications

[79] J. L. LAWALL, R. R. HANSEN, N. PALIX, G. MULLER. *Improving the security of infrastructure software using Coccinelle*, October 2010, n^{o} 83.