IN PARTNERSHIP WITH:
CNRS

Université Pierre et Marie Curie (Paris 6)

Activity Report 2016

Project-Team REGAL

Large-Scale Distributed Systems and Applications

IN COLLABORATION WITH: Laboratoire d'informatique de Paris 6 (LIP6)

RESEARCH CENTER
Paris

THEME
Distributed Systems and middleware

Table of contents

Project-Team REGAL

Creation of the Project-Team: 2005 July 01

Keywords:

Computer Science and Digital Science:

1.1.1. - Multicore
1.1.6. - Cloud
1.1.7. - Peer to peer
1.1.9. - Fault tolerant systems
1.1.13. - Virtualization
1.3. - Distributed Systems
1.6. - Green Computing
2.6. - Infrastructure software
2.6.1. - Operating systems
2.6.2. - Middleware
2.6.3. - Virtual machines
3.1.3. - Distributed data
3.1.8. - Big data (production, storage, transfer)
7.1. - Parallel and distributed algorithms

Other Research Topics and Application Domains:

4.5. - Energy consumption
6.4. - Internet of things
8.2. - Connected city
9.2.3. - Video games
9.4.1. - Computer science

1. Members

Research Scientists

Mesaac Makpangou [Inria, Researcher, HDR]
Marc Shapiro [Inria, Senior Researcher, HDR]

Faculty Members

Pierre Sens [Team leader, Univ. Paris VI, Professor, HDR]
Luciana Bezerra Arantes [Univ. Paris VI, Associate Professor]
Swan Dubois [Univ. Paris VI, Associate Professor]
Jonathan Lejeune [Univ. Paris VI, Associate Professor, since Sep 2016]
Sebastien Monnet [Univ. Paris VI, Associate Professor, until Sep 2016]
Franck Petit [Univ. Paris VI, Professor, HDR]
Julien Sopena [Univ. Paris VI, Associate Professor]

Engineers

Salvatore Pileggi [Inria, unded by FP7 SyncFree, until Sep 2016]
Tyler Crain [Inria, post-doc funded by FP7 SyncFree, until Sep 2016]

PhD Students

Sébastien Bouchard [Inria, from Oct 2016]

Marjorie Bournat [Univ. Paris VI]
Michael Damien Carver [Magency, under CIFRE grant]
Joao Paulo de Araujo [Univ. Paris VI]
Lyes Hamidouche [Magency, under CIFRE grant]
Denis Jeanneau [Univ. Paris VI]
Mahsa Najafzadeh [Inria, until May 2016]
Vinh Tao Thanh [Scality, under CIFRE grant]
Alejandro Tomsic [Inria]
Guillaume Turchini [Univ. Paris VI]
Gauthier Voron [Univ. Paris VI]
Bassirou Ngom [Univ. Cheikh Anta Diop de Dakar / Univ. Paris VI, until Sep 2016]

Visiting Scientist
Rostom Mennour [Univ. Constantine]

Administrative Assistant
Helene Milome [Inria]

2. Overall Objectives

2.1. Overall Objectives

The research of the Regal team addresses the theory and practice of *Computer Systems,* including multicore computers, clusters, networks, peer-to-peer systems, cloud computing systems, and other communicating entities such as swarms of robots. It addresses the challenges of communicating, sharing information, and computing correctly in such large-scale, highly dynamic computer systems. This includes addressing the core problems of communication, consensus and fault detection, scalability, replication and consistency of shared data, information sharing in collaborative groups, dynamic content distribution, and multi- and many-core concurrent algorithms.

Regal is a joint research team between LIP6 and Inria Paris.

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.

4. Highlights of the Year

4.1. Highlights of the Year

- We initiate a collaboration with ICL Lab (University of Tennessee) to study failure detection in Exascale computing. We designed and evaluated a new robust failure detector. This result is published at SC 2016 [26].

5. New Software and Platforms

5.1. Antidote

Functional Description

Antidote is the flexible cloud database platform currently under development in the SyncFree European project. Antidote aims to be both a research platform for studying replication and consistency at the large scale, and an instrument for exploiting research results. The platform supports replication of CRDTs, in and between sharded (partitioned) data centres (DCs). The current stable version supports strong transactional consistency inside a DC, and causal transactional consistency between DCs. Ongoing research includes support for explicit consistency [23] , for elastic version management, for adaptive replication, for partial replication, and for reconfigurable sharding.

- Participants: Tyler Crain, Marc Shapiro and Alejandro Tomsic
- Contact: Tyler Crain
- URL: https://github.com/SyncFree/

5.2. G-DUR

Functional Description

A large family of distributed transactional protocols have a common structure, called Deferred Update Replication (DUR). DUR provides dependability by replicating data, and performance by not re-executing transactions but only applying their updates. Protocols of the DUR family differ only in behaviors of few generic functions. Based on this insight, we offer a generic DUR middleware, called G-DUR, along with a library of finely-optimized plug-in implementations of the required behaviors.

- Participants: Marc Shapiro and Masoud Saeida Ardekani
- Contact: Marc Shapiro
- URL: https://github.com/msaeida/jessy

5.3. NumaGIC

FUNCTIONAL DESCRIPTION

NumaGiC is a version of the HotSpot garbage collector (GC) adapted to many-core computers with very large main memories. In order to maximise GC throughput, it manages the trade-off between memory locality (local scans) and parallelism (work stealing) in a self-balancing manner. Furthemore, the collector features several memory placement heuristics that improve locality.

- Participants: Lokesh Gidra, Marc Shapiro, Julien Sopena and Gaël Thomas
- Contact: Lokesh Gidra
- URL: http://gforge.inria.fr/projects/transgc/

6. New Results

6.1. Distributed Algorithms for Dynamic Networks and Fault Tolerance

Participants: Luciana Bezerra Arantes [correspondent], Sébastien Bouchart, Marjorie Bournat, Swan Dubois, Denis Jeanneau, Mohamed Hamza Kaaouachi, Sébastien Monnet, Franck Petit [correspondent], Pierre Sens, Julien Sopena.

Nowadays, distributed systems are more and more heterogeneous and versatile. Computing units can join, leave or move inside a global infrastructure. These features require the implementation of *dynamic* systems, that is to say they can cope autonomously with changes in their structure in terms of physical facilities and software. It therefore becomes necessary to define, develop, and validate distributed algorithms able to managed such dynamic and large scale systems, for instance mobile *ad hoc* networks, (mobile) sensor networks, P2P systems, Cloud environments, robot networks, to quote only a few.

The fact that computing units may leave, join, or move may result of an intentional behavior or not. In the latter case, the system may be subject to disruptions due to component faults that can be permanent, transient, exogenous, evil-minded, etc. It is therefore crucial to come up with solutions tolerating some types of faults.

We address both system dynamic and fault tolerance through various aspects: (1) Fault Detection, (2) Self-Stabilization, and (3) Dynamic System Design. Our approach covers the whole spectrum from theory to experimentation. We design algorithms, prove them correct, implement them, and evaluate them within simulation plateforms.

6.1.1. Failure detection

Since 2013, we address both theoretical and practical aspects of failure detector. The failure detector (FD) abstraction has been used to solve agreement problems in asynchronous systems prone to crash failures, but so far it has mostly been used in static and complete networks. FDs are distributed oracles that provide processes with unreliable information on process failures, often in the form of a list of trusted process identities. In 2016 we obtain the following results.

We propose in [31] a new failure detector that expresses the confidence with regard to the system as a whole. Similarly to a reputation approach, it is possible to indicate the relative importance of each process of the system, while a threshold offers a degree of flexibility for failures and false suspicions. Performance evaluation results, based on real PlanetLab traces, confirm the degree of flexible of the failure detector. By logically organizing nodes in a distributed hypercube, denoted VCube, which dynamically re-organizes itself in case of node failures, detected by a hierarchical perfect failure, we have proposed a autonomic distributed quorum algorithm [35]. By replacing the perfect failure detector by another one that offers eventual strong completeness, we have presented in [33] a second autonomic reliable broadcast protocol.

In the context of large networks, we propose Internet Failure Detector Service (IFDS) [16] for processes running in the Internet on multiple autonomous systems. The failure detection service is adaptive, and can be easily integrated into applications that require configurable QoS guarantees. The service is based on monitors which are capable of providing global process state information through a SNMP MIB. Monitors at different networks communicate across the Internet using Web Services. The system was implemented and evaluated for monitored processes running both on single LAN and on PlanetLab. Experimental results are presented, showing the performance of the detector, in particular the advantages of using the self-tuning strategies to address the requirements of multiple concurrent applications running on a dynamic environment.

Finally, in collaboration with ICL Lab. (University of Tennessee), we study failure detection in the context of ExaScale computing. We designed and evaluated a new robust failure detector, able to maintain and distribute the correct list of alive resources within proven and scalable bounds. The detection and distribution of the fault information follow different overlay topologies that together guarantee minimal disturbance to the applications. A virtual observation ring minimizes the overhead by allowing each node to be observed by another single node, providing an unobtrusive behavior. The propagation stage is using a non-uniform variant of a reliable broadcast over a circulant graph overlay network, and guarantees a logarithmic fault propagation. Extensive simulations, together with experiments on the Titan ORNL supercomputer, show that the algorithm performs extremely well, and exhibits all the desired properties of an Exascale-ready algorithm. This work has been published at SC 2016 conference [26].

6.1.2. Self-Stabilization

Regardless its initial state, a *self-stabilizing* system has the ability to reach a correct behavior in finite time. Self-stabilization is a generic paradigm to tolerate transient faults (*i.e.*, faults of finite duration) in distributed systems. Self-stabilization is also a suitable approach to design reliable solutions for dynamic systems. Results obtained in this area by Regal members in 2016 follow.

In [8], we address the ability to maintain distributed structures at large scale. Among the many different structures proposed in this context, The prefix tree structure is a good candidate for indexing and retrieving information. One weakness of using such a distributed structure stands in its poor native fault tolerance, leading to the use of preventive costly mechanisms such as replication. We focus on making tries self-stabilizing over such platforms, and propose a self-stabilizing maintenance algorithm for a prefix tree using a message passing model. The proof of self-stabilization is provided, and simulation results are given, to better capture its performances.

In [4], we propose a silent self-stabilizing leader election algorithm for bidirectional connected identified networks of arbitrary topology. Written in the locally shared memory model, it assumes the distributed unfair daemon, *i.e.*, the most general scheduling hypothesis of the model. Our algorithm requires no global knowledge on the network (such as an upper bound on the diameter or the number of processes). We show that its stabilization time is in $\Theta(n^3)$ steps in the worst case, where n is the number of processes. Its memory requirement is asymptotically optimal, *i.e.*, $\Theta(\log n)$ bits per processes. Its round complexity is of the same order of magnitude — *i.e.*, $\Theta(n)$ rounds — as the best existing algorithms designed with similar settings. To the best of our knowledge, this is the first asynchronous self-stabilizing leader election algorithm for arbitrary identified networks that is proven to achieve a stabilization time polynomial in steps. By contrast, we show that the previous best existing algorithms stabilize in a non polynomial number of steps in the worst case.

A *snap-stabilizing* protocol, regardless of the initial configuration of the system, guarantees that it always behaves according to its specification. In [9], we consider the locally shared memory model. In this model, we propose a snap-stabilizing Propagation of Information with Feedback (PIF) protocol for rooted networks of arbitrary topology. Then, we use the proposed PIF protocol as a key module in the design of snap-stabilizing solutions for some fundamental problems in distributed systems, such as Leader Election, Reset, Snapshot, and Termination Detection. Finally, we show that in the locally shared memory model, snap-stabilization is as expressive as self-stabilization by designing a universal transformer to provide a snap-stabilizing version of any protocol that can be (automatically) self-stabilized. Since by definition, a snap-stabilizing algorithm is self-stabilizing, self- and snap-stabilization have the same expressiveness in the locally shared memory model.

In [6], we address the *committee coordination problem*: A committee consists of a set of professors and committee meetings are synchronized, so that each professor participates in at most one committee meeting at a time. We propose two snap-stabilizing distributed algorithms for the committee coordination. They are enriched with some desirable properties related to concurrency, (weak) fairness, and a stronger synchronization mechanism called 2-Phase Discussion. Existing work in the literature has shown that (1) in general, fairness cannot be achieved in committee coordination, and (2) it becomes feasible if each professor waits for meetings infinitely often. Nevertheless, we show that even under this latter assumption, it is impossible to implement a fair solution that allows maximal concurrency. Hence, we propose two orthogonal snap-stabilizing algorithms, each satisfying 2-phase discussion, and either maximal concurrency or fairness.

6.1.3. Dynamic Distributed Systems

In [19], we introduce the notion of *gradually stabilizing* algorithm as any self-stabilizing algorithm with the following additional feature: if at most τ *dynamic steps*—a dynamic step is a step containing topological changes—occur starting from a legitimate configuration, it first quickly recovers to a configuration from which a minimum quality of service is satisfied and then gradually converges to stronger and stronger safety guarantees until reaching a legitimate configuration again. We illustrate this new property by proposing a gradually stabilizing unison algorithm, that consists in synchronizing logical clocks locally maintained by the processes.

The next results consider highly dynamic distributed systems modelled by time-varying graphs (TVGs). In [7], we first address proof of impossibility results that often use informal arguments about convergence. We provide a general framework that formally proves the convergence of the sequence of executions of any deterministic algorithm over TVGs of any convergent sequence of TVGs. Next, we focus of the weakest class of long-lived TVGs, *i.e.*, the class of TVGs where any node can communicate any other node infinitely often. We illustrate the relevance of our result by showing that no deterministic algorithm is able to compute various distributed covering structure on any TVG of this class. Namely, our impossibility results focus on the eventual footprint, the minimal dominating set and the maximal matching problems.

We also study the k-set agreement problem, a generalization of the consensus problem where processes can decide up to k different values. Very few papers have tackled this problem in dynamic networks. Exploiting the formalism of TVGs, we propose in [11] a new quorum-based failure detector for solving k-set agreement in dynamic networks with asynchronous communications. We present two algorithms that implement this new failure detector using graph connectivity and message pattern assumptions. We also provide an algorithm for solving k-set agreement using our new failure detector.

Finally, in [22], we deal with the classical problem of exploring a ring by a cohort of synchronous robots. We focus on the perpetual version of this problem in which it is required that each node of the ring is visited by a robot infinitely often. We assume that the robots evolve in ring-shape TVGs, *i.e.*, the static graph made of the same set of nodes and that includes all edges that are present at least once over time forms a ring of arbitrary size. We also assume that each node is infinitely often reachable from any other node. In this context, we aim at providing a self-stabilizing algorithm to the robots (*i.e.*, the algorithm must guarantee an eventual correct behavior regardless of the initial state and positions of the robots). We show that this problem is deterministically solvable in this harsh environment by providing a self-stabilizing algorithm for three robots.

6.2. Large scale data distribution

Participants: Luciana Arantes [correspondent], Rudyar Cortes, Mesaac Makpangou, Sébastien Monnet, Pierre Sens.

The proliferation of GPS-enabled devices leads to the massive generation of geotagged data sets recently known as Big Location Data. It allows users to explore and analyse data in space and time, and requires an architecture that scales with the insertions and location-temporal queries workload from thousands to millions of users. Most large scale key-value data storage solutions only provide a single one-dimensional index which does not natively support efficient multidimensional queries. In 2016, we propose GeoTrie [29], a scalable architecture built by coalescing any number of machines organized on top of a Distributed Hash Table. The key idea of our approach is to provide a distributed global index which scales with the number of nodes and provides natural load balancing for insertions and location-temporal range queries. We assess our solution using the largest public multimedia data set released by Yahoo! which includes millions of geotagged multimedia files.

We also propose ECHO [10], a novel and lightweight solution that efficiently supports range queries over a ring-like Distributed Hash Table (DHT) structure. By implementing a tree-based index structure and an effective query routing strategy, ECHO provides low-latency and low-overhead query searches by exploiting the Tabu Search principle. Load balancing is also improved reducing the traditional bottleneck problems arising in upper level nodes of tree-based index structures such as PHT. Furthermore, ECHO copes with DHT churn problems as its index exploits logical information as opposed to static reference cache approaches or replication techniques. The performance evaluation results obtained using PeerSim simulator show that ECHO achieves efficient performance compared other solutions such as the PHT strategy and its optimized version which includes a query cache.

6.3. Consistency protocols

Participants: Marc Shapiro [correspondent], Tyler Crain, Mahsa Najafzadeh, Marek Zawirski, Alejandro Tomsic.

6.3.1. Static Reasoning About Consistency, and associated tools

Large-scale distributed systems often rely on replicated databases that allow a programmer to request different data consistency guarantees for different operations, and thereby control their performance. Using such databases is far from trivial: requesting stronger consistency in too many places may hurt performance, and requesting it in too few places may violate correctness. To help programmers in this task, we propose the first proof rule for establishing that a particular choice of consistency guarantees for various operations on a replicated database is enough to ensure the preservation of a given data integrity invariant. Our rule is modular: it allows reasoning about the behaviour of every operation separately under some assumption on the behaviour of other operations. This leads to simple reasoning, which we have automated in an SMT-based tool. We present a nontrivial proof of soundness of our rule and illustrate its use on several examples.

The intuition was presented at EuroSys 2015 [47]. We present the full theory and proofs in the POPL 2016 paper "'Cause I'm Strong Enough: Reasoning about Consistency Choices in Distributed Systems" [30]. The proof procedure and tool are described in PaPoC 2016 paper "The CISE Tool: Proving Weakly-Consistent Applications Correct" [34] and a YouTube video [48]. It is also the focus of Mahsa Najafzadeh's PhD thesis [3].

6.3.2. Scalable consistency protocols

Developers of cloud-scale applications face a difficult decision of which kind of storage to use, summarised by the CAP theorem. Currently the choice is between classical CP databases, which provide strong guarantees but are slow, expensive, and unavailable under partition; and NoSQL-style AP databases, which are fast and available, but too hard to program against. We present an alternative: Cure provides the highest level of guarantees that remains compatible with availability. These guarantees include: causal consistency (no ordering anomalies), atomicity (consistent multi-key updates), and support for high-level data types (developer friendly API) with safe resolution of concurrent updates (guaranteeing convergence). These guarantees minimise the anomalies caused by parallelism and distribution, thus facilitating the development of applications. This paper presents the protocols for highly available transactions, and an experimental evaluation showing that Cure is able to achieve scalability similar to eventually- consistent NoSQL databases, while providing stronger guarantees.

This work is published under the title "Cure: Strong semantics meets high availability and low latency" at ICDCS 2016 [18].

6.3.3. Lightweight, correct causal consistency

Non-Monotonic Snapshot Isolation (NMSI), a variant of the widely deployed Snapshot Isolation (SI), aims at improving scalability by relaxing snapshots. In contrast to SI, NMSI snapshots are causally consistent, which allows for more parallelism and a reduced abort rate.

This work documents the design of PhysiCS-NMSI, a transactional protocol implementing NMSI in a partitioned data store. It is the first protocol to rely on a single scalar taken from a physical clock for tracking causal dependencies and building causally consistent snapshots. Its commit protocol ensures atomicity and the absence of write-write conflicts. Our PhysiCS-NMSI approach increases concurrency and reduces abort rate and metadata overhead as compared to state-of-art systems.

The paper "PhysiCS-NMSI: efficient consistent snapshots for scalable snapshot isolation" is published at PaPoC 2016 [36].

6.3.4. Reconciling consistency and scalability

Geo-replicated storage systems are at the core of current Internet services. Unfortunately, there exists a fundamental tension between consistency and performance for offering scalable geo-replication. Weakening consistency semantics leads to less coordination and consequently a good user experience, but it may introduce anomalies such as state divergence and invariant violation. In contrast, maintaining stronger consistency precludes anomalies but requires more coordination. This paper discusses two main contributions to address this tension. First, RedBlue Consistency enables blue operations to be fast (and weakly consistent) while the remaining red operations are strongly consistent (and slow). We identify sufficient conditions for determining when operations can be blue or must be red. Second, Explicit Consistency further increases the space of operations that can be fast by restricting the concurrent execution of only the operations that can break application-defined invariants. We further show how to allow operations to complete locally in the common case, by relying on a reservation system that moves coordination off the critical path of operation execution.

The paper "Geo-Replication: Fast If Possible, Consistent If Necessary" is published in the IEEE CS Data Engineering Bulletin of March 2016 [5].

6.3.5. Consistency in 3D

Comparisons of different consistency models often try to place them in a linear strong-to-weak order. However this view is clearly inadequate, since it is well known, for instance, that Snapshot Isolation and Serialisability are incomparable. In the interest of a better understanding, we propose a new classification, along three dimensions, related to: a total order of writes, a causal order of reads, and transactional composition of multiple operations. A model may be stronger than another on one dimension and weaker on another. We believe that this new classification scheme is both scientifically sound and has good explicative value. We presents the three-dimensional design space intuitively.

This work was presented as an invited keynote paper at Concur 2016 [17].

6.3.6. Scalable consistency protocols

Collaborative text editing systems allow users to concurrently edit a shared document, inserting and deleting elements (e.g., characters or lines). There are a number of protocols for collaborative text editing, but so far there has been no precise specification of their desired behavior, and several of these protocols have been shown not to satisfy even basic expectations. This work provides a precise specification of a replicated list object, which models the core functionality of replicated systems for collaborative text editing. We define a strong list specification, which we prove is implemented by an existing protocol, as well as a weak list specification, which admits additional protocol behaviors.

A major factor determining the efficiency and practical feasibility of a collaborative text editing protocol is the space overhead of the metadata that the protocol must maintain to ensure correctness. We show that for a large class of list protocols, implementing either the strong or the weak list specification requires a metadata overhead that is at least linear in the number of elements deleted from the list. The class of protocols to which this lower bound applies includes all list protocols that we are aware of, and we show that one of these protocols almost matches the bound.

This work is published at PODC 2016 [21].

6.3.7. Highly-responsive CRDTs for group editing

Group editing is a crucial feature for many end-user applications. It requires high responsiveness, which can be provided only by optimistic replication algorithms, which come in two classes: classical Operational Transformation (OT), or more recent Conflict-Free Replicated Data Types (CRDTs).

Typically, CRDTs perform better on *downstream* operations, i.e., when merging concurrent operations than OT, because the former have logarithmic complexity and the latter quadratic. However, CRDTs are often less responsive, because their *upstream* complexity is linear. To improve this, this paper proposes to interpose an auxiliary data structure, called the *identifier data structure* in front of the base CRDT. The identifier structure ensures logarithmic complexity and does not require replication or synchronization. Combined with a block-wise storage approach, this approach improves upstream execution time by several orders of magnitude, with negligeable impact on memory occupation, network bandwidth, and downstream execution performance.

This work is published at ACM Group 2016 [27].

6.4. Memory management for multicores

Participants: Antoine Blin, Damien Carver, Maxime Lorrillere, Sébastien Monnet, Julien Sopena [correspondent].

Regal co-advises with Whisper team the PhD of Antoine Blin. The thesis focusses on modern complex embedded systems that involve a mix of real-time and best effort applications. The recent emergence of low-cost multicore processors raises the possibility of running both kinds of applications on a single machine, with virtualization ensuring isolation. Nevertheless, memory contention can introduce other sources of delay, that can lead to missed deadlines. We first investigated the source of memory contention for the Mibench benchmark in a paper published at ETYS 2016 [25]. Then, in a paper published at ECRTS 2016 [24], we present a combined offline/online memory bandwidth monitoring approach. Our approach estimates and limits the impact of the memory contention incurred by the best-effort applications on the execution time of the real-time application. Using our approach, the system designer can limit the overhead on the real-time application to under 5% of its expected execution time, while still enabling progress of the best-effort applications.

Another memory management challenge for multi-cores is the fragmentation induced by the virtualized environments. Previously, we proposed Puma (for Pooling Unused Memory in Virtual Machines) which allows I/O intensive applications running on top of VMs to benefit of large caches. This was realized by providing a remote caching mechanism that provides the ability for any VM to extend its cache using the memory of other VMs located either in the same or in a different host. This work was defended by Maxime Lorrillere in April 2016 [2].

More recently, we study the memory arbitration between containers. In the Damien Carver's PhD thesis (started in October 2015), we are designing ACDC (Advanced Consolidation for Dynamic Containers), a kernel-level mechanisms that automatically provides more memory to the most active containers.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

7.1.1. Joint industrial PhD with Renault

- Renault, 2014-2016, 45 000 euros. The purpose of this contract is to develop solutions for running a mix of real-time and best-effort applications on a small embedded multicore architecture. Our goal is to optimize the usage of the processor resource. The PhD of Antoine Blin is supported by a CIFRE fellowship with Renault.

7.1.2. Joint industrial PhD with Scality SA: CRDTs for Large-Scale Storage Systems

This year, we continued the joint CIFRE (industrial PhD) research of Tao Thanh Vinh, with the French start-up company Scality, as described above (under "Large-Scale File Systems").

The objective of this research is to design new algorithms for file and block storage systems, considering both the issues of scaling the file naming tree to a very large size, and the issue of conflicting updates to files or to the name tree, in the case of high latency or disconnected work.

A new CIFRE agreement with Scality is awaiting the agreement from ANRT and will start ASAP. The PhD student is Dimitrios Vasilas, and his topic is "Scalable indexing for large-scale distributed storage systems."

7.1.3. Joint industrial PhDs: data sharing in mobile networks and automatic resizing of shared I/O caches, with Magency

Magency organizes large events during which participants can use mobile devices to access related data and interact together.

The thesis of Lyes Hamidouche concerns efficient data sharing among a large number of mobile devices. Magency brings traces captured during real events (data accesses and user mobility). We are jointly working on the design of algorithms allowing a large number of mobile devices to efficiently access remote data.

Magency also runs servers. A server is used before an event in order to be prepared and tested, and then, during the event to serve the numerous mobile devices accesses. Many servers are run on a single physical machine using containers. Using this configuration, the memory is partitioned, leading to poor performances for applications that need a large amount of memory for caching purpose. In the context of Damien Carver's PhD thesis, we are designing kernel-level mechanisms that automatically give more memory to the most active containers, leveraging the expertise acquired during Maxime Lorrillere's PhD thesis.

7.1.4. EMR CREDIT, with Thales

Franck Petit and Swan Dubois participate to the creation of the EMR (Equipe Mixte de Recherche) *CREDIT*, (Compréhension, Représentation et Exploitation Des Interactions Temporelles) between LIP6/UPMC and Thales.

Nowadays, networks are the field of temporal interactions that occur in many settings networks, including security issues. The amount and the speed of such interactions increases everyday. Until recently, the dynamics of these objects was little studied due to the lack of appropriate tools and methods. However, it becomes crucial to understand the dynamics of these interactions. Typically, how can we detect failures or attacks in network traffic, fraud in financial transactions, bugs or attacks traces of software execution. More generally, we seek to identify patterns in the dynamics of interactions. Recently, several different approaches have been proposed to study such interactions. For instance, by merging all interactions taking place over a period (e.g. one day) in a graph that are studied thereafter (evolving graphs). Another approach was to built meta-objects by duplicating entities at each unit of time of their activity, and by connecting them together.

The goal of the EMR is to join both teams of LIP6 and Thales on these issues. More specifically, we hope to make significant progress on security issues such as anomaly detection. This requires the use of a formalism sufficiently expressive to formulate complex temporal properties. Recently, a vast collection of concepts, formalisms, and models has been unified in a framework called Time-Varying Graphs. We want to pursuit that way. In the short run, the challenges facing us are: (1) refine the model to capture some interaction patterns, (2) design of algorithms to separate sequences of interactions, (3) Identify classes of entities playing a particular role in the dynamics, such as bridges between communities, or sources and sinks.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. Labex SMART - (2012–2019)

Members: ISIR (UPMC/CNRS), LIP6 (UPMC/CNRS), LIB (UPMC/INSERM), LJLL (UPMC/CNRS), LTCI (Institut Mines-Télécom/CNRS), CHArt-LUTIN (Univ. Paris 8/EPHE), L2E (UPMC), STMS (IRCAM/CNRS).

Funding: Sorbonne Universités, ANR.

Description: The SMART Labex project aims globally to enhancing the quality of life in our digital societies by building the foundational bases for facilitating the inclusion of intelligent artifacts in our daily life for service and assistance. The project addresses underlying scientific questions raised by the development of Human-centered digital systems and artifacts in a comprehensive way. The research program is organized along five axes and Regal is responsible of the axe "Autonomic Distributed Environments for Mobility."

The project involves a PhD grant of 100 000 euros over 3 years.

8.1.2. ESTATE - (2016–2020)

Members: LIP6 (Regal, project leader), LaBRI (Univ. de Bordeaux); Verimag (Univ. de Grenoble).

Funding: ESTATE is funded by ANR (PRC) for a total of about 544 000 euros, of which 233 376 euros for Regal.

Objectives: The core of ESTATE consists in laying the foundations of a new algorithmic framework for enabling Autonomic Computing in distributed and highly dynamic systems and networks. We plan to design a model that includes the minimal algorithmic basis allowing the emergence of dynamic distributed systems with self-* capabilities, *e.g.*, self-organization, self-healing, self-configuration, self-management, self-optimization, self-adaptiveness, or self-repair. In order to do this, we consider three main research streams:

(i) building the theoretical foundations of autonomic computing in dynamic systems, (ii) enhancing the safety in some cases by establishing the minimum requirements in terms of amount or type of dynamics to allow some strong safety guarantees, (iii) providing additional formal guarantees by proposing a general framework based on the Coq proof assistant to (semi-)automatically construct certified proofs.

The coordinator of ESTATE is Franck Petit.

8.1.3. RainbowFS - (2016–2020)

Members: LIP6 (Regal, project leader), Scality SA, CNRS-LIG, Télécom Sud-Paris.

Funding: is funded by ANR (PRC) for a total of 919 534 euros, of which 359 554 euros for Regal.

Objectives: RainbowFS proposes a "just-right" approach to storage and consistency, for developing distributed, cloud-scale applications. Existing approaches shoehorn the application design to some predefined consistency model, but no single model is appropriate for all uses. Instead, we propose tools to co-design the application and its consistency protocol. Our approach reconciles the conflicting requirements of availability and performance vs. safety: common-case operations are designed to be asynchronous; synchronisation is used only when strictly necessary to satisfy the application's integrity invariants. Furthermore, we deconstruct classical consistency models into orthogonal primitives that the developer can compose efficiently, and provide a number of tools for quick, efficient and correct cloud-scale deployment and execution. Using this methodology, we will develop an entreprise-grade, highly-scalable file system, exploring the rainbow of possible semantics, and we demonstrate it in a massive experiment.

The coordinator of RainbowFS is Marc Shapiro.

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

8.2.1.1. SyncFree

Title: Large-scale computation without synchronisation

Programm: FP7

Duration: October 2013 – December 2016

Coordinator: Inria

Partners:

Basho Technologies (United Kingdom)

Faculdade de Ciencias E Tecnologiada Universidade Nova de Lisboa (Portugal)

Koc University (Turkey)

Rovio Entertainment OY (Finland)

Trifork AS (Denmark)

Université Catholique de Louvain (Belgium)

Technische Universitaet Kaiserslautern (Germany)

Erlang Solutions Ltd (United Kingdom).

Inria contact: Marc Shapiro

The goal of SyncFree is to enable large-scale distributed applications without global synchronisation, by exploiting the recent concept of Conflict-free Replicated Data Types (CRDTs). CRDTs allow unsynchronised concurrent updates, yet ensure data consistency. This revolutionary approach maximises responsiveness and availability; it enables locating data near its users, in decentralised clouds.

Global-scale applications, such as virtual wallets, advertising platforms, social networks, online games, or collaboration networks, require consistency across distributed data items. As networked users, objects, devices, and sensors proliferate, the consistency issue is increasingly acute for the software industry. Current alternatives are both unsatisfactory: either to rely on synchronisation to ensure strong consistency, or to forfeit synchronisation and consistency altogether with ad-hoc eventual consistency. The former approach does not scale beyond a single data centre and is expensive. The latter is extremely difficult to understand, and remains error-prone, even for highly-skilled programmers.

SyncFree avoids both global synchronisation and the complexities of ad-hoc eventual consistency by leveraging the formal properties of CRDTs. CRDTs are designed so that unsynchronised concurrent updates do not conflict and have well-defined semantics. By combining CRDT objects from a standard library of proven datatypes (counters, sets, graphs, sequences, etc.), large-scale distributed programming is simpler and less error-prone. CRDTs are a practical and cost-effective approach.

The SyncFree project will develop both theoretical and practical understanding of large-scale synchronisation-free programming based on CRDTs. Project results will be new industrial applications, new application architectures, large-scale evaluation of both, programming models and algorithms for large-scale applications, and advanced scientific understanding.

8.2.1.2. LightKone

Title: Lightweight Computation for Networks at the Edge

Programm: H2020-ICT-2016-2017

Duration: January 2017 - December 2019

Coordinator: Université Catholique de Louvain

Partners:

Université Catholique de Louvain (Belgium)

Technische Universitaet Kaiserslautern (Germany)

INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciencia (Portugal)

Faculdade de Ciencias E Tecnologiada Universidade Nova de Lisboa (Portugal)

Universitat Politecnica De Catalunya (Spain)

Scality (France)

Gluk Advice B.V. (Netherlands)

Inria contact: Marc Shapiro

The goal of LightKone is to develop a scientifically sound and industrially validated model for doing general-purpose computation on edge networks. An edge network consists of a large set of heterogeneous, loosely coupled computing nodes situated at the logical extreme of a network. Common examples are networks of Internet of Things, mobile devices, personal computers, and points of presence including Mobile Edge Computing. Internet applications are increasingly running on edge networks, to reduce latency, increase scalability, resilience, and security, and permit local decision making. However, today's state of the art, the gossip and peer-to-peer models, give no solution for defining general-purpose computations on edge networks, i.e., computation with shared mutable state. LightKone will solve this problem by combining two recent advances in distributed computing, namely synchronisation-free programming and hybrid gossip algorithms, both of which are successfully used separately in industry. Together, they are a natural combination for edge computing. We will cover edge networks both with and without data center nodes, and applications focused on collaboration, computation, and both. Project results will be new programming models and algorithms that advance scientific understanding, implemented in new industrial applications and a startup company, and evaluated in large-scale realistic settings.

8.3. International Initiatives

8.3.1. Inria International Labs

Inria Chile
Associate Team involved in the International Lab:

8.3.1.1. ARMADA

Title: hARnessing MAssive DAta flows

International Partner (Institution - Laboratory - Researcher):

Universidad Tecnica Federico Santa Maria (Chile) - Department of Computer Science (Department of Comput) - Xavier Bonnaire

Start year: 2014

See also: http://web.inria-armada.org

The ARMADA project aims at designing and implementing a reliable framework for the management and processing of massive dynamic dataflows. The project is two-pronged: fault-tolerant middleware support for processing massive continuous input, and a redundant storage service for mutable data on a massive scale.

8.3.2. Participation in Other International Programs

8.3.2.1. CNRS-Inria-FAP's

Title: Autonomic and Scalable Algorithms for Building Resilient Distributed Systems

International Partner (Institution - Laboratory - Researcher):

Universida de Federal do Paraná (UFPR), Brazil, Prof. Elias Duarte

Duration: 2015–2017

In the context of autonomic computing systems that detect and diagnose problems, self-adapting themselves, the VCube (Virtual Cube), proposed by Prof. Elias Duarte , is a distributed diagnosis algorithm that organizes the system nodes on a virtual hypercube topology. VCube has logarithmic properties: when all nodes are fault-free, processes are virtually connected to form a perfect hypercube; as soon as one or more failures are detected, links are automatically reconnected to remove the faulty nodes and the resulting topology, connecting only fault-free nodes, keeps the logarithmic properties. The goal of this project is to exploit the autonomic and logarithmic properties of the VCube by proposing self-adapting and self-configurable services.

8.3.2.2. Capes-Cofecub

Title: CHOOSING - Cooperation on Hybrid cOmputing clOuds for energy SavING

French Partners: Paris XI (LRI), Regal, LIG, SUPELEC

International Partners (Institution - Laboratory - Researcher):

Universidade de São Paulo - Instituto de Matemática e Estatística - Brazil, Unicamp - Instituto de Computação - Brazil

Duration: 2014–2018

The cloud computing is an important factor for environmentally sustainable development. If, in the one hand, the increasing demand of users drive the creation of large datacenters, in the other hand, cloud computing's "multitenancy" trait allows the reduction of physical hardware and, therefore, the saving of energy. Thus, it is imperative to optimize the energy consumption corresponding to the datacenter's activities. Three elements are crucial on energy consumption of a cloud platform: computation (processing), storage and network infrastructure. Therefore, the aim of this project is to provide different techniques to reduce energy consumption regarding these three elements. Our work mainly focuses on energy saving aspects based on virtualization, i.e., pursuing the idea of the intensive migration of classical storage/processing systems to virtual ones. We will study how different organizations (whose resources are combined as hybrid clouds) can cooperate with each other in order to minimize the energy consumption without the detriment of client requirements or quality of service. Then, we intend to propose efficient algorithmic solutions and design new coordination mechanisms that incentive cloud providers to collaborate.

8.3.2.3. Spanish research ministry project

Title: BFT-DYNASTIE - Byzantine Fault Tolerance: Dynamic Adaptive Services for Partitionable Systems

French Partners: Labri, Irisa, LIP6

International Partners (Institution - Laboratory - Researcher):

University of the Basque Country UPV - Spain, EPFL - LSD - Switzerland, Friedrich-Alexander-Universitat Erlangen-Nurenberg - Deutschland, University of Sydney - Australia

Duration: 2017–2019

The project BFT-DYNASTIE is aimed at extending the model based on the alternation of periods of stable and unstable behavior to all aspects of fault-tolerant distributed systems, including synchrony models, process and communication channel failure models, system membership, node mobility, and network partitioning. The two main and new challenges of this project are: the consideration of the most general and complex to address failure model, known as Byzantine, arbitrary or malicious, which requires qualified majorities and the use of techniques form the security area; and the operation of the system in partitioned mode, which requires adequate reconciliation mechanisms when two partitions merge.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

Ajoy Kumar Datta

Date: May 2016 - June 2016

Institution: University of Nevada, Las Vegas (USA)

João Barreto

Date: April 2016 - September 2016

Institution: Instituto Superior Técnico, Lisbon, INESC-ID (Portugal)

8.4.1.1. Internships

Alvarez Colombo Santiago Javier

Date: Jul 2015 - Jan 2016

Institution: Universidad de Buenos Aires (Argentina)

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Swan Dubois: Co-Chair of the first Workshop DGDC, collocated with DISC 2016

Franck Petit: General Chair of the 17^{th} International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2016), ed. LNCS, Lyon, France.

9.1.1.2. Member of the Organizing Committees

Swan Dubois: Member of the Organizing Committee and Publicity Chair of the 30^{th} International Symposium on DIStributed Computing (DISC 2016)

9.1.2. Scientific Events Selection

9.1.2.1. Member of Steering Committee

- Marc Shapiro, Steering Committee of OPODIS 2013–2016.

9.1.2.2. Chair of Conference Program Committees

Pierre Sens: Chair of Distributed Systems and Networks track, 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD 2016), Los Angeles, USA.

9.1.2.3. Member of the Conference Program Committees

Luciana Arantes: IFIP International Conference on Distributed Applications and Interoperable Systems (DAIS 2016), International Workshop on Recent Advances in the Dependability Assessment of Complex System (Radiance 2016), IEEE International Symposium on Network Computing and Applications (NCA 2016), 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD 2016), 7th Latin-American Symposium on Dependable Computing (LADC 2016), 18th International Conferences on High Performance Computing and Communications (HPCC 2016), 11th International Conference on Green, Pervasive and Cloud Computing (GPC-2016), 35th International Conference of the Chilean Computer Science Society (SCCC 2016).

Swan Dubois: 30^{th} International Symposium on DIStributed Computing (DISC 2016), 17^{th} International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2016), 18^{th} *Workshop on Advances in Parallel and Distributed Computational Models* (APDCM 2016), 18^{eme} Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL 2016)

Franck Petit: 18^{th} *Workshop on Advances in Parallel and Distributed Computational Models* (APDCM 2016)

Pierre Sens: 27th International Symposium on Software Reliability Engineering (ISSRE 2016), 12th European Dependable Computing Conference (EDCC 2016) 30th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2016), IEEE International Symposium on Network Computing and Applications (NCA 2016).

Marc Shapiro: Int. Conf. on Architectural Support for Programming Languages and Systems (ASPLOS 2016), Int. Conf. on Principles of Distributed Systems (OPODIS 2016), Selection committee of Enterprise Research Challenge Contest (ERCC) at Int. Symp. on Stabilization, Safety, and Security of Dist. Sys. (SSS 2016), W. on Planet-Scale Distributed Systems (W-PSDS 2016).

9.1.2.4. Reviewer

Swan Dubois: 20th International Conference on Principles of Distributed Systems (OPODIS 2016), 4th International Conference on Networked Systems (NETYS 2016)

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

Franck Petit: co-editor of the proceedings of the 18th International Symposium on Stabilization, Safety, and Security of Distributed Systems, LNCS 10083.

Pierre Sens: Associate editor of International Journal of High Performance Computing and Networking (IJHPCN)

9.1.3.2. Reviewer - Reviewing Activities

Swan Dubois: ACM Transactions on Storage (TOS), SIAM Journal on Computing (SICOMP)

9.1.4. Invited Talks

Pierre Sens. Resource Management in Large Scale Distributed Systems, Dec. 2016, University of Santiago

Marc Shapiro. KTH, Stockholm, Sweden, Nov. 2016. AntidoteDB: a Planet-Scale Available Transactional Database With Strong Semantics

Marc Shapiro. Paris-Descartes University, Nov. 2016. Just-Right Consistency

Nuno Preguiça. Invariant-Preserving Applications for Weakly-consistent Replicated Databases. OPODIS PC workshop, October 2016, Lugano, Switzerland.

Marc Shapiro. Consistency in 3D. OPODIS PC workshop, October 2016, Lugano, Switzerland.

Marc Shapiro. "Consistency in 3D" [17]. Invited talk at Int. Conf. on Concurrency Theory (CONCUR), Québec, Canada, Aug. 2016.

Marc Shapiro. "Just-Right Consistency: Static analysis for minimal synchronisation." Technical talk at Critéo, Paris, France, March 2016.

Marc Shapiro. Keynote Speach at Laboratoire d'Informatique de Grenoble. "Just-Right Consistency: Static analysis for minimal synchronisation." March 2016.

Marc Shapiro, "Reconciling performance and safety in the Antidote widely-geo-replicated database," invited talk at GDR RSD and ASF Winter School 2016, le Pleynet les Sept Laux, France, March 2016.

Marc Shapiro. Invited talk at Inria/EPFL Workshop "Consistency in three dimensions (It's the invariants, stupid)." Rennes, Jan. 2016.

9.1.5. Scientific Expertise

Franck Petit, 2016: Expertise for HCERES

9.1.6. Research Administration

Franck Petit, 2014-2018: Deputy Director of LIP6

Franck Petit, since 2012: Member of the Executive Committee of Labex SMART, Co-Chair (with P. Sens) of Track 4, Autonomic Distributed Environments for Mobility.

Franck Petit, 2016–2020: Coordinator, ANR project ESTATE.

Pierre Sens, since 2016: Member of Section 6 of the national committee for scientific research CoNRS

Pierre Sens, since 2012: Member of the Executive Committee of Labex SMART, Co-Chair (with F. Petit) of Track 4, Autonomic Distributed Environments for Mobility.

Pierre Sens, 2012–2016: Member of the scientific council of UPMC

Pierre Sens, since 2015, officer at scientific research vice presidency UPMC

Pierre Sens, since 2014: Member of Steering Committee of International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD).

Marc Shapiro, since 2009: Member of Scientific Advisory Board of the Center of Informatics and Information Technology (CITI), then Lab. for Informatics and Comp. Sc. (NOVA-LINCS) Universidade Nova de Lisboa, Portugal.

Marc Shapiro, since 2012: Member ERC (European Research Council) Consolidator Grants panel PE6.

Marc Shapiro, since 2013: member of the board of *Société Informatique de France*. Vice-Chair for Research since 2016.

Marc Shapiro, since 2016: Member of Advisory Board of IEEE Computer Society Special Technical Community on Operating Systems (STCOS).

Marc Shapiro, since 2014: Member of Steering Committee of Int. Conf. on Principles of Distributed Systems (OPODIS).

Marc Shapiro, 2016–2020: Coordinator, ANR project RainbowFS.

Marc Shapiro, 2013–2016: Coordinator, FP7 project SyncFree.

Julien Sopena, Member of "Directoire des formations et de l'insertion professionnelle" of UPMC Sorbonne Universités, France

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Julien Sopena is Member of "Directoire des formations et de l'insertion professionnelle" of UPMC Sorbonne Universités, France

Master: Julien Sopena is responsible of Computer Science Master's degree in Distributed systems and applications (in French, SAR), UPMC Sorbonne Universités, France

Master: Luciana Arantes, Swan Dubois, Sébastien Monnet, Franck Petit, Pierre Sens, Julien Sopena, Advanced distributed algorithms, M2, UPMC Sorbonne Universités, France

Master: Jonathan Lejeune, Designing Large-Scale Distributed Applications, M2, UPMC Sorbonne Universités, France

Master: Maxime Lorrillere, Julien Sopena, Linux Kernel Programming, M1, UPMC Sorbonne Universités, France

Master: Luciana Arantes, Sébastien Monnet, Pierre Sens, Julien Sopena, Operating systems kernel, M1, UPMC Sorbonne Universités, France

Master: Luciana Arantes, Swan Dubois, System distributed Programming, M1, UPMC Sorbonne Universités, France

Master: Luciana Arantes, Swan Dubois, Franck Petit, Distributed Algorithms, M1, UPMC Sorbonne Universités, France

Master: Sébastien Monnet, Julien Sopena, Client-server distributed systems, M1, UPMC Sorbonne Universités, France

Licence: Pierre Sens, Luciana Arantes, Julien Sopena, Principles of operating systems, L3, UPMC Sorbonne Universités, France

Licence: Swan Dubois, Initiation to operating systems, L3, UPMC Sorbonne Universités, France

Licence: Swan Dubois, Franck Petit, Advanced C Programming, L2, UPMC Sorbonne Universités, France

Licence: Swan Dubois, Sébastien Monnet, Introduction to operating systems, L2, UPMC Sorbonne Universités, France

Licence: Mesaac Makpangou, C Programming Language, 27 h, L2, UPMC Sorbonne Universités, France

Ingénieur 4ème année : Marc Shapiro, Introduction aux systèmes d'exploitation, 26 h, M1, Polytech UPMC Sorbonne Universités, France.

9.2.2. Supervision

PhD: Mohamed Hamza Kaaouachi, "Autonomic Distributed Environments for Mobility", UPMC/Chart-LUTIN (Labex SMART), Franck Petit, Swan Dubois, and Francois Jouen (Chart), 01/12/2016.

PhD: Maxime Lorrillere, "A kernel cooperative cache for virtualized environments", UPMC, Sébastien Monnet, Julien Sopena, Pierre Sens, 02/04/2016.

PhD: Mahsa Najafzadeh, UPMC, funded by Inria competitive grant (Cordi-S), since Nov. 2012, Marc Shapiro, 22 April 2016.

PhD in Progress: Joao Paulo de Araujo, "L'exécution efficace d'algorithmes distribués dans les réseaux véhiculaires", funded by CNPq (Brésil), since Nov.2015, Pierre Sens and Luciana Arantes.

PhD in progress : Antoine Blin, "Execution of real-time applications on a small multicore embedded system", since April 2012, Gilles Muller (Whisper) and Julien Sopena, CIFRE Renault

PhD in progress: Sébastien Bouchard, "Gathering with faulty robots", UPMC, since Oct. 2016, Swan Dubois, Franck Petit, Yoann Dieudonné (University of Picardy Jules Verne)

PhD in progress: Marjorie Bournat, "Exploration with robots in dynamic networks", UPMC, since Sep. 2015, Swan Dubois, Franck Petit, Yoann Dieudonné (University of Picardy Jules Verne)

PhD in progress: Damien Carver, "HACHE : HorizontAl Cache cHorEgraphy - Toward automatic resizing of shared I/O caches.", UPMC, CIFRE, since Jan. 2015, Sébastien Monnet, Pierre Sens, Julien Sopena, Dimitri Refauvelet (Magency).

PhD in Progress: Florent Coriat, "Géolocalisation et routage en situation de crise" since Sept 2014, UPMC, Anne Fladenmuller (NPA-LIP6) and Luciana Arantes.

PhD in progress: Rudyar Cortes,"Un Environnement à grande échelle pour le traitement de flots massifs de données," UPMC, funded by Chile government, since Sep. 2013, Olivier Marin, Luciana Arantes, Pierre Sens.

PhD in progress: Lyes Hamidouche, "Data replication and data sharing in mobile networks", UPMC, CIFRE, since Nov. 2014, Sébastien Monnet, Pierre Sens, Dimitri Refauvelet (Magency).

PhD in progress: Denis Jeanneau,"Problèmes d'accord et détecteurs de défaillances dans les réseaux dynamique," UPMC, funded by Labex Smart, since Oct. 2015, Luciana Arantes, Pierre Sens.

PhD in progress: Yoann Péron, "Development of an adaptive recommendation system", UPMC/Makazi, since May 2014, Franck Petit, Patrick Gallinari, Matthias Oehler (Makazi).

PhD in progress: Alejandro Z. Tomsic, UPMC, funded by SyncFree, since Feb. 2014, Marc Shapiro.

PhD in progress: Tao Thanh Vinh, UPMC, CIFRE, since Feb. 2014, Marc Shapiro, Vianney Rancurel (Scality).

PhD in progress: Gauthier Voron, "Big-Os : un OS pour les grands volumes de données,", UPMC, since Sep. 2014, Gaël Thomas, Pierre Sens.

9.2.3. PhD committees

Franck Petit was the reviewer of:

- Mohamed Khaled, PhD UPJV (Advisors: F. Levé and V. Villain)

Franck Petit was Chair of:

- Quentin Bramas, UPMC (Advisor: S. Tixeuil)
- Fadwa Boubekeur, UPMC (Advisor: L. Blin)

Pierre Sens was the reviewer of:

- S. Chiky, HDR ISEP, Paris
- P. Li, PhD Bordeaux (Advisors: R. Namyst, E. Brunnet)
- J. Olivares, PhD Rennes 1 (Advisor: A-M Kermarrec)
- J. Pastor, PhD ENM, Nantes (Advisors: A. Lebre, F. Desprez)
- C. Sauvagnat, PhD IRIT, Toulouse (Advisor: M. Kaaniche)

Pierre Sens was Chair of

- C. Concolato, HDR Telecom Paris
- T. Ecarot, PhD, Telecom SupParis, UPMC (Advisor: D. Zeghlache)

Marc Shapiro was on the following committees:

- Kirill Bogdanov (Advisors: Dejan Kostic, Jeff Maguire), licentiate thesis, KTH, Stockholm, Nov. 2016.
- Brice Nédélec (Advisors: Pascal Molli, Achour Mostefaoui), PhD thesis, Nantes, oct. 2016.

9.3. Popularization

Julien Sopena animated an activity during the Science Festival 2016 at UPMC

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. H. KAAOUACHI. *A distributed approach to covering problems in highly dynamic systems*, UPMC - Paris 6 Sorbonne Universités, January 2016, https://tel.archives-ouvertes.fr/tel-01289153

[2] M. LORRILLERE. *A kernel cooperative cache for virtualized environments*, Université Pierre et Marie Curie - Paris VI, February 2016, https://hal.inria.fr/tel-01273367

[3] M. NAJAFZADEH. *The Analysis and Co-design of Weakly-Consistent Applications*, Université Pierre et Marie Curie, April 2016, https://hal.inria.fr/tel-01351187

Articles in International Peer-Reviewed Journals

[4] K. ALTISEN, A. COURNIER, S. DEVISMES, A. DURAND, F. PETIT. *Self-Stabilizing Leader Election in Polynomial Steps*, in "Information and Computation", 2016, http://hal.upmc.fr/hal-01347471

[5] V. BALEGAS, C. LI, M. NAJAFZADEH, D. PORTO, A. CLEMENT, S. DUARTE, C. FERREIRA, J. GEHRKE, J. LEITÃO, N. PREGUIÇA, R. RODRIGUES, M. SHAPIRO, V. VAFEIADIS. *Geo-Replication: Fast If Possible, Consistent If Necessary*, in "IEEE Data Engineering Bulletin", March 2016, vol. 39, n^{o} 1, 12 p. , https://hal.inria.fr/hal-01350652

[6] B. BONAKDARPOUR, S. DEVISMES, F. PETIT. *Snap-Stabilizing Committee Coordination*, in "Journal of Parallel and Distributed Computing", January 2016, vol. 87, pp. 26-42 [*DOI :* 10.1016/J.JPDC.2015.09.004], http://hal.upmc.fr/hal-01347461

[7] N. BRAUD-SANTONI, S. DUBOIS, M. H. KAAOUACHI, F. PETIT. *The Next 700 Impossibility Results in Time-Varying Graphs*, in "International Journal of Networking and Computing", 2016, vol. 6, n^{o} 1, pp. 27-41, https://hal.inria.fr/hal-01344422

[8] E. CARON, A. K. DATTA, C. TEDESCHI, F. PETIT. *Self-Stabilizing Prefix Tree Based Overlay Networks*, in "International Journal of Foundations of Computer Science", 2016, vol. 27, n^{o} 5, pp. 607–630 [*DOI :* 10.1142/S0129054116500192], http://hal.upmc.fr/hal-01347457

[9] A. COURNIER, A. K. DATTA, S. DEVISMES, F. PETIT, V. VILLAIN. *The expressive power of snap-stabilization*, in "Theoretical Computer Science", May 2016, vol. 626, pp. 40–66 [*DOI :* 10.1016/J.TCS.2016.01.036], http://hal.upmc.fr/hal-01292988

[10] N. HIDALGO, L. ARANTES, P. SENS, X. BONNAIRE. *ECHO: Efficient Complex Query over DHT overlays*, in "Journal of Parallel and Distributed Computing", February 2016, vol. 88, pp. 31-45 [*DOI :* 10.1016/J.JPDC.2015.10.007], https://hal.archives-ouvertes.fr/hal-01340589

[11] D. JEANNEAU, T. RIEUTORD, L. ARANTES, P. SENS. *Solving k-Set Agreement Using Failure Detectors in Unknown Dynamic Networks*, in "IEEE Transactions on Parallel and Distributed Systems", 2016, https://hal.inria.fr/hal-01354864

[12] S. F. PILEGGI. *Probabilistic Semantics*, in "Procedia Computer Science", 2016, vol. 80, pp. 1834–1845 [*DOI :* 10.1016/J.PROCS.2016.05.472], http://hal.upmc.fr/hal-01327353

[13] E. ROSAS, N. HIDALGO, V. GIL COSTA, M. MARIN, C. BONACIC, H. SENGER, L. ARANTES, C. MARCONDES, O. MARIN. *Survey on Simulation for Mobile Ad-Hoc Communication for Disaster Scenarios*, in "Journal of Computer Science and Technology", 2016, vol. 31, n° 2 [*DOI :* 10.1007/S11390-016-1630-X], http://hal.upmc.fr/hal-01354787

[14] H. SENGER, V. GIL-COSTA, L. ARANTES, C. A. MARCONDES, M. MARIN, L. M. SATO, F. A. B. DA SILVA. *BSP Cost and Scalability Analysis for MapReduce Operations*, in "Concurrency and Computation: Practice and Experience", June 2016, vol. 28, n° 8, pp. 2503-2527 [*DOI :* 10.1002/CPE.3628], https://hal.inria.fr/hal-01254275

[15] D. SERRANO, S. BOUCHENAK, Y. KOUKI, F. ALVARES DE OLIVEIRA JR., T. LEDOUX, J. LEJEUNE, J. SOPENA, L. ARANTES, P. SENS. *SLA guarantees for cloud services*, in "Future Generation Computer Systems", January 2016, vol. 54, pp. 233–246 [*DOI :* 10.1016/J.FUTURE.2015.03.018], https://hal.archives-ouvertes.fr/hal-01162654

[16] R. C. TURCHETTI, E. P. DUARTE, L. ARANTES, P. SENS. *A QoS-configurable failure detection service for internet applications*, in "Journal of Internet Services and Applications", 2016, vol. 7, n° 9, 14 p. [*DOI :* 10.1186/S13174-016-0051-Y], https://hal.inria.fr/hal-01405120

Invited Conferences

[17] M. SHAPIRO, M. SAEIDA ARDEKANI, G. PETRI. *Consistency in 3D*, in "Int. Conf. on Concurrency Theory (CONCUR) 2016", Québec, Canada, LIPICS (editor), Int. Conf. on Concurrency Theory (CONCUR) 2016, August 2016, vol. CONCUR 2016, 15 p. , https://hal.inria.fr/hal-01350668

International Conferences with Proceedings

[18] D. D. AKKOORATH, A. TOMSIC, M. BRAVO, Z. LI, T. CRAIN, A. BIENIUSA, N. PREGUIÇA, M. SHAPIRO. *Cure: Strong semantics meets high availability and low latency*, in "Int. Conf. on Distributed Computing Systems (ICDCS 2016)", Nara, Japan, IEEE, June 2016, pp. 405-414 [*DOI :* 10.1109/ICDCS.2016.98], https://hal.inria.fr/hal-01350558

[19] K. ALTISEN, S. DEVISMES, A. DURAND, F. PETIT. *Gradual Stabilization under τ-Dynamics*, in "22nd International European Conference on Parallel and Distributed Computing (Euro-Par 2016)", Grenoble, France, August 2016, http://hal.upmc.fr/hal-01348413

[20] L. ARANTES, R. FRIEDMAN, O. MARIN, P. SENS. *Probabilistic Byzantine Tolerance Scheduling in Hybrid Cloud Environments*, in "18th International Conference on Distributed Computing and Networking (ICDCN 2017)", Hyderabad, India, January 2017 [*DOI :* 10.1145/1235], https://hal.inria.fr/hal-01399026

[21] H. ATTIYA, S. BURCKHARDT, A. GOTSMAN, A. MORRISON, H. YANG, M. ZAWIRSKI. *Specification and Complexity of Collaborative Text Editing*, in "Int. Symp. on Principles of Distributed Computing (PODC) 2016", Chicago, IL, United States, ACM (editor), Int. Symp. on Principles of Distributed Computing (PODC) 2016, July 2016, vol. PODC 2016, 10 p. [*DOI :* 10.1145/2933057.2933090], https://hal.inria.fr/hal-01351512

[22] M. BOURNAT, A. K. DATTA, S. DUBOIS. *Self-Stabilizing Robots in Highly Dynamic Environments*, in "SSS 2016 - 18th International Symposium Stabilization, Safety, and Security of Distributed Systems", Lyon, France, Lecture Notes in Computer Science, Springer, November 2016, vol. 10083, pp. 54-69 [*DOI :* 10.1007/978-3-319-49259-9_5], https://hal.archives-ouvertes.fr/hal-01416308

[23] L. BLIN, F. BOUBEKEUR, S. DUBOIS. *Algorithme auto-stabilisant efficace en mémoire pour la construction d'un arbre couvrant de diamètre minimum*, in "ALGOTEL 2016 - 18èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications", Bayonne, France, May 2016, https://hal.archives-ouvertes.fr/hal-01302779

[24] A. BLIN, C. COURTAUD, J. SOPENA, J. LAWALL, G. MULLER. *Maximizing Parallelism without Exploding Deadlines in a Mixed Criticality Embedded System*, in "28th EUROMICRO Conference on Real-Time Systems (ECRTS'16)", Toulouse, France, July 2016, https://hal.inria.fr/hal-01346979

[25] A. BLIN, C. COURTAUD, J. SOPENA, J. LAWALL, G. MULLER. *Understanding the Memory Consumption of the MiBench Embedded Benchmark*, in "Netys", Marakech, Morocco, May 2016, https://hal.inria.fr/hal-01349421

[26] G. BOSILCA, A. BOUTEILLER, A. GUERMOUCHE, T. HÉRAULT, Y. ROBERT, P. SENS, J. DONGARRA. *Failure Detection and Propagation in HPC systems*, in "SC 2016 - The International Conference for High Performance Computing, Networking, Storage and Analysis", Salt Lake City, United States, November 2016, https://hal.inria.fr/hal-01352109

[27] L. BRIOT, P. URSO, M. SHAPIRO. *High Responsiveness for Group Editing CRDTs*, in "ACM International Conference on Supporting Group Work", Sanibel Island, FL, United States, November 2016 [*DOI :* 10.1145/2957276.2957300], https://hal.inria.fr/hal-01343941

[28] F. CORIAT, A. FLADENMULLER, L. ARANTES, O. MARIN. *Crowdsourcing-based architecture for post-disaster geolocation: A comparative performance evaluation*, in "The 15th IEEE International Symposium on Network Computing and Applications (NCA 2016)", Cambridge, MA, United States, October 2016, pp. 1 - 9 [*DOI :* 10.1109/NCA.2016.7778583], http://hal.upmc.fr/hal-01416297

[29] R. CORTÉS, X. BONNAIRE, O. MARIN, L. ARANTES, P. SENS. *GeoTrie: A Scalable Architecture for Location-Temporal Range Queries over Massive GeoTagged Data Sets*, in "The 15th IEEE International Symposium on Network Computing and Applications (NCA 2016)", Cambridge, MA, United States, 2016, https://hal.inria.fr/hal-01388949

[30] A. GOTSMAN, H. YANG, C. FERREIRA, M. NAJAFZADEH, M. SHAPIRO. *'Cause I'm Strong Enough: Reasoning about Consistency Choices in Distributed Systems*, in "Symposium on Principles of Programming Languages", Saint Petersburg, FL, United States, January 2016, pp. 371–384 [*DOI :* 10.1145/2837614.2837625], https://hal.inria.fr/hal-01243192

[31] A. GRACIELA DE MORAES ROSSETTO, C. R. GEYER, L. ARANTES, P. SENS. *Implementing a flexible failure detector that expresses the confidence in the system*, in "LADC 2016 - 7th Latin-American Symposium on Dependable Computing", Cali, Colombia, October 2016, https://hal.inria.fr/hal-01352162

[32] S. ISKOUNEN, T.-M.-T. NGUYEN, S. MONNET. *WiFi-Direct Simulation for INET in OMNeT++*, in "The 3rd OMNeT++ Community Summit", Brno, Czech Republic, September 2016, http://hal.upmc.fr/hal-01389581

[33] D. JEANNEAU, L. A. RODRIGUES, E. DUARTE, L. ARANTES. *An Autonomic Hierarchical Reliable Broadcast Protocol for Asynchronous Distributed Systems with Failure Detector*, in "Latin-American Symposium on Dependable Computing (LADC)", Cali, Colombia, 2016, https://hal.inria.fr/hal-01355322

[34] M. NAJAFZADEH, A. GOTSMAN, H. YANG, C. FERREIRA, M. SHAPIRO. *The CISE Tool: Proving Weakly-Consistent Applications Correct*, in "PaPoC 2016 - 2nd Workshop on the Principles and Practice of Consistency for Distributed Data", Londres, United Kingdom, ACM, April 2016 [*DOI :* 10.1145/2911151.2911160], https://hal.inria.fr/hal-01350636

[35] L. A. RODRIGUES, L. ARANTES, E. DUARTE. *An Autonomic Majority Quorum System*, in "30th IEEE International Conference on Advanced Information Networking and Applications (AINA)", Crans-Montana, Switzerland, 2016 [*DOI :* 10.1016/J.JPDC.2015.10.007], http://hal.upmc.fr/hal-01349513

[36] A. TOMSIC, T. CRAIN, M. SHAPIRO. *PhysiCS-NMSI: efficient consistent snapshots for scalable snapshot isolation*, in "Int. W. on Principles and Practice of Consistency for Distr. Data (PaPoC) 2016", London, United Kingdom, ACM (editor), Int. W. on Principles and Practice of Consistency for Distr. Data (PaPoC) 2016, April 2016, vol. PaPoC 2016, 4 p. [*DOI :* 10.1145/2911151.2911166], https://hal.inria.fr/hal-01350657

[37] M. ZAWIRSKI, C. BAQUERO, A. BIENIUSA, N. PREGUIÇA, M. SHAPIRO. *Eventually Consistent Register Revisited*, in "Int. W. on Principles and Practice of Consistency for Distributed Data (PaPoC)", London, United Kingdom, ACM (editor), Int. W. on Principles and Practice of Consistency for Distributed Data (PaPoC), ACM Sigops / EuroSys, April 2016, vol. PaPoC 2016, 7 p. [*DOI :* 10.1145/2911151.2911157], https://hal.inria.fr/hal-01242700

Scientific Books (or Scientific Book chapters)

[38] A. GRACIELA DE MORAES ROSSETTO, C. ROLIM, V. LEITHARDT, C. R. GEYER, L. ARANTES, P. SENS. *A failure detector based on processes' relevance and the confidence degree in the system for self-healing in ubiquitous environments*, in "Pervasive Computing", C. DOBRE, F. XHAFA (editors), Next Generation Platforms for Intelligent Data Collection, 2016, 23 p. , https://hal.inria.fr/hal-01355320

Research Reports

[39] D. D. AKKOORATH, A. TOMSIC, M. BRAVO, Z. LI, T. CRAIN, A. BIENIUSA, N. PREGUIÇA, M. SHAPIRO. *Cure: Strong semantics meets high availability and low latency*, Inria ; Paris 6, February 2016, n^o RR-8858, https://hal.inria.fr/hal-01270776

[40] A. BLIN, C. COURTAUD, J. SOPENA, J. LAWALL, G. MULLER. *Maximizing Parallelism without Exploding Deadlines in a Mixed Criticality Embedded System*, Inria, February 2016, nº RR-8838, https://hal.inria.fr/hal-01268078

[41] M. BOURNAT, A. K. DATTA, S. DUBOIS. *Self-Stabilizing Robots in Highly Dynamic Environments* , LIP6 UMR 7606, Inria, UPMC Sorbonne Universités, France ; University of Nevada, Las Vegas, United States, September 2016, https://hal.inria.fr/hal-01368920

[42] M. BOURNAT, S. DUBOIS, F. PETIT. *Computability of Perpetual Exploration in Highly Dynamic Rings*, Laboratoire d'informatique de Paris 6 [LIP6] ; UPMC Sorbonne Universités/CNRS/Inria - EPI REGAL, December 2016, https://hal.archives-ouvertes.fr/hal-01417565

[43] M. NAJAFZADEH, A. GOTSMAN, H. YANG, C. FERREIRA, M. SHAPIRO. *The CISE Tool: Proving Weakly-Consistent Applications Correct*, Inria Paris Rocquencourt, February 2016, n^o RR-8870, https://hal.archives-ouvertes.fr/hal-01279495

[44] A. G. M. ROSSETTO, L. ARANTES, P. SENS, C. R. GEYER. *Impact: an Unreliable Failure Detector Based on Processes' Relevance and the Confidence Degree in the System*, Université Pierre et Marie Curie ; Inria Paris-Rocquencourt - Regal ; Universidade Federal do Rio Grande do Sul, September 2016, https://hal.inria.fr/hal-01136595

[45] M. SHAPIRO, M. SAEIDA ARDEKANI, G. PETRI. *Consistency in 3D*, Institut National de la Recherche en Informatique et Automatique (Inria), July 2016, n^o RR-8932, https://hal.archives-ouvertes.fr/hal-01343592

Other Publications

[46] P. DARCHE. *Évolution des mémoires à semi-conducteurs à accès aléatoire*, February 2017, article E2491 des techniques de l'ingenieur p. , article de référence sur le domaine, http://hal.upmc.fr/hal-01341972

References in notes

[47] V. BALEGAS, S. DUARTE, C. FERREIRA, R. RODRIGUES, N. PREGUIÇA, M. NAJAFZADEH, M. SHAPIRO. *Putting Consistency back into Eventual Consistency*, in "Euro. Conf. on Comp. Sys. (EuroSys)", Bordeaux, France, ACM, 2015, pp. 6:1–6:16 [*DOI :* 10.1145/2741948.2741972], https://hal.inria.fr/hal-01248191

[48] M. NAJAFZADEH, M. SHAPIRO. *Demo of the CISE tool*, November 2015, https://youtu.be/HJjWqNDh-GA