RESEARCH CENTRE

Paris

IN PARTNERSHIP WITH:

Sorbonne Université

2021
ACTIVITY REPORT

Project-Team
DELYS

DistributEd aLgorithms and sYStems

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

DOMAIN

Networks, Systems and Services,
Distributed Computing

THEME

Distributed Systems and middleware

Contents

Project-Team DELYS

Creation of the Project-Team: 2019 January 01

Keywords

Computer sciences and digital sciences

A1.1.1. – Multicore, Manycore

A1.1.9. – Fault tolerant systems

A1.1.13. – Virtualization

A1.2.5. – Internet of things

A1.3.2. – Mobile distributed systems

A1.3.3. – Blockchain

A1.3.4. – Peer to peer

A1.3.5. – Cloud

A1.3.6. – Fog, Edge

A1.5.2. – Communicating systems

A2.6. – Infrastructure software

A2.6.1. – Operating systems

A2.6.2. – Middleware

A2.6.3. – Virtual machines

A2.6.4. – Ressource management

A3.1.3. – Distributed data

A3.1.8. – Big data (production, storage, transfer)

A7.1.1. – Distributed algorithms

Other research topics and application domains

B6.4. – Internet of things

1 Team members, visitors, external collaborators

Research Scientists

- Mesaac Makpangou [Inria, Researcher, HDR]
- Marc Shapiro [Inria, Emeritus, HDR]

Faculty Members

- Pierre Sens [Team leader, Sorbonne Université, Professor, HDR]
- Luciana Bezerra Arantes [Sorbonne Université, Associate Professor]
- Philippe Darche [Université de Paris, Associate Professor]
- Swan Dubois [Sorbonne Université, Associate Professor]
- Bertil Folliot [Sorbonne Université]
- Colette Johnen [Université de Bordeaux, Professor, until Aug 2021, HDR]
- Jonathan Lejeune [Sorbonne Université, Associate Professor]
- Franck Petit [Sorbonne Université, Professor, HDR]
- Julien Sopena [Sorbonne Université, Associate Professor]

PhD Students

- Aymeric Agon-Rambosson [Sorbonne Université, From Oct 2021]
- Jose Jurandir Alves Esteves [Orange Labs, Until Dec. 2021]
- Arnaud Favier [Inria]
- Saalik Hatia [Sorbonne Université]
- Gabriel Le Bouder [École Nationale Supérieure de Chimie et de Physique de Bordeaux]
- Celia Mahamdi [Sorbonne Université]
- Benoit Martin [Sorbonne Université]
- Sreeja Nair [Sorbonne Université, until Aug 2021]
- Laurent Prosperi [Inria (Cordi-S)]
- Jonathan Sid-Otmane [Orange Labs, CIFRE]
- Ilyas Toumlilt [Sorbonne Université]
- Dimitrios Vasilas [Sorbonne Université, until Apr 2021]
- Daniel Wilhelm [Sorbonne Université]

Technical Staff

- Yannick Li [Inria, Engineer, until Oct 2021]

Administrative Assistants

- Christine Anocq [Inria]
- Nelly Maloisel [Inria]

External Collaborator

- Sebastien Monnet [Université Savoie Mont-Blanc]

2 Overall objectives

The research of the Delys team addresses the theory and practice of computer systems, with a focus on distributed systems. This includes multicore computers, clusters, networks, peer-to-peer systems, cloud, fog end edge computing systems, and other communicating entities such as swarms of robots. Delys addresses the challenges of correctly communicating, sharing information, and computing in large-scale, highly dynamic computer systems. Core problems of interest include communication, consensus and fault detection, scalability, resource management, data management, data sharing (replication and consistency), group collaboration, dynamic content distribution, and multi- and many-core concurrent algorithms.

Delys is a joint research team between LIP6 (Sorbonne-Université/CNRS) and Inria Paris.

3 Research program

3.1 Research rationale

Delys addresses the theoretical and practical issues of *Computer Systems*, leveraging our dual expertise in theoretical and experimental research. Our major focus is the sharing of information and guaranteeing correct execution of highly-dynamic computer systems. Our research covers a large spectrum of distributed computer systems including multicore computers, mobile networks, cloud computing systems, or dynamic communicating entities. This holistic approach enables handling related problems at different levels. Among such problems we can highlight consensus, fault detection, scalability, search of information, resource allocation, replication and consistency of shared data, dynamic content distribution, and concurrent and parallel algorithms.

Our approach aims to establish a principled "virtuous cycle." In response to some a concrete issue in a real system, we might design an algorithm. We prove it correct and we evaluate it theoretically, and also implement and test it experimentally The conclusions feed back to algorithm design and to theory.

Two current evolutions in the Computer Systems area strongly influence our research project:

1. **A modern computer system is increasingly distributed and dynamic**. It is composed of multiple devices, geographically spread over heterogeneous platforms, spanning multiple management domains. Years of fundamental research in the field are now coming to fruition, and are being used by millions of users of web systems, peer-to-peer systems, gaming and social applications, cloud computing, and now edge computing. These new uses bring new challenges, such as the adaptation to dynamically-changing conditions, where knowledge of the system state can only be partial and incomplete.

2. **Heterogeneous architectures and virtualisation are everywhere.** The parallelism offered by distributed clusters and *multicore* architectures is opening highly parallel computing to new application areas. To be successful, however, many issues need to be addressed. Challenges include obtaining a consistent view of shared resources, such as memory, and optimally distributing computations among heterogeneous architectures. These issues arise at a more fine-grained level than before, leading to the need for different solutions down to OS level itself.

The scientific challenges of the distributed computing systems are subject to many important features which include scalability, fault tolerance, dynamics, emergent behaviour, heterogeneity, and virtualisation at many levels. 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. Sometimes, classical "*static*" problems, (*e.g.*, Leader Election, Spanning Tree Construction, ...) even need to be redefined to consider the unstable nature of the distributed system.

In particular, Delys focuses on a number of key challenges:

Consistency in geo-scale and edge systems. Such systems need to scale to large geographies and large numbers of attached devices, while executing in an untamed, unstable environment. This poses difficult scientific challenges, which are all the more pressing as the cloud moves more and more towards the edge, IoT and mobile computing. A key issue is how to share data effectively and consistently across the whole spectrum. Delys has made several key contributions, including CRDTs, the Transactional Causal Consistency Plus model, the AntidoteDB geo-distributed database, its edge extension Colony, and the startup concordant.io.

Rethinking distributed algorithms. From a theoretical point of view the key question is how to adapt the fundamental building blocks to new architectures. More specifically, how to rethink classical distributed algorithms, to take into account the dynamics of advanced modern systems. The recent literature on dynamic systems proposes many different ad-hoc models, one for each setting, without unification. Furthermore, the models are often unrealistic. A key challenge is to identify which assumptions make sense in new distributed systems. The research objectives of Delys are then (1) to identify under which realistic assumptions a given fundamental problem such as mutual exclusion, consensus or leader election can be solved, and (2) to design efficient algorithms under these assumptions.

Resource management in heterogeneous systems. A key practical issue is how to manage resources on large and heterogeneous configurations. Managing resources in such systems requires fully decentralized solutions, and to rethink the way various platforms can collaborate and interoperate with each other. In this context, data management is a key component. The fundamental issue we address is how to efficiently and reliably share information in highly distributed environments.

Adaptation of runtimes. The OS community faces the challenge of adapting runtime support to new architectures. With the increasingly widespread use of multicore architectures and virtualised environments, internal runtime protocols need to be revisited. Especially, memory management is crucial in OS and virtualisation technologies have highly impact on it. On one hand, the isolation property of virtualisation has severe side effects on the efficiency of memory allocation since it needs to be constantly balanced between hosted OSs. On the other hand, by hiding the physical machine to OSs, virtualisation prevents them to efficiently place their data in memory on different cores. Our research will thus focus on providing solutions to efficiently share memory between OSs without jeopardizing isolation properties.

4 Application domains

We target highly distributed infrastructures composed of multiple devices geographically spread over heterogeneous platforms including cloud, fog computing and IoT.

At OS level, we study multicore architectures and virtualized environments based on VM hypervisors and containers. Our research focuses on providing solutions to efficiently share memory between virtualized environments.

5 Highlights of the year

5.1 Awards

1. Best student paper award at 20th IEEE International Symposium on Network Computing and Applications, Nov 2021, Cambridge, Boston, United States (NCA 2021) [15]

2. Best student paper award at Conference on Principles of Distributed Systems (OPODIS 2021), Dec 2021, Strasbourg, France [14]

3. Best demo paper award at CNSM 2021 – 17th International Conference on Network and Service Management (CNSM 2021) [12]

4. The Painless (PArallel INstantiabLE Sat Solver) software was awarded First Prize in the Parallel Track of the 2021 SAT Competition. It is developed jointly by LRDE and by the MoVe and DELYS teams of LIP6.

6 New software and platforms

6.1 New software

6.1.1 Colony

Name: Extending TCC+ and stronger guarantees to the network far edge

Keywords: Edge Computing, Causal consistency, Replication and consistency, P2P, Hybrid consistency

Functional Description: Colony is developed as an extension to AntidoteDB. It guarantees Transactional Causal Plus Consistency (TCC+) globally, based on a forest topology, rooted at the AntidoteDB data centres. It features P2P edge groups. An edge group enjoys strong consistency internally, functions with or without a cloud connection, and can disconnect and migrate to a different subtree in the topology. Colony features a dynamic access control system, leveraging the TCC+ guarantees between data and their associated access control lists.

Contact: Ilyas Toumlilt

Partner: Technische Universität Kaiserslautern (UniKL), Allemagne

7 New results

7.1 Distributed Algorithms for Dynamic Networks and Fault Tolerance

Participants: Luciana Bezerra Arantes, Swan Dubois, Arnaud Favier, Colette Johnen, Jonathan Lejeune, Célia Mahamdi, Mesaac Makpangou, Franck Petit, 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.

In 2021, we obtained the following results.

Ordered Message broadcast. FIFO broadcast provides application ordering semantics of messages broadcast by the same sender and have been mostly implemented on top of unreliable static networks. In [18], we proposed a round-based FIFO broadcast algorithm with both termination detection and bounded message size for dynamic networks with recurrent connectivity. Initially, processes only know the number of processes N in the system and their identifier. Due to the dynamics of the network links, messages can be lost. Since no unbounded timestamp is used to identify a message, its size is bounded to $2N + O(log(N)) + msgSize$ bits where $msgSize$ is the bound size in bits of the broadcast data. We also proposed a FIFO atomic broadcast algorithm for dynamic networks with recurrent connectivity that uses the proposed FIFO broadcast and deliver primitives. This algorithm provides causal total order broadcast primitives.

Topology aware Leader election. Eventual leader election is an essential service for many reliable applications that require coordination actions on top of asynchronous fail-prone distributed systems.In [15], we propose CEL, a new distributed eventual leader election algorithm for dynamic networks, which exploits topological information to improve the choice of a central leader and reduce message exchanges. The algorithm has a crosslayer neighbors detection, with a neighbor-aware mechanism, to improve the sharing of topological knowledge and elect a central leader faster. It uses a self-pruning mechanism based on topological knowledge, combined with probabilistic gossip, to improve the performance of broadcast propagation. Evaluations were conducted on the OMNeT++ environment, simulating realistic MANET with interference, collision, and messages loss. The results show our approach reduces the number of messages sent and provides a stable leader and short paths to the leader.

Self-Stabilizing Leader election in Dynamic Networks. Essentially, self-stabilizing algorithms tolerate *transient failures*, since by definition such failures last a finite time (as opposed to crash failures, for example) and their frequency is low (as opposed to intermittent failures). Self-stabilization is also an approach to tolerate topological changes of the interconnection network. In that case, topological changes are considered as transient failures of links. So, as for other transient faults, the safety cannot be guaranteed all the time. Such approach becomes totally ineffective when the frequency of topological is very high. In this case, the network dynamics should be no more considered as an anomaly but rather as an integral part of the system nature.

We studied conditions under which stabilizing leader election can be solved in highly dynamic identified message passing systems. In [11], we provide necessary and sufficient conditions under which the problem can be solved assuming that every process can reach all the others at least once through a journey—a journey can be thought as a path over time from a source to a destination. In [10], we weaken the model: we consider dynamic systems where some processes may not be either sources or recipients. We show that self-stabilizing leader election can only be achieved if all processes are both sources and recipients—*i.e.*, the models studied in [11]. Otherwise (processes are either only sources or recipients), we show that even pseudo-stabilization—a weaker form of stabilization—cannot be achieved, except in a special case where each source can always reach all other processes within some bounded time.

Optimal Space Lower Bound for Deterministic Self-Stabilizing Leader Election Algorithms. Given a boolean predicate Π on labeled networks (*e.g.*, proper coloring, leader election, etc.), a self-stabilizing algorithm for Π is a distributed algorithm that can start from any initial configuration of the network (*i.e.*, every node has an arbitrary value assigned to each of its variables), and eventually converge to a configuration satisfying Π. It is known that leader election does not have a deterministic self-stabilizing algorithm using a constant-size register at each node, *i.e.*, for some networks, some of their nodes must have registers whose sizes grow with the size n of the networks. On the other hand, it is also known that leader election can be solved by a deterministic self-stabilizing algorithm using registers of $O(\log\log n)$ bits per node in any n-node bounded-degree network. In [14], we show that this latter space complexity is optimal. Specifically, we prove that every deterministic self-stabilizing algorithm solving leader election must use $\Omega(\log\log n)$-bit per node registers in some n-node networks. In addition, we show that our lower bounds go beyond leader election, and apply to all problems that cannot be solved by anonymous algorithms.

This last paper [14] received the Best Student Paper Award at OPODIS 2021.

7.2 Distributed systems and Large-scale data distribution

Participants: Luciana Arantes, Jose Jurandir Alves Esteves, Saalik Hatia, Jonathan Sid-Otmane, Pierre Sens, Marc Shapiro, Julien Sopena, Ilyas Toumlilt, Dimitrios Vasilas, Daniel Wladdimiro.

Resource management in large networks. Network Operators expect to accurately satisfy a wide range of user's needs by providing fully customized services relying on Network Slicing. The efficiency of

Network Slicing depends on an optimized management of network resources and Quality of Service (QoS). We focus on Network Slice placement optimization problem.

In [13] and [12], we consider online learning for optimal network slice placement under the assumption that slice requests arrive according to a non-stationary Poisson process. We propose a framework based on Deep Reinforcement Learning (DRL) combined with a heuristic to design algorithms. We specifically design two pure-DRL algorithms and two families of hybrid DRL-heuristic algorithms. To validate their performance, we perform extensive simulations in the context of a large-scale operator infrastructure. The evaluation results show that the proposed hybrid DRL-heuristic algorithms require three orders of magnitude of learning episodes less than pure-DRL to achieve convergence. This result indicates that the proposed hybrid DRLheuristic approach is more reliable than pure-DRL in a real non-stationary network scenario.

Adaptive stream processing systems. In [22], we propose a new adaptive Stream Processing System (SPS) for real-time processing that, based on input data rate variation, dynamically adapts the number of active operator replicas. Our SPS extends Storm by pre-allocating, for each operator, a set of inactive replicas which are activated (or deactivated) when necessary without the Storm reconfiguration cost. We exploit the MAPE model and define a new metric that aggregates the value of multiple metrics to dynamically changes the number of replicas of an operator. We deploy our SPS over Google Cloud Platform and results confirm that our metric can tolerate highly dynamic conditions, improving resource usage while preserving high throughput and low latency

Edge-first collaborative data systems. Distributing and replicating data at the edge enables immediate response, autonomy and availability in edge applications, such as gaming, cooperative engineering, or in-the-field information sharing. However, application developers and users demand the highest possible consistency guarantees, and specific support for group collaboration. To address this challenge, we designed the Colony system. It guarantees Transactional Causal Plus Consistency (TCC+) globally, dovetailing with Snapshot Isolation within edge groups. To help with scalability, fault tolerance and security, its logical communication topology is tree-like, with replicated roots in the core cloud, but with the flexibility to migrate a node or a group. Despite this hybrid approach, applications enjoy the same semantics everywhere in the topology. Our experiments show that peer groups improve throughput and response time significantly, performance is not affected in offline mode, and that migration is seamless.

This work, joint with Pierre Sutra of Télécom SudParis, is published at Middleware 2021 [19] and is the thesis topic of Ilyas Toumlilt [34].

The concepts explored in this work are currently being industrialised, thanks to support from Inria Startup Studio, in the start-up Concordant.

Distributed and federated indexing This work studies how to support efficient query processing, where users and data are distributed across multiple geographic locations. In particular, we study how the placement of indexes or materialized views, and the query/storage communication patterns, affect the metrics of performance, freshness, and resource consumption. We propose a query engine architecture that enables the administrator to to make the appropriate placement decisions on a case-by-case basis.

The enabling technique is a composition-based design. Our architecture consists of building block components, with a uniform interface and interaction semantics. Each block encapsulates some primitive query processing task. The administrator composes a system by instantiating pertinent blocks, connected in a directed acyclic graph. She chooses and places the blocks to provide some higher-order query processing capabilities, while satisfying the administrator's performance objectives.

We propose an implementation of the proposed approach, in the form of a framework for constructing and deployment application-specific query engines, called Proteus. The experimental evaluation supports the theoretical analysis of the trade-offs involved in query processing state placement, and suggests that Proteus can effectively occupy multiple different points in the design space of geo-distributed query processing.

This joint work with B. King, through a Cifre grant with Scality, is the thesis topic of D. Vasilas [35].

Exploring the coordination design space. Static analysis tools, for instance Soteria, make it possible to identify which updates conflict and need to be coordinated in order to maintain the data invariants. The coordination can be implemented in many ways, trading overhead against parallelism. The design space is multi-dimensional: locks can have various levels of granularity; different types of lock can be used (for example, mutex vs. shared/exclusive locks); the placement of the lock object has a significant impact. Furthermore, the performance of an option depends on the workload.

To systematize the dimensions of coordination, we construct a coordination lattice, which enables to systematically navigate the concurrency control dimensions of granularity, mode, and placement. The choice of granularity affects the cost of both lock acquisition and of lock contention (the coarser the lock, the lesser the cost of acquisition, but the higher the contention). Placement affects only the lock acquisition costs, and mode affects lock contention. Therefore the major dimension of the coordination lattice is granularity with mode and placement as secondary dimensions. Navigating any dimension has an impact on the overhead of locking. Accordingly, we propose a systematic approach to the design of correct coordination protocols, enabling the designer to select one according to performance metrics.

This work is published in the PhD thesis of Sreeja Nair [32].

A highly-available tree data type. The tree is an essential data structure in many applications. In a distributed application, such as a distributed file system, the tree is replicated. To improve performance and availability, different clients should be able to update their replicas concurrently and without coordination. Such concurrent updates converge if the effects commute, but nonetheless, concurrent moves can lead to incorrect states and even data loss. Such a severe issue cannot be ignored; ultimately, only one of the conflicting moves may be allowed to take effect. However, as such conflicts are infrequent, a solution should be lightweight. Previous approaches would require preventative cross-replica coordination, or totally order move operations after-the-fact, requiring roll-back and compensation operations.

In this work, we present a novel replicated tree that supports coordination-free concurrent atomic moves, and provably maintains the tree invariant. Our analysis identifies cases where concurrent moves are inherently safe, and we devise a lightweight, coordination-free, rollback-free algorithm for the remaining cases, such that a maximal safe subset of moves takes effect.

We present a detailed analysis of the concurrency issues with trees, justifying our replicated tree data structure. We provide mechanized proof that the data structure is convergent and maintains the tree invariant. Finally, we compare the response time and availability of our design against the literature.

This is joint work with Carla Ferreira of Universidade NOVA de Lisboa and her group. It is described in an Inria technical report [37] and in the PhD thesis of Sreeja Nair [32].

Highly-available file system. Building scalable and highly available geo-replicated file systems is hard. These systems need to resolve conflicts that emerge in concurrent operations in a way that maintains file system invariants, is meaningful to the user, and does not depart from the traditional file system interface. Conflict resolution in existing systems often leads to unexpected or inconsistent results. We design and implement ElmerFS, a geo-replicated, truly concurrent file system designed with the aim of addressing these challenges. ElmerFS is based on two key ideas: (1) the use of Conflict-Free Replicated Data Types (CRDTs) for representing file system structures, which ensures that replicas converge to a correct state, and (2) conflict resolution rules, which are determined by the choice of CRDT types and their composition, designed with the principle of being intuitive to the user.

This joint work with Romain Vaillant, Brad King and Dimitri Vasilas of Scality, was presented at HotStorage 2021 [20].

A formally verified geo-distributed database, from model to implementation. We are leading an informal international collaboration to design and implement a full-featured geo-distributed database that is correct by construction. We adopt a stepwise approach. The initial step studies a simplistic concurrent database system; we describe its key invariants, formalise an operational semantics, and provide a reference implementation. From there, we plan to show that the semantic model satisfies the invariants (proof in Coq), that the implementation respects the model, and that the implementation passes litmus-test cases. Each following step adds a single feature, such as a cache, journaling, sharding, or garbage collection. We formalise the feature, and expect to show formally and experimentally that the improved system simulates the simpler one. We expect that the features compose, so that the correctness of the full system with all the features follows from the individual proofs. At the time of writing, the operational semantics for several steps is complete; the translation to Coq and the reference implementation are ongoing. This is joint work with Gustavo Petri (ARM Research Cambridge UK), Annette Bieniusa (TU Kaiserslautern) and Carla Ferreira (Universidade NOVA de Lisboa). A preliminary technical report is available [38]. This is the PhD topic of Saalik Hatia.

Developing, deploying and running correct distributed software by composition. Modern applications are highly distributed and data-intensive. Programming a distributed system is challenging because of asynchrony, failures and trade-offs. In addition, application requirements vary with the use-case and throughout the development cycle. Moreover, existing tools come with restricted expressiveness or limited runtime customisability. This work aims to address this by improving reuse while maintaining fine-grain control and enhancing dependability. We argue that an environment for composable distributed computing will facilitate the process of developing distributed systems. We use high-level composable specification, verification tools and a distributed runtime.

This work will be the thesis topic of Benoît Martin and Laurent Prosperi (who is co-advised by Ahmed Bouajjani of Université de Paris).

Decentralised limiters for 5G Slicing. Meeting the goals of 5G networks —high bandwidth, low latency, massive connectivity, and resiliency— demands improvements to the infrastructure that hosts the network components. Mobile Network Operators will rely on a geographically distributed and highly scalable infrastructure that must handle and replicate user data consistently. This paper explores the management of user data with regards to data consistency in the first 5G specification. In particular we will focus on how the 5G system procedures handle and update data, and discuss failure scenarios where the correctness properties of the user data may be violated. In this work we present the necessary properties that an underlying data store must deliver in order to maintain correctness in the presence of failures.

The 5G specification describes a true geo-distributed system; we study the specification from the perspective of the consistency issues it raises. We currently focus on the specific use case of limiting resource usage in 5G slices (a slice is a virtual network involving geo-distributed users). We compare different consistency approaches and limitation algorithms, to study the trade-off between consistency, cost, and faithfulness to the prescribed limits.

This is joint work with Sofiane Imdali and Frédéric Martelli, through a Cifre grant with Orange Labs, and is the PhD topic of Jonathan Sid-Otmane [33].

7.3 Resource management in system software

Participants: Jonathan Lejeune, Marc Shapiro, Julien Sopena, Yoann Ghigoff.

In-kernel Caching In 2021, we studied in-memory key-value stores aiming at improving their performances. In-memory key-value stores are critical components that help scale large internet services by providing low-latency access to popular data. Memcached, one of the most popular key-value stores, suffers from performance limitations inherent to the Linux networking stack and fails to achieve high performance when using high-speed network interfaces. While the Linux network stack can be bypassed using DPDK based solutions, such approaches require a complete redesign of the software stack and induce high CPU utilization even when client load is low. To overcome these limitations, we propose in [16], published at NSDI'21, BMC an in-kernel cache for Memcached that serves requests before the execution of the standard network stack. Requests to the BMC cache are processed as part of the NIC interrupts, which allows performance to scale with the number of cores serving the NIC queues. To ensure safety, BMC is implemented using eBPF. On small requests, our evaluations show that BMC improves throughput by up to 18x compared to the vanilla Memcached application and up to 6x compared to an optimized version of Memcached. In addition, our results also show that BMC has negligible overhead and does not decrease the throughput for larger requests.

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Participants: José Jurandir Alves Esteves, Pierre Sens, Marc Shapiro, Jonathan Sid-Otmane, Dimitrios Vasilas.

DELYS has a CIFRE contract with Scality SA:

- Dimitrios Vasilas is advised by Marc Shapiro and Brad King. He works on secondary indexing in large-scale storage systems under weak consistency. He obtained his PhD in July 2021.

DELYS has two contracts with Orange within the I/O Lab joint laboratory:

- Jonathan Sid-Otmane is advised by Marc Shapiro. He studies the applications of distributed databases to the needs of the telco industry in the context of 5G. He obtained his PhD in December 2021.
- José Jurandir Alves Esteves is advised by Pierre Sens. He works on network slice placement stategies. He obtained his PhD in December 2021.

8.2 Startup support from Inria

Marc Shapiro received support from Inria Startup Studio to incubate start-up concordant.io, developing CRDT-based solutions for geo-scale and edge distribution of data. ISS supports two software engineers for 12 months.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 STIC/MATH/CLIMAT AmSud project

ReMatch

Title: ReMatch : Resource Management in Clouds for Executing High Performance Applications

Local supervisor: Pierre Sens

Partners:

- Capes-Print
- Universidade Federal Fluminense (UFF)
- Université d'Avignon
- Mine Paristech
- Université de Bordeaux.
- Université de Montpellier

Inria contact: *Pierre Sens*

Summary: This project aims to solve the problem of resource allocation and management in cloud computing for HPC applications, minimizing execution time, power consumption and maximizing fault tolerance without violating SLA, and using applications from the field of biology as a case study. Cloud computing has traditionally been used for data sharing and general purpose services, but more recently it has begun to emerge as a promising alternative for High Performance Computing (HPC) applications. This computational paradigm offers several advantages when compared to a dedicated infrastructure, such as rapid provisioning of resources and significant reduction of operational costs. However, some challenges must be overcome to bridge the gap between the performance offered by a dedicated infrastructure and the clouds. Overheads introduced by the virtualization layer, hardware heterogeneity and high network latencies negatively affect the performance of HPC applications. In addition, cloud providers generally adopt resource-sharing policies that can further reduce the performance of such applications. Typically, a physical server can host multiple virtual machines that can cause contention in accessing shared resources, such as cache and main memory, significantly reducing their performance. In addition, the selection of virtual machines and their manual configuration is a rather complex task for scientists who develop HPC applications and are not experts in cloud administration tools. This problem becomes even more complex if we consider scenarios where scientists must perform a number of HPC applications with data dependence (i.e. workflow). Application schedulers that have multiple policies that vary according to the objective function such as minimizing the total execution time, minimizing the demand for energy, maintaining a guarantee of service level with the user, among others, play a fundamental role in ensuring the efficiency of the execution of such applications. In order to leverage the use of clouds to execute HPC applications, this project aims to address these various aspects. The importance of using clouds to run HPC applications can be observed by some initiatives, such as UberCloud, which has offered HPC cloud service where users can discuss the experience of using such an environment. As a case study, we consider mainly experiments in the area of bioinformatics and, in particular, comparative genomics.

ADMITS

Title: Architecting Distributed Monitoring and Analytics for IoT in Disaster Scenarios

Begin date: Wed Jan 01 2020

Local supervisor: Luciana Arantes

Partners:

- Universidad Tecnica Federico Santa Maria
- Universidade Federal do Rio Grande do Norte
- Universidad de la Republica Uruguay

Inria contact: *Luciana Arantes*

Summary: The ADMITS (Architecting Distributed Monitoring and analytics for IoT in disaster Scenarios) project aims to develop algorithms, protocols and architectures to enable a decentralized distributed computing environment to provide support for monitoring, failure detection, and analytics in IoT disaster scenarios. We face a context where every year, millions of people are affected by natural and man-made disasters, whereby governments all around the world spend huge amounts of resources on preparation, immediate response, and reconstruction. Since November 2015, severe weather brought on by El Ni~no Southern Oscillation (ENSO), including heavy rains, floods, flash floods and landslides significantly hit South America, causing thousands of homelessness and deaths. In Brazil, the Emergency Management Service early reports thousands of households affected by the rainstorms, Landslides, drought, and, very recently, the devastating mudflows caused by Mariana and Brumadinho dam disasters. In Uruguay, the National Emergency System's (SINAE) reports thousands of people were displaced by flooding caused by heavy rains, as well as a tornado, which destroyed homes in various areas in the country. Chile attains special attention in Latin America as it is by far the most natural disaster-prone country. Chile is one of the most earthquake-prone countries in the world, mainly due to its location along the Pacific Ring of Fire, an area of intense volcanic activity and earthquakes. Chile is affected by drought, floods, tsunamis, volcanic eruptions, forest fires, earthquakes (in April 2014, a powerful 8.2 magnitude earthquake struck near Chile's northern coast prompted a tsunami and strong aftershocks) and wildfire (the most devastating in its history was in January 2017). Recently, the Internet of Things (IoT) paradigm has been extensively used for efficiently managing disaster scenarios, such as volcanic disasters, floods, forest fire, landslides, earthquakes, urban disasters, industrial and terrorists attacks, and so on. The IoT support can provide key capabilities to localize victims, achieve situation awareness, and monitor/actuate the environment. However, in a disaster scenario the communication/processing infrastructure and the devices themselves may fail producing either temporary or permanent network partitions and loss of information. Moreover, it is expected that in the years to come, IoT will generate large amounts of data everyday, making data processing and analysis very difficult and challenging in time-critical applications

9.2 International research visitors

9.2.1 Informal international cooperation

Let us report on the following informal international collaborations:

- Annette Bieniusa (TU Kaiserslautern, Germany) on the design and implementation of an efficient database backend. She co-advised the Masters' internship of Ayush Pandey.
- Annette Bieniusa (TU Kaiserslautern, Germany), Carla Ferreira (Universidade NOVA de Lisboa, Portugal) and Gustavo Petri (Université Paris-Diderot, then ARM Research, Cambridge, UK), on the formalisation and proof of an advanced distributed database. This is part of the thesis topic of Saalik Hatia.
- Annette Bieniusa (TU Kaiserslautern, Germany), Nuno Preguiça and João Leitão (Universidade NOVA de Lisboa, Portugal), and Carlos Baquero (Universidade do Minho, Portugal), on distributed data management and consistency. This is relevant to several research projects in the group.
- Carla Ferreira (Universidade NOVA de Lisboa, Portugal) on the design and proof of a highly-available replicated tree data structure. This work is reported in the PhD thesis of Sreeja Nair [32].

9.2.2 Visits of international scientists

Other international visits to the team

Maria Clicia STELLING DE CASTRO

Status: Professor

Institution of origin: UERJ - Universidade do Estado do Rio de Janeiro

Country: Brazil

Dates: From September to December

Context of the visit: ReMatch project

Mobility program/type of mobility: research stay

Lucia DRUMMOND

Status: Professor

Institution of origin: UFF- Universidade Federal Fluminense

Country: Brazil

Dates: October

Context of the visit: ReMatch project

Mobility program/type of mobility: research stay

9.3 National initiatives

9.3.1 ANR

AdeCoDS (2019–2023)

Title: Programming, verifying, and synthesizing Adequately-Consistent Distributed Systems (AdeCoDS).

Members: Université de Paris (project leader), Sorbonne-Université LIP6, ARM, Orange.

Funding: The total funding of AdeCoDS from ANR is 523 471 euros, of which 162 500 euros for Delys.

Objectives The goal of the project is to provide a framework for programming distributed systems that are both correct and efficient (available and performant). The idea is to offer to developers a programming framework where it is possible, for a given application, (1) to build implementations that are correct under specific assumptions on the consistency level guaranteed by the infrastructure (e.g., databases and libraries of data structures), and (2) to discover in a systematic way the different trade-offs between the consistency level guaranteed by the infrastructure and the type and the amount of synchronization they need to use in their implementation in order ensure its correctness. For that, the project will develop a methodology based on combining (1) automated verification and synthesis methods, (2) language-based methods for correct programming, and (3) techniques for efficient system design.

ESTATE - (2016–2021)

Members: LIP6 (DELYS, 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 DELYS.

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.

RainbowFS - (2016–2022)

Members: LIP6 (DELYS, project leader), Scality SA, CNRS-LIG, Télécom Sud-Paris, Université Savoie-Mont-Blanc.

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

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.

SeMaFoR - (2021–2024)

Members: LS2N-IMT Atlantique (project leader), LIP6 (DELYS), AlterWay.

Funding: is funded by ANR (PRCE) for a total of 506 787 euros, of which 157 896 euros for DELYS.

Objectives: The goal is to propose an autonomic Fog system designed in a generic way. To this end, we will address several open challenges: 1) Provide an Architecture Description Language (ADL) for modeling Fog systems and their specific features such as the locality concept, QoS constraints applied on resources and their dependencies, the dynamicity of considered workloads, etc. This ADL should be generic and customizable to address any possible kind of Fog system. 2) Support collaborative decision-making between a fleet of small autonomic controllers distributed over the Fog. Tackling the convergence of local decisions to obtain a shared and consistent decision among these autonomic controllers requires new distributed agreement protocols based on distributed consensus algorithms. 3) Support the automatic generation and coordination of reconfiguration plans between the autonomic controllers. Even if each controller gets a new local target configuration to apply from the consensus, the execution plan of the overall reconfiguration needs to be generated and coordinated to minimize the disruption time and avoid failures. 4) Design and implement a fully open source framework usable in a standalone way or integrated with standard solutions (e.g., Kubernetes). The project targets in particular the future generation of Fog architects, DevOps engineers. We plan to evaluate the solution both on simulated Fog infrastructures as well as real infrastructures.

The local coordinator of SeMaFor in Delys is Jonathan Lejeune.

9.3.2 Informal national cooperation

Let us report on some additional national collaborations:

- Ahmed Bouajjani (Université de Paris), on the design of a programming environment for building distributed systems by correct composition. Ahmed Bouajjani is PhD co-advisor of Laurent Prosperi.
- Pierre Sutra (Télécom SudParis, France), on consistency protocols, especially for edge databases. Pierre Sutra contributed to the Colony project [19].

10 Dissemination

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- Chair of Steering Committee of Workshop on Principles and Practice of Consistency for distributed Data (PaPoC), M. Shapiro.

10.1.2 Scientific events: selection

Member of the conference program committees

- Marc Shapiro, European Conference on Computer Systems 2021 and 2022 (EuroSys, ACM).
- Marc Shapiro, Operating Systems Design and Implementation 2022 (OSDI, ACM, Usenix).
- Franck Petit, 35th International Symposium on Distributed Computing (DISC 2021).
- Pierre Sens, 20th IEEE International Symposium on Network Computing and Applications (NCA 2021).
- Luciana Arantes 41st IEEE International Conference on Distributed Computing Systems (ICDCS 2021).
- Luciana Arantes, 20th IEEE International Symposium on Network Computing and Applications (NCA 2021).
- Luciana Arantes 10th Latin-American Symposium on Dependable Computing (LADC 2021).
- Luciana Arantes 18th Annual IFIP International Conference on Network and Parallel Computing (IFIP NPC 2021).

10.1.3 Journal

Member of the editorial boards

- Pierre Sens, International Journal of High Performance Computing and Networking (IJHPCN).
- Luciana, Journal of Parallel and Distributed Computing (JPDC).

10.2 Service and responsibilities

- Marc Shapiro, Member of the Board of Société informatique de France (SiF), the French learned society in informatics.
- Marc Shapiro, member of ACM Europe working group on European Research Visibility (RAISE).
- Marc Shapiro, member of working group on PhD Studies and Professional Integration of Cossaf, the federation of French academic learned societies.

10.2.1 Research administration

- Colette Johnen, since 2020: Member of section 27 of Conseil national des Universités.
- Pierre Sens, until Aug 2021: Member of Section 6 of the national committee for scientific research CoNRS.
- Franck Petit, Pierre Sens, since 2012: Member of the Executive Committee of Labex SMART, CoChairs of Track 4, Autonomic Distributed Environments for Mobility.

10.3 Teaching - Supervision - Juries

10.3.1 Teaching

- Julien Sopena is Member of "Directoire des formations et de l'insertion professionnelle" of Sorbonne Université, France
- Master: Julien Sopena is responsible of Computer Science Master's degree in Distributed systems and applications (in French, SAR), Sorbonne Universités, France

- Master: Luciana Arantes, Swan Dubois, Jonathan Lejeune, Franck Petit, Pierre Sens, Julien Sopena, Advanced distributed algorithms, M2, Sorbonne Université, France
- Master: Jonathan Lejeune, Designing Large-Scale Distributed Applications, M2, Sorbonne Université, France
- Master: Maxime Lorrillere, Julien Sopena, Linux Kernel Programming, M1, Sorbonne Université, France
- Master: Luciana Arantes, Swan Dubois, Jonathan Lejeune, Pierre Sens, Julien Sopena, Operating systems kernel, M1, Sorbonne Université, France
- Master: Luciana Arantes, Swan Dubois, Franck Petit, Distributed Algorithms, M1, Sorbonne Université, France
- Master: Franck Petit, Autonomic Networks, M2, Sorbonne Université, France
- Master: Franck Petit, Distributed Algorithms for Networks, M1, Sorbonne Université, France
- Master: Jonathan Lejeune, Julien Sopena, Client-server distributed systems, M1, Sorbonne Université, France.
- Master: Luciana Arantes, Pierre Sens, Franck Petit. Cloud Computing, M1, EIT Digital Master, France.
- Master: Julien Sopena, Marc Shapiro, Ilyas Toumlilt, Francis Laniel. Kernels and virtual machines (*Noyaux et machines virtuelles*, NMV), M2, Sorbonne Université, France.
- Licence: Pierre Sens, Luciana Arantes, Julien Sopena, Principles of operating systems, L3, Sorbonne Université, France
- Licence: Swan Dubois, Initiation to operating systems, L3, Sorbonne Université, France
- Licence: Swan Dubois, Multi-threaded Programming, L3, Sorbonne Université, France
- Licence: Jonathan Lejeune, Oriented-Object Programming, L3, Sorbonne Université, France
- Licence: Franck Petit, Advanced C Programming, L2, Sorbonne Université, France
- Licence: Swan Dubois, Jonathan Lejeune, Franck Petit, Julien Sopena, Introduction to operating systems, L2, Sorbonne Université, France
- Licence: Mesaac Makpangou, C Programming Language, 27 h, L2, Sorbonne Université, France
- Ingénieur 4ème année : Marc Shapiro, Introduction aux systèmes d'exploitation, 26 h, M1, Polytech Sorbonne Université, France.
- Licence : Philippe Darche (coordinator), Architecture of Internet of Things (IoT), 2 × 32h, L3, Institut Universitaire Technologique (IUT) Paris Descartes, France.
- Engineering School: Philippe Darche (coordinator), Solid-State Memories, 4th year, ESIEE, France.
- DUT: Philippe Darche (coordinator), Introduction to Computer Systems - Data representation, 60h, Institut Universitaire Technologique (IUT) Paris Descartes, France.
- DUT: Philippe Darche (coordinator), Computer Architecture, 32h, Institut Universitaire Technologique (IUT) Paris Descartes, France.
- DUT: Philippe Darche (coordinator), Computer Systems Programming, 80h, Institut Universitaire Technologique (IUT) Paris Descartes, France.

10.3.2 PhD Advising

- CIFRE PhD: José Alves Esteves, "Adaptation dynamique en environnements répartis contraints", Sorbonne Univ. Dec. 2021. Advised by Pierre Sens and Amina Boubendir Orange Labs.
- PhD: Sreeja Nair, “Designing safe and highly available distributed applications,” Sorbonne Univ., Jul. 2021. Advised by Marc Shapiro.
- CIFRE PhD: Jonathan Sid-Otmane. “Étude des contraintes de cohérence des données dans la 5G, appliquée aux limitations d’usage de ressources dans les slices réseau,” Dec. 2021. Advised by Marc Shapiro with Sofiane Imadali and Frédéric Martelli, Orange Labs.
- PhD: Ilyas Toumlilt, “Colony: A Hybrid Consistency System for Highly-Available Collaborative Edge Computing,” Sorbonne Univ., Dec. 2021. Advised by Marc Shapiro.
- CIFRE PhD: Dimitrios Vasilas, “A flexible and decentralised approach to query processing for geo-distributed data systems”, Sorbonne Univ., Jul. 2021. Advised by Marc Shapiro, with Brad King, Scality.
- PhD in progress: Aymeric Agon-Rambosson, "Maintien du groupes dans un environnement hautement hétérogène et dynamique", Sorbonne Univ., since Oct. 2021. Advised by Pierre Sens and Jonathan Lejeune.
- PhD in progress: Arnaud Favier, "Election de leader dans les réseaux dynamiques", Sorbonne Univ., since Sep. 2018. Advised by Pierre Sens and Luciana Arantes.
- PhD in progress: Célia Mahamdi, "Prise de décision collaborative dans un système distribué et dynamique", Sorbonne Univ., since Sep. 2020. Advised by Mesaac Makpongou and Jonathan Lejeune.
- PhD in progress: Saalik Hatia, “Efficient management of memory and storage for CRDTs,” Sorbonne Univ., since Oct. 2018. Advised by Marc Shapiro.
- PhD in progress: Gabriel Le Bouder, "Autonomic synchronization", Sorbonne Univ., since Sep. 2019. Advised by Franck Petit.
- PhD in progress: Benoît Martin, “Protocol de cohérence hybride: de la cohérence causale à la cohérence forte,” Sorbonne Univ., since Sep. 2019. Advised by Mesaac Makpangou and Marc Shapiro.
- PhD in progress: Laurent Prosperi, “Abstractions, langage et runtime pour les systèmes distribués,” Sorbonne Univ., since Sep. 2019. Advised by Marc Shapiro.
- PhD in progress: Daniel Wladdimiro, "Adaptation dynamique en environnements répartis contraints", Sorbonne Univ., since Sep. 2019. Advised by Pierre Sens and Luciana Arantes.
- PhD in progress: Daniel Wilhelm, "Algorithmes de diffusion causale dans les systèmes répartis dynamique", Sorbonne Univ., since Oct. 2019, Pierre Sens and Luciana Arantes.

10.3.3 Juries

Pierre Sens was the reviewer of

- Flavien Vernier, HDR, LISTIC, Univ. Savoie Mont Blanc
- Mozhdeh Farhadi, PhD, IRISA, Univ. Rennes 1
- Pedro Penna, PhD, LIG, Univ. Grenoble Alpes
- Ugaitz Amozarrain, PhD, San Sebastien, Univ. Basque (Espagne)

Pierre Sens was Chair of

- Tuanir França Rezende, PhD, SAMOVAR, Inst. Polytechnique Paris
- Jonathan Sid-Otmane, PhD, LIP6, Sorbonne Univ.

11 Scientific production

11.1 Major publications

[1] V. Balegas, N. Preguiça, R. Rodrigues, S. Duarte, C. Ferreira, M. Najafzadeh and M. Shapiro. 'Putting Consistency back into Eventual Consistency'. In: *euroconfon # Comp.\Sys.\(EuroSys)*. Bordeaux, France, Apr. 2015, 6:1–6:16. DOI: 10.1145/2741948.2741972. URL: https://doi.org/10.1145/2741948.2741972.

[2] L. Blin, L. Feuilloley and G. Le Bouder. 'Optimal Space Lower Bound for Deterministic Self-Stabilizing Leader Election Algorithms'. In: *OPODIS*. OPODIS 2021 - International Conference on Principles of Distributed Systems. Strasbourg, France, Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03536828.

[3] S. Dubois, R. Guerraoui, P. Kuznetsov, F. Petit and P. Sens. 'The weakest failure detector for eventual consistency'. In: *Distributed Computing* 32.6 (Dec. 2019), pp. 479–492. DOI: 10.1007/s00446-016-0292-9. URL: https://hal.inria.fr/hal-02413314.

[4] A. Gotsman, H. Yang, C. Ferreira, M. Najafzadeh and M. Shapiro. ''Cause I'm Strong Enough: Reasoning about Consistency Choices in Distributed Systems'. In: *sympon # Principles of Prog.\Lang.\(POPL)*. St.~Petersburg, FL, USA, 2016, pp. 371–384. DOI: 10.1145/2837614.2837625. URL: http://dx.doi.org/10.1145/2837614.2837625.

[5] B. Lepers, R. Gouicem, D. Carver, J.-P. Lozi, N. Palix, M.-V. Aponte, W. Zwaenepoel, J. Sopena, J. Lawall and G. Muller. 'Provable Multicore Schedulers with Ipanema: Application to Work Conservation'. In: Eurosys 2020 - European Conference on Computer Systems. Heraklion / Virtual, Greece, 27th Apr. 2020. DOI: 10.1145/3342195.3387544. URL: https://hal.inria.fr/hal-02554342.

[6] J. Peeters, N. Ventroux, T. Sassolas and M. Shapiro. *Distributing computing system implementing a non-speculative hardware transactional memory and a method for using same for distributed computing*. Patent awarded US 10 416 925 B2. United States Patent and Trademark Office (USPTO), Sept. 2019.

[7] M. Shapiro, N. Preguiça, C. Baquero and M. Zawirski. 'Conflict-free Replicated Data Types'. In: *intsympon # Stabilization, Safety, and Security of Dist.\Sys.\(SSS)*. Ed. by X. Défago, F. Petit and V. Villain. Vol. 6976. Lecture Notes in Comp.\Sc. Grenoble, France: Springer-Verlag, Oct. 2011, pp. 386–400. URL: http://lip6.fr/Marc.Shapiro/papers/CRDTs%5C_SSS-2011.pdf.

[8] M. Zawirski, N. Preguiça, S. Duarte, A. Bieniusa, V. Balegas and M. Shapiro. 'Write Fast, Read in the Past: Causal Consistency for Client-side Applications'. In: *intconfon # Middleware (MIDDLEWARE)*. ACM/IFIP/Usenix. Vancouver, BC, Canada, Dec. 2015, pp. 75–87.

11.2 Publications of the year

International journals

[9] S. Devismes, A. Lamani, F. Petit, P. Raymond and S. Tixeuil. 'Terminating Exploration Of A Grid By An Optimal Number Of Asynchronous Oblivious Robots'. In: *The Computer Journal*. The Computer Journal 64.1 (Jan. 2021), pp. 132–154. DOI: 10.1093/comjnl/bxz166. URL: https://hal.archives-ouvertes.fr/hal-02363013.

International peer-reviewed conferences

[10] K. Altisen, S. Devismes, A. Durand, C. Johnen and F. Petit. 'On Implementing Stabilizing Leader Election with Weak Assumptions on Network Dynamics'. In: PODC '21: ACM Symposium on Principles of Distributed Computing. Virtual Event, Italy: ACM, 26th July 2021, pp. 21–31. DOI: 10.1145/3465084.3467917. URL: https://hal.archives-ouvertes.fr/hal-03346225.

[11] K. Altisen, S. Devismes, A. Durand, C. Johnen and F. Petit. ‘Self-stabilizing Systems in Spite of High Dynamics’. In: 22nd International Conference on Distributed Computing and Networking, ICDCN’21. ICDCN ’21: International Conference on Distributed Computing and Networking 2021. Nara, Japan, Jan. 2021, pp. 156–165. DOI: 10.1145/3427796.3427838. URL: https://hal.archives-ouvertes.fr/hal-02376832.

[12] ***Best Paper***
J. J. Alves Esteves, A. Boubendir, F. Guillemin and P. Sens. ‘DRL-based Slice Placement under Realistic Network Load Conditions’. In: CNSM 2021 - 17th International Conference on Network and Service Management. Izmir, Turkey, 25th Oct. 2021. URL: https://hal.inria.fr/hal-03516310.

[13] J. J. Alves Esteves, A. Boubendir, F. Guillemin and P. Sens. ‘DRL-based Slice Placement Under Non-Stationary Conditions’. In: CNSM 2021 - 17th International Conference on Network and Service Management. Izmir, Turkey, 25th Oct. 2021. URL: https://hal.inria.fr/hal-03332502.

[14] ***Best Paper***
L. Blin, L. Feuilloley and G. Le Bouder. ‘Optimal Space Lower Bound for Deterministic Self-Stabilizing Leader Election Algorithms’. In: *OPODIS*. OPODIS 2021 - International Conference on Principles of Distributed Systems. LIPIcs. Strasbourg, France, Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03536828.

[15] ***Best Paper***
A. Favier, L. Arantes, J. Lejeune and P. Sens. ‘Centrality-Based Eventual Leader Election in Dynamic Networks’. In: NCA 2021C- 20th IEEE International Symposium on Network Computing and Applications. Cambridge, Boston, United States, 23rd Nov. 2021. URL: https://hal.inria.fr/hal-03452072.

[16] Y. Ghigoff, J. Sopena, K. Lazri, A. Blin and G. Muller. ‘BMC: Accelerating Memcached using Safe In-kernel Caching and Pre-stack Processing’. In: NSDI’21 - 18th USENIX Symposium on Networked Systems Design and Implementation. Virtual event, United States: USENIX Association, 12th Apr. 2021, pp. 487–501. URL: https://hal.inria.fr/hal-03361644.

[17] H. Heydari, G. Silvestre and L. Arantes. ‘Efficient Consensus-Free Weight Reassignment for Atomic Storage’. In: NCA 2021 - 20th IEEE International Symposium on Network Computing and Applications. Virtual, France, 23rd Nov. 2021. URL: https://hal-enac.archives-ouvertes.fr/hal-03454633.

[18] C. Johnen, L. Arantes and P. Sens. ‘FIFO and Atomic broadcast algorithms with bounded message size for dynamic systems’. In: SRDS 2021 - 40th International Symposium on Reliable Distributed Systems. Chicago / Virtual, United States, 20th Sept. 2021. URL: https://hal.inria.fr/hal-03332423.

[19] I. Toumlilt, P. Sutra and M. Shapiro. ‘Highly-available and consistent group collaboration at the edge with colony’. In: *Middleware ’21: Proceedings of the 22nd International Middleware Conference.* Middleware 2021: 22nd International Middleware Conference. Québec / Virtual, Canada: ACM, 2nd Dec. 2021, pp. 336–351. DOI: 10.1145/3464298.3493405. URL: https://hal.inria.fr/hal-03353663.

[20] R. Vaillant, D. Vasilas, M. Shapiro and T. L. Nguyen. ‘CRDTs for truly concurrent file systems’. In: HotStorage ’21 -13th ACM Workshop on Hot Topics in Storage and File Systems. Virtual, France, 27th July 2021. URL: https://hal.inria.fr/hal-03278658.

[21] D. Wilhelm, L. Arantes and P. Sens. ‘A scalable causal broadcast that tolerates dynamics of mobile networks’. In: 23rd International Conference on Distributed Computing and Networking (ICDCN). New Delhi / Virtual, India, 4th Jan. 2022. URL: https://hal.inria.fr/hal-03524944.

[22] D. Wladdimiro, L. Arantes, P. Sens and N. Hidalgo. ‘A Multi-Metric Adaptive Stream Processing System’. In: NCA 2021 - 20th IEEE International Symposium on Network Computing and Applications. Cambridge, Boston, United States, 23rd Nov. 2021. URL: https://hal.inria.fr/hal-03516376.

Conferences without proceedings

[23] N. Maurice, J. Sopena and L. Lacassagne. 'Un nouvel algorithme efficace de Split & Merge pour systèmes embarqués'. In: COMPAS 2021 - Conférence francophone d'informatique en Parallélisme, Architecture et Système. Lyon, France, 5th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03330463.

Scientific books

[24] P. Darche. *Le Microprocesseur 1 : fonctions de calcul et de mémorisation, modèles de calcul et architecture des ordinateurs.* ISTE Ltd, 1st June 2021. URL: https://hal.sorbonne-universite.fr/hal-03260877.

[25] P. Darche. *Le Microprocesseur 2 : communication dans un système numérique.* ISTE Ltd, 1st June 2021. URL: https://hal.sorbonne-universite.fr/hal-03260885.

[26] P. Darche. *Le Microprocesseur 3 : aspects matériels.* ISTE Ltd, 1st June 2021. URL: https://hal.sorbonne-universite.fr/hal-03260888.

[27] P. Darche. *Le Microprocesseur 4 : aspects logiciels.* ISTE Ltd, 1st June 2021. URL: https://hal.sorbonne-universite.fr/hal-03260893.

[28] P. Darche. *Le Microprocesseur 5 : aspects logiciels et matériels du développement, du débogage et du test.* ISTE Ltd, 1st June 2021. URL: https://hal.sorbonne-universite.fr/hal-03260897.

[29] P. Darche. *Microprocessor 4. Core Concepts: Software Aspects.* ISTE Ltd and John Wiley & Sons, Inc., 1st Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03120713.

[30] P. Darche. *Microprocessor 5. Software and Hardware Aspects of Development, Debugging and Testing – The Microcomputer.* ISTE Ltd and John Wiley & Sons, Inc., 1st Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03120718.

Doctoral dissertations and habilitation theses

[31] J. J. Alves Esteves. 'Optimization of Network Slice Placement in Distributed Large Scale Infrastructures From Heuristics to Controlled Deep Reinforcement Learning'. Sorbonne Universites, UPMC University of Paris 6; Orange Labs, 13th Dec. 2021. URL: https://hal.inria.fr/tel-03500387.

[32] S. S. Nair. 'Designing safe and highly available distributed applications'. Sorbonne Université, 1st July 2021. URL: https://tel.archives-ouvertes.fr/tel-03339393.

[33] J. Sid-Otmane. 'A study of data consistency constraints in 5G, applied to limiting resource usage in network slices'. Sorbonne Universite, 13th Dec. 2021. URL: https://tel.archives-ouvertes.fr/tel-03539545.

[34] I. Toumlilt. 'Colony: A Hybrid Consistency System for Highly-Available Collaborative Edge Computing'. Sorbonne Université, 21st Dec. 2021. URL: https://hal.inria.fr/tel-03538565.

[35] D. Vasilas. 'A flexible and decentralised approach to query processing for geo-distributed data systems'. Sorbonne Université, 19th Feb. 2021. URL: https://hal.inria.fr/tel-03272208.

Reports & preprints

[36] C. Johnen and M. Haddad. *Efficient self-stabilizing construction of disjoint MDSs in distance-2 model.* Inria Paris, Sorbonne Université; LaBRI, CNRS UMR 5800; LIRIS UMR CNRS 5205, 11th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03138979.

[37] S. S. Nair, F. Meirim, M. Pereira, C. Ferreira and M. Shapiro. *A coordination-free, convergent, and safe replicated tree.* RR-9395. LIP6, Sorbonne Université, Inria de Paris; Universidade nova de Lisboa, 23rd Feb. 2021, p. 36. URL: https://hal.archives-ouvertes.fr/hal-03150817.

11.3 Cited publications

[38] S. Hatia and M. Shapiro. *Specification of a Transactionally and Causally-Consistent (TCC) database.* Research Report RR-9355. DELYS ; LIP6, Sorbonne Université, Inria, Paris, France, July 2020. URL: https://hal.inria.fr/hal-02902474.